Photometric redshifts and clustering statistics of the large-scale structure

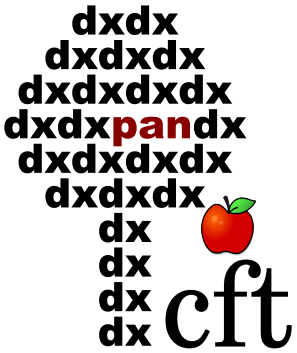
Author:

Anjitha John William

Supervisors:

dr. hab. Maciej Bilicki, prof. CFT PAN

dr. hab. Wojciech A. Hellwing, prof. CFT PAN



Center for Theoretical Physics

Polish Academy of Sciences

Thesis submitted to the Center for Theoretical Physics of the Polish Academy of Sciences in accordance with the requirements for the degree of Doctor of Philosophy in Physics

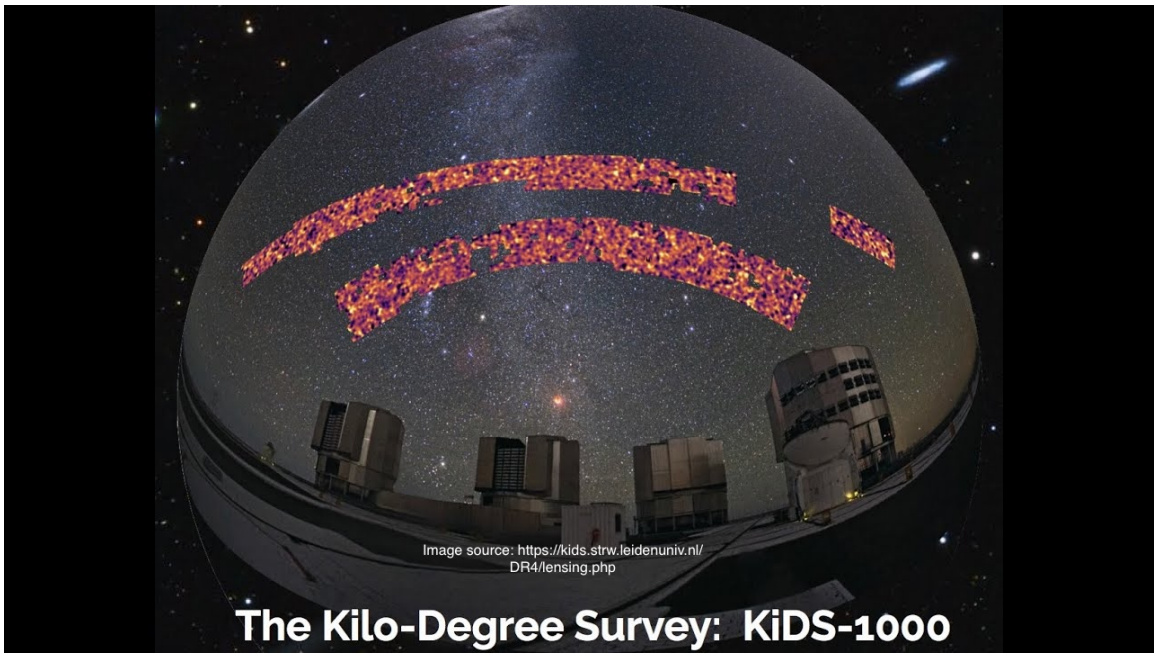
April, 2026

Dedicated to my mother

For your love, support and sacrifices

"The eternal mystery of the world is its comprehensibility."

- Albert Einstein [43]



Abstract

According to the standard ΛCDM model, the Universe is composed of dark energy, radiation, dark matter (DM), and baryonic matter. The distribution of total matter (DM + baryons) can be traced through the clustering of large-scale structure (LSS) such as galaxies and quasars. LSS surveys are conducted using both photometric and spectroscopic techniques. Spectroscopic surveys provide precise redshift measurements (spec-zs) from spectra and enable the three-dimensional mapping of LSS. However, they are observationally expensive and time-consuming. In contrast, photometric surveys can observe large areas of the sky, detecting millions of objects to great depths within a relatively short time. Their main limitation is that they do not provide direct redshift measurements. Obtaining spec-zs for all objects in photometric surveys is impractical due to the long exposure times required, which poses a significant challenge for using photometric data in cosmological analyses where redshift is a fundamental quantity.

This thesis addresses this challenge by estimating redshifts from photometric quantities using advanced machine learning algorithms. These estimates, known as photometric redshifts (photo-zs), are less precise than spec-zs. Although photo-zs are not reliable for full three-dimensional clustering due to their uncertainties, they can be effectively used for angular (two-dimensional) clustering measurements. Although redshift information is not needed to measure the angular clustering in observed data, the redshift distribution is essential for modelling this observable. This distribution is derived from photo-zs. Since galaxies are biased tracers of total matter, while dark matter itself is not directly observable, this modelling allows us to estimate the bias and better understand the relationship between visible structures and the total matter distribution in our Universe.

A central contribution of this thesis is the development of the deep learning framework `Hybrid-z` for photo-z estimation. This model combines four-band optical imaging data with nine-band photometric magnitudes through a hybrid architecture of Convolutional Neural Networks and Artificial Neural Networks. Applied to the Kilo-Degree Survey Data Release 4 (KiDS-DR4) bright galaxy catalogue, `Hybrid-z` significantly improves the photo-zs. It achieves a ∼20% reduction in scatter compared to the previous method that relies only on photometric magnitudes rather than imaging data. The model is publicly released as an open-source tool, and the resulting photo-z catalogue for the KiDS-DR4 bright galaxy sample is made available to the community.

`Hybrid-z` is further applied to KiDS-DR4 quasars to construct redshift distributions and analyse their clustering using the angular two-point correlation function. The best-fit scale-independent quasar bias increases from $b \approx 1.6$ at $z \approx 0.6$ to $b \approx 4.0$ at $z \approx 2.2$, following a quadratic trend with redshift. The clustering results suggest that quasars inhabit DM halos with masses $\log_{10}(M_{\rm eff}/h^{-1}M_{\odot}) \sim 12.7$–$12.9$ and peak heights $\nu_{\rm eff}$ rising from $\sim$ 1.5 to 2.9 over the considered redshift range. We also studied the systematic

effects on bias calculation. It shows that stellar contamination has a negligible effect, while the redshift distribution significantly affects the inferred bias, highlighting the need for accurate redshift calibration.

Beyond two-point statistics, this thesis studies the non-Gaussian features of the matter distribution using higher-order statistics derived from the Count-in-Cells method. Measurements of average correlation function and reduced cumulants in photo-z bins of KiDS-DR4 bright sample reveal significant non-Gaussianity and its evolution with redshift. The clustering signal is strongest on small angular scales and decreases with increasing scale, with indications of observational systematics on large scales. Finally, the dependence of clustering on galaxy properties is examined. Galaxies with higher stellar mass and redder colours exhibit stronger clustering, indicating that they preferentially inhabit denser environments.

These results provide insights into the relationship between luminous tracers in KiDS-DR4 and the underlying matter distribution, and demonstrate the potential of using machine learning for photo-z estimation with clustering analysis for next-generation LSS surveys.

The abstract is translated into Polish by Piotr Płonka.

Streszczenie

Zgodnie ze standardowym modelem ΛCDM, Wszechświat składa się z ciemnej energii, promieniowania, ciemnej materii oraz materii barionowej. Rozkład gęstości materii (ciemnej materii i barionowej) można śledzić poprzez analizę grupowania się struktur wielkoskalowych (LSS), takich jak galaktyki i kwazary. Badania tych struktur przeprowadza się przy użyciu zarówno technik fotometrycznych, jak i spektroskopowych. Przeglądy spektroskopowe dostarczają precyzyjnych pomiarów przesunięć ku czerwieni (spec-zs) na podstawie widm i umożliwiają trójwymiarowe mapowanie wielkoskalowych struktur kosmicznych. Są one jednak kosztowne pod względem obserwacyjnym i czasochłonne. Z drugiej strony, badania fotometryczne pozwalają obserwować duże obszary nieba, wykrywając miliony obiektów na dużych głębokościach w stosunkowo krótkim czasie. Ich głównym ograniczeniem jest to, że nie zapewniają one bezpośrednich pomiarów przesunięcia ku czerwieni. Uzyskanie wartości spec-zs dla wszystkich obiektów w badaniach fotometrycznych jest niepraktyczne ze względu na wymagane długie czasy naświetlania, co stanowi poważne wyzwanie dla wykorzystania danych fotometrycznych w analizach kosmologicznych, w których przesunięcie ku czerwieni jest kluczową wielkością.

W niniejszej pracy podjęto to wyzwanie poprzez oszacowanie przesunięć ku czerwieni na podstawie danych fotometrycznych przy użyciu zaawansowanych algorytmów uczenia maszynowego. Szacunki te, znane jako fotometryczne przesunięcia ku czerwieni (photo-z), są jednak mniej precyzyjne niż spec-z. Chociaż wartości photo-z nie są dośc precyzyjne w przypadku pełnego trójwymiarowego grupowania się galaktyk ze względu na związane z nimi niepewności, można je skutecznie wykorzystać do pomiarów grupowania się kątowego (dwuwymiarowego). Chociaż informacje o przesunięciu ku czerwieni nie są potrzebne do pomiaru grupowania kątowego w obserwowanych danych, rozkład przesunięć ku czerwieni jest niezbędny do modelowania tej wielkości obserwowalnej. Rozkład ten jest wyznaczany na podstawie wartości photo-zs. Ponieważ galaktyki są obciążonymi wskaźnikami całkowitego rozkładu materii, podczas gdy sama ciemna materia nie jest bezpośrednio obserwowalna, modelowanie to pozwala nam oszacować tzw. *bias* ("parametr obciążenia") i lepiej zrozumieć związek między widocznymi strukturami a rozkładem gęstości materii we Wszechświecie.

Głównym osiągnięciem niniejszej pracy jest opracowanie platformy uczenia głębokiego `Hybrid-z` służącej do szacowania przesunięcia ku czerwieni na podstawie danych optycznych. Model ten łączy dane obrazowe z czterech pasm optycznych z dziewięciopasmowymi wielkościami fotometrycznymi za pomocą hybrydowej architektury opartej na konwolucyjnych sieciach neuronowych i sztucznych sieciach neuronowych. Zastosowany do katalogu jasnych galaktyk Kilo-Degree Survey Data Release 4 (KiDS-DR4), `Hybrid-z`

znacznie poprawia oszacowania foto-z. Osiąga on około 20% redukcję rozrzutu w porównaniu z poprzednią metodą, która opierała się wyłącznie na wielkościach fotometrycznych, a nie na danych z obrazów. Model został udostępniony publicznie jako narzędzie open-source, a wynikowy katalog foto-z dla próbki jasnych galaktyk KiDS-DR4 jest dostępny dla społeczności.

Metodę `Hybrid-z` zastosowano również do kwazarów z katalogu KiDS-DR4 w celu sporządzenia rozkładów przesunięć ku czerwieni oraz analizy ich grupowania przy użyciu kątowej funkcji korelacji dwupunktowej. Najlepiej dopsaowany, niezależny od skali *bias* kwazarów wzrasta z $b \approx 1,6$ dla $z \approx 0,6$ do $b \approx 4,0$ dla $z \approx 2,2$, wykazując kwadratową zależność od przesunięcia ku czerwieni. Wyniki analizy grupowania sugerują, że kwazary znajdują się w halach ciemnej materii o masach $\log_{10}(M_{\rm eff}/h^{-1}M_{\odot}) \sim 12,7$–$12,9$ i wysokościach pików $\nu_{\rm eff}$ rosnących od $\sim 1,5$ do $2,9$ w rozpatrywanym zakresie przesunięcia ku czerwieni. Zbadano również efekty systematyczne w obliczeniach *bias*. Wynika z nich, że zanieczyszczenie gwiazdami ma znikomy wpływ, podczas gdy rozkład przesunięcia ku czerwieni znacząco wpływa na wyznaczony *bias*, co podkreśla potrzebę dokładnej kalibracji przesunięcia ku czerwieni.

W niniejszej pracy, wykraczając poza statystykę dwupunktową, zbadano niegaussowskie cechy rozkładu materii przy użyciu statystyk wyższego rzędu uzyskanych metodą zliczeń w komórkach. Pomiary uśrednionej funkcji korelacji oraz zredukowanych kumulantów w przedziałach fotometrycznych z jasnej próbki KiDS-DR4 ujawniają znaczącą niegaussowskość oraz jej ewolucję wraz z przesunięciem ku czerwieni. Sygnał grupowania jest najsilniejszy w małych skalach kątowych i słabnie wraz ze wzrostem skali, co wskazuje na systematykę obserwacyjną w dużych skalach. Na koniec zbadano zależność grupowania od właściwości galaktyk. Galaktyki o większej masie gwiazdowej i bardziej czerwonych barwach wykazują silniejsze grupowanie, co wskazuje, że preferują one gęstsze środowiska.

Wyniki te dostarczają wglądu w związek między obiketami pozagalaktycznymi w KiDS-DR4 a leżącym u ich podstaw rozkładem materii oraz pokazują potencjał wykorzystania uczenia maszynowego do szacowania foto-z w połączeniu z analizą grupowania w badaniach LSS nowej generacji.

Acknowledgements

As Isaac Newton once said, "If I have seen further, it is by standing on the shoulders of giants." This work has greatly benefited from the contributions of many researchers whose papers and books I have read throughout my PhD journey. While I have made every effort to acknowledge all relevant sources, I apologize for any that may have been unintentionally overlooked.

I would like to express my sincere gratitude to my supervisors, **Prof. Maciej Bilicki** and **Prof. Wojciech A. Hellwing**, for their continuous guidance, insightful feedback, and support throughout my doctoral studies. Their expertise, patience, and encouragement have been invaluable throughout this journey. Our regular meetings and discussions have played a crucial role in shaping my scientific thinking and approach to research problems. I am also grateful for the opportunity to be part of a culturally diverse research environment.

I would also like to thank my colleagues and administration at CFT for a friendly and supportive academic environment. I also extend my gratitude to Post-docs in our group - Mariana, Priyanka, Oliver, and Szymon. I thank the members of the Kilo-Degree Survey collaboration for their collaborative and welcoming spirit. I am also grateful for the people I met during my PhD-Feven, Gursharanjit, Paweł, Suhani, and Midhun- whose presence, friendship and support have meant a lot to me. I gratefully acknowledge the financial support provided by the **National Science Center**, the **Warsaw4PhD doctoral school**, and **CFT PAN**, without which this research would not have been possible.

This acknowledgement would be incomplete without mentioning my two friends who have been by my side since high school. I fondly recall the three of us as school children, engaging in discussion about Newton's laws of motion, a shared curiosity that later guided us through our Bachelor's and Master's studies in Physics. Thirteen years on, I am truly happy to acknowledge them in my PhD thesis and to extend my heartfelt thanks to Adarsha and Archana for their genuine friendship and support throughout all these years. I am also deeply grateful to my childhood friend Annie for her friendship, constant support, and the effort she has made to keep our bond strong over the years. I would also like to thank Aparna, Anjana, Aswani, Gauri, and Sharath, who entered my life at various stages and provided invaluable support and kindness, and took the time to listen throughout this PhD journey, even from a distance of $\sim 10^{-16}$ Mpc. I also thank Juby and Jithya for their sincere friendship and emotional support during some of the most crucial moments of my PhD.

I would especially like to express my gratitude to my best friend and companion, Sarath, for his love, support, and friendship, as well as for the many insightful discussions on life, arts, music, physics, and mathematics. I extend my thanks to my brother and my cousin Kukku for their love, encouragement, and constant support.

Finally, I would like to express my deepest gratitude to my mother for her unconditional love, patience, and encouragement. Her belief in me has always been my greatest source of strength. Thanks for giving me the freedom to grow into myself, without pressure, and for always trusting me to find my own path. Everything I have achieved is built upon her sacrifices and support.

Declaration

The research presented in this thesis was carried out between 2021 and 2026 during my doctoral studies under the supervision of Prof. Maciej A. Bilicki and Prof. Wojciech A. Hellwing at the Center for Theoretical Physics of the Polish Academy of Sciences in Warsaw, Poland. No part of this thesis has been submitted previously for any other degree. The thesis includes following works:

- **John William, Anjitha** ; Jalan, Priyanka; Bilicki, Maciej; Hellwing, Wojciech A.; Thuruthipilly, Hareesh; Nakoneczny, Szymon J.

  Hybrid-z: Enhancing the Kilo-Degree Survey bright galaxy sample photometric redshifts with deep learning

  A&A, 698, A276 (2025)

  https://doi.org/10.1051/0004-6361/202453576

  I am the first author of the included publication and led the research. I designed the methodology, developed the core codebase, and built the `Hybrid-z` model architecture. I performed all data analysis, ran the computations, and generated all figures, except the following.

  The original photometric and spectroscopic dataset, test sample crossmatch with spectroscopic surveys (Table 2) used in this work was provided by Maciej Bilicki. The Lorentzian fitting of the photometric redshift error distribution, including the corresponding figure, was performed by Maciej Bilicki. Initial data preprocessing for the deep learning model was helped by Priyanka Jalan. The idea of smoothing the redshift distribution was proposed by Wojciech A. Hellwing, and its implementation was carried out by Priyanka Jalan. All authors contributed to discussions, interpretation of results, and the writing of the manuscript.

- **John William, Anjitha**; Bilicki, Maciej ; Hellwing, Wojciech A. ; Nakoneczny, Szymon J. ; Jalan, Priyanka

  Angular clustering and bias of photometric quasars in the Kilo-Degree Survey Data Release 4

  This work is currently under review by the Astronomy & Astrophysics Journal.

  https://doi:10.48550/arXiv.2511.17311

  I am the first author of the publication and led the research. I estimated the photometric redshifts of quasars, performed the full clustering analysis, and calculated the quasar bias. I also generated all figures and carried out the associated computations. The original photometric dataset used in this work was provided by Szymon.J Nakoneczny. The training dataset is from Maciej Bilicki. The calculation of halo

mass and peak height were performed by Wojciech A. Hellwing. All authors contributed to scientific discussions, interpretation of the results, and the writing of the manuscript.

Other contribution

- **John William, Anjitha** ; Jalan, Priyanka ; Bilicki, Maciej; Hellwing, Wojciech Deep Learning Based Photometric Redshifts for the Kilo-Degree Survey Bright Galaxy Sample

  Proceedings of the Polish Astronomical Society, vol. 13, 243-248 (2024)

  I am the lead author. I designed the methodology, developed the core codebase, and built the model architecture. I performed all data analysis, ran the computations, and generated all figures. The dataset used in this work was provided by Maciej Bilicki. All authors contributed to discussions, interpretation of results, and the writing of the manuscript.

Contents

1 Introduction **1**

1.1 Classical field theory of gravitation 1

1.1.1 FLRW metric and Friedmann equations 2

1.1.1.1 Friedmann equations 3

1.2 Standard model of cosmology 4

1.3 Redshift 5

1.3.1 Hubble-Lemaître's law 7

1.4 Theory of structure formation 8

1.4.1 Newtonian perturbation theory 9

1.4.2 Linear matter power spectrum 12

1.4.3 Non-linear matter power spectrum 13

1.5 Clustering statistics 14

1.5.1 Clustering of matter overdensity field 15

1.5.1.1 Two-point correlation function of matter density field 16

1.5.2 Clustering of tracer overdensity field 18

1.5.3 Correlation function estimators and bias 19

1.5.4 Angular two-point correlation function 21

1.5.5 Count-in-Cells statistics 24

1.5.5.1 Count-in-Cells statistics of matter over density field 25

1.5.5.2 Count-in-Cells estimator for photometric galaxy sample 27

1.6 Summary 28

2 Photometric Redshift Estimation **30**

2.1 Redshift and luminosity 30

2.2 Spectroscopic and photometric redshifts 31

2.3 Photometric redshift estimation methods 33

2.3.1 Template-based methods 35

2.3.2 Data-driven methods 36

2.3.3 Machine learning 37

2.3.4 Supervised learning 37
2.3.5 Galaxy images and photometric redshift estimation 38
2.3.6 Artificial Neural Networks 39
2.3.7 Universal Approximation Theorem 40
2.3.8 Training of Neural Networks 41
2.3.9 Bias-variance tradeoff 44
2.3.10 Convolutional Neural Networks 45
2.4 Summary 48

3 Hybrid-z: Enhancing the Kilo-Degree Survey bright galaxy sample photometric redshifts with deep learning 50
3.1 Introduction 50

4 Angular clustering and bias of photometric quasars in the Kilo-Degree Survey Data Release 4 65
4.1 Introduction 65

5 Higher order clustering in the Kilo-Degree Survey bright galaxy sample 79
5.1 Data 80
5.2 Count-in-Cells estimator 82
5.3 Results and discussion 83
5.3.1 Redshift bins 83
5.3.2 Dependence on colour and stellar mass 87
5.3.3 Transition to Gaussian behaviour 91
5.3.4 Upturn behaviour 91
5.3.5 Uncertainties 92
5.4 Summary 96
5.5 Future directions 97

6 Summary and future prospects 99

Appendix A Scalar Perturbation 108

Appendix B Striding, Padding, and Pooling operations 110

Part I

Introduction

This chapter introduces the cosmological framework and the clustering statistics employed in the study of structure formation in the universe.

One of the foundational pillars of modern cosmology is the assumption that the Universe is both *homogeneous* and *isotropic* on sufficiently large-scales. *Homogeneity* refers to the property that the Universe appears statistically the same at every point in space, while *isotropy* implies that it looks the same in all directions to any observer. These two assumptions are well supported by cosmological observations.

Gravity is the fundamental interaction in our Universe. Consequently, our exploration of cosmological models is conducted within the theoretical framework of gravity, and the best one we have is Einstein's General Theory of Relativity [42]. Newtonian gravity is insufficient for a fully relativistic description of an expanding Universe on horizon scales; however, Newtonian and weak-field approximations remain accurate and widely used for many large-scale structure (LSS) applications well inside the horizon.

1.1 Classical field theory of gravitation

Newton formulated the theory of gravity and published it in 1686. Newton's universal theory of gravitation is not adequate for cosmology because it assumes instantaneous action at a distance and breaks down in describing the dynamics of an expanding, relativistic Universe. General relativity (GR), put forward in 1915, is a classical field theory of gravity. In GR, the metric is the classical field; more precisely, it is a rank-2 symmetric tensor field. This tensor field determines the distance between points in spacetime.

Euclid's parallelism postulate states that parallel lines will continue to be parallel out to infinity. It was later recognised, in the 18th and 19th centuries, that this holds only in flat (Euclidean) geometry. The convergence or divergence of parallel lines is possible in curved spacetime. The deviation of parallel lines from paralleism provides an intuitive test for the curvature of spacetime. Curvature was mathematically formalised within two-dimensional differential geometry, where Gaussian curvature describes how a surface locally departs from flatness. This is extended to four-dimensional spacetime through the concept of parallel transport [73]. Curvature manifests explicitly when a vector is parallel

transported around an infinitesimal closed loop, resulting in a rotation or displacement upon return to the starting point, reflecting nontrivial holonomy. Mathematically, this change in vector under parallel transport is encoded in the Riemann curvature tensor [88].

The spacetime metric is directly related to the distribution of energy and momentum by Einstein's gravitational Field Equations (EFEs) [118].

$$\begin{aligned} \mathcal{R}_{\mu\nu} - \frac{1}{2}\mathcal{R}g_{\mu\nu} &= \frac{8\pi G}{c^4}T_{\mu\nu} \\ G_{\mu\nu} &= \frac{8\pi G}{c^4}T_{\mu\nu} \end{aligned} \tag{1.1}$$

$\mathcal{R}_{\mu\nu}$ is the Ricci curvature tensor, $\mathcal{R}$ is the contraction of the Ricci tensor and is called the Ricci scalar and $G_{\mu\nu}$ is known as the Einstein tensor. While $T_{\mu\nu}$, the energy-momentum tensor, is a measure of the total energy, momentum and stress of the spacetime. The right-hand side of EFEs is the total energy-momentum content of the Universe. $G \approx 6.67 \times 10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$ is Newton's gravitational constant and $c \approx 3 \times 10^8\,\mathrm{m\,s^{-1}}$ is the speed of light.

1.1.1 FLRW metric and Friedmann equations

Einstein's field equations are second-order, coupled, non-linear partial differential equations for the spacetime metric components. Obtaining exact solutions in full generality is difficult; however, many physically relevant exact solutions exist under symmetry assumptions. While many solutions reveal the mathematical richness of GR, several lack clear physical interpretation; others, however, play central roles in astrophysics and cosmology. Well-known examples include the Schwarzschild solution for the exterior of spherical masses [107] and the Kerr solution for rotating compact objects [68], as well as interior stellar models such as the Tolman and Buchdahl families used to describe neutron stars. In cosmology, more general inhomogeneous or anisotropic solutions exist, such as the Lemaître–Tolman–Bondi models, Szekeres solutions, and Bianchi cosmologies [62]. However, the metric obtained by adopting a symmetric ansatz based on homogeneity and isotropy, Friedmann–Lemaître–Robertson–Walker (FLRW) metric, is a simple, physically meaningful model and also consistent with large-scale matter distribution in the Universe. Whether the Universe is expanding, contracting, or static depends on the physical properties of its matter and energy content. The FLRW metric in polar coordinates (r, θ, ϕ) is,

$$ds^2 = -c^2dt^2 + a^2(t)\left[\frac{dr^2}{1 - kr^2/R_0^2} + r^2 d\Omega^2\right] \tag{1.2}$$

where r is the comoving radial coordinate, $d\Omega^2 = d\theta^2 + \sin^2\theta d\phi^2$, R_0 is the curvature scale. $a(t)$ is the arbitrary function called scale factor that describes the expansion of the Universe and it tells us how physical separations (r_{phy}) are growing with time.

$$r_{\mathrm{phy}} = a(t)r \tag{1.3}$$

k is the curvature parameter corresponding to three discrete possibilities of spatial geometry ($k = -1, 0, +1$ for open, flat, and closed Universe, respectively). The comoving radial component can be redefined as a parameter χ and obtained by integrating the following equation,

$$d\chi \equiv \frac{dr}{\sqrt{1 - kr^2/R_0^2}} \tag{1.4}$$

The line element becomes,

$$ds^2 = -c^2dt^2 + a^2(t)\left[d\chi^2 + S_K^2(\chi)d\Omega^2\right], \tag{1.5}$$

$$S_k(\chi) \equiv \begin{cases} R_0 \sinh(\chi/R_0), & k = -1, \\ \chi, & k = 0, \\ R_0 \sin(\chi/R_0), & k = +1. \end{cases} \tag{1.6}$$

There is no difference between the comoving radial coordinate r and χ for flat Universe. Observations indicate that the Universe is very close to spatially flat ($k \approx 0$). This conclusion is strongly supported by the measurements of the cosmic microwave background (CMB), baryon acoustic oscillations (BAO), and LSS [36].

1.1.1.1 Friedmann equations

Having introduced the FLRW metric as the standard form for a homogeneous and isotropic Universe, we can now proceed to determine the time evolution of the scale factor $a(t)$, which is the unknown function in Eq.1.2. FLRW metric determines the left-hand side of the field equations, while the energy-momentum tensor is obtained by specifying the matter-energy content in the Universe. The homogeneous and isotropic assumption implies that the $T_{\mu\nu}$ of the total matter content of necessity has to have perfect fluid form (i.e., a fluid without any dissipation) [44]. Consider the Universe with both vacuum energy (energy density Λ) and a perfect fluid [118]. The cosmological constant is conventionally treated as part of spacetime geometry rather than matter. EFEs become,

$$G_{\mu\nu} + \Lambda g_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu} \tag{1.7}$$

However, since the term $\Lambda g_{\mu\nu}$ can be mathematically moved to the matter side of Einstein's equations, it is equivalent to interpret it as a perfect fluid with constant energy density and negative pressure. Substituting the FLRW metric and a perfect-fluid energy-momentum tensor into Eq.1.7 yields two independent equations for $a(t)$, the Friedmann equations. The temporal part of EFEs for the FLRW metric leads to the first Friedmann equation, which gives the evolution of the scale factor.

$$G_0^0 = \frac{8\pi G}{c^4} T_0^0$$

Then,

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{kc^2}{a^2} + \frac{\Lambda c^2}{3} \tag{1.8}$$

ρ is the sum of all contributions to the energy density of the Universe, and Λ is called the cosmological constant. The spatial component of EFEs lead to the second Friedmann equation, known as acceleration equation, which is expressed as follows.

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\rho + \frac{3P}{c^2}\right) + \frac{\Lambda c^2}{3} \tag{1.9}$$

Here P is the pressure of contents in the Universe. This equation determines whether the Universe's expansion is accelerating or decelerating. The first Friedmann Eq.(1.8) can be written in terms of Hubble parameter, $H = \frac{\dot{a}}{a}$,

$$H^2 = \frac{8\pi G}{3}\rho - \frac{kc^2}{a^2} + \frac{\Lambda c^2}{3} \tag{1.10}$$

The critical density is defined as,

$$\rho_c = \frac{3H_0^2}{8\pi G} \tag{1.11}$$

H_0 is the present time Hubble parameter value and is referred to as Hubble's constant. In summary, the first Friedmann equation is the "expansion law" of the Universe. It tells us the value of the Hubble parameter as a function of the cosmic energy content. The second Friedmann equation is the "acceleration law". It tells us whether the expansion of the Universe accelerates or decelerates.

1.2 Standard model of cosmology

Finding the specific solution to the Friedmann equations requires the imposition of initial conditions, such as the scale factor $a(t_0)$ and the Hubble parameter $H(t_0) = H_0$, in order to uniquely the integration constant. Here, t_0 is the present time. In addition to the initial conditions, solving these differential equations requires knowledge of the energy content of the Universe, the curvature parameter k, and the cosmological constant Λ.

From a cosmological perspective, components are most usefully distinguished by their equation of state and clustering behaviour (relativistic vs non-relativistic), rather than by particle identity alone [102]. These include relativistic particles like photons and neutrinos (collectively referred to as radiation), baryonic matter (baryons and leptons), cold dark matter (CDM) being responsible for structure formation, and the dark energy or cosmological constant Λ. This phenomenological model of the Universe is called the ΛCDM model. It is convenient to express the energy densities of all components in relation to the critical density and define the following dimensionless density parameters.

$$\Omega_r = \frac{\rho_r}{\rho_c}, \quad \Omega_m = \frac{\rho_m}{\rho_c}, \quad \Omega_\Lambda = \frac{\rho_\Lambda}{\rho_c} = \frac{\Lambda c^2}{3H_0^2} \tag{1.12}$$

where ρ_r, ρ_m and ρ_Λ are the energy densities of radiation, matter, and vacuum energy, respectively. Here, matter represents both ordinary and CDM. The Friedmann equation (Eq.1.10) can be written as,

$$\frac{H^2}{H_0^2} = \Omega_r a^{-4} + \Omega_m a^{-3} + \Omega_k a^{-2} + \Omega_\Lambda \tag{1.13}$$

Here, Ω_k is the curvature density parameter $\Omega_k \equiv -kc^2/(R_0 H_0)^2$. Evaluating both sides of the Friedmann equations with the initial condition, $a(t_0) \equiv 1$, results in the following constraint.

$$1 = \Omega_r + \Omega_m + \Omega_\Lambda + \Omega_k = \Omega_0 + \Omega_k \tag{1.14}$$

In its minimal formulation, the ΛCDM model is fully specified by the following cosmological parameters. The parameters, $(\Omega_m, \Omega_b, \Omega_\Lambda, H_0)$, describe the evolution of the homogeneous background Universe, which is assumed to be a spatially flat FLRW spacetime. Two additional parameters characterise the initial perturbations (namely, the amplitude (A_s) and the spectral index (n_s) of the primordial power spectrum, (see Section 1.4.2), and one further parameter called the reionisation parameter which captures the integrated Thomson scattering opacity from reionisation and affects the large-scale CMB polarisation [36].

1.3 Redshift

The large-scale structure of the Universe consists of dark matter (DM) and baryons arranged in a cosmic web of filaments, clusters, and voids. As DM is not directly observable, this structure is traced by luminous objects, primarily galaxies, which are gravitationally bound systems of stars, gas, dust, and DM. These galaxies exhibit a range of absorption and emission lines in their spectra, each with well-known characteristic frequencies. When a galaxy is moving away from us, the characteristic lines in its spectrum shift towards the longer wavelengths and are referred to as redshift (z). Conversely, if a galaxy moves towards us, the frequency of the light waves increases, a phenomenon known as blueshift. These frequency shifts were used by Vesto Slipher in 1912 to measure the observed galaxies' apparent radial velocity [110] and then systematically applied by Edwin Hubble [59] to determine the galaxy distances. The shift in the frequency of light signals can arise from several physical effects, including the Doppler effect, gravitational redshift, and the expansion of the Universe. Among these, the dominant contribution to the observed redshift on cosmological scales is due to the expansion of the Universe. In practice, the observed redshift combines the cosmological expansion and (typically smaller) contributions from peculiar velocities and gravitational potentials; in the low-velocity limit one can write $1 + z_{\rm obs} \simeq (1 + z_{\rm cos})(1 + z_{\rm pec})$. Here, the focus is on $z_{\rm cos}$ from cosmic expansion.

To compute this frequency shift, we can adopt the FLRW metric and consider a light ray propagating towards the observer along a radial direction $d\theta = d\phi = 0$. This consideration of light propagation does not affect the final expression of redshift. Consider an observer at comoving coordinate $(t_{obs}, 0, 0, 0)$, observing the light signals from a distant source in the Universe at comoving radial coordinate r_{source}. The emission of crests of a light wave occurs at coordinate $(t_{em}, r_{source}, 0, 0)$. Light travels on a null geodesic, $ds^2 = 0$. Therefore, the equation of motion for the crests of a light wave is,

$$c^2dt^2 - a(t)^2\frac{dr^2}{1-kr^2} = 0 \tag{1.15}$$

Integrating along the null geodesic taken by the light wave,

$$\int_{t_{em}}^{t_{obs}} \frac{cdt}{a(t)} = \int_0^{r_{source}} \frac{dr}{(1-kr^2)^{1/2}} \tag{1.16}$$

Imagine another light wave crest emitted from the same source at time $t_{em} + \delta t_{em}$ and observed at $t_{obs} + \delta t_{obs}$.

$$\int_{t_{em}+\delta t_{em}}^{t_{obs}+\delta t_{obs}} \frac{cdt}{a(t)} = \int_0^{r_{source}} \frac{dr}{(1-kr^2)^{1/2}} \tag{1.17}$$

The integral over the radial coordinate on the right-hand side of Eq.1.16 and 1.17 is identical. We can equate the left-hand-side of these equations and get,

$$\int_{t_{em}}^{t_{obs}} \frac{cdt}{a(t)} = \int_{t_{em}+\delta t_{em}}^{t_{obs}+\delta t_{obs}} \frac{cdt}{a(t)} \tag{1.18}$$

The time interval between the emission of the two light wave crests is assumed to be negligible on cosmological timescales. And the integrands are approximately constant over the integration limits. As a result, the integrals can be approximated by the following expression.

$$\frac{\delta t_{em}}{a(t_{em})} = \frac{\delta t_{obs}}{a(t_{obs})} \tag{1.19}$$

δt denotes the proper time interval between successive oscillations of the light wave (i.e., the separation between wave crests, with $\delta t_{\rm em}$ at emission and $\delta t_{\rm obs}$ at observation), For comoving observers in an FLRW spacetime, the cosmic time coordinate t coincides with the observer's proper time, so δt can be interpreted as the proper-time separation between successive wave crests. this ratio can be expressed in terms of the corresponding frequencies,

$$\frac{f_{obs}}{f_{em}} = \frac{a(t_{em})}{a(t_{obs})} \tag{1.20}$$

Frequency shift depends only on the scale factor $a(t)$ at the time of emission and of observation. Frequency (f) of the light wave is inversely proportional to the scale factor

implying that its wavelength (λ) is proportional to the scale factor since

$$f = \frac{c}{\lambda}$$

Therefore,the redshift parameter can be written as,

$$z = \frac{\lambda_{obs} - \lambda_{em}}{\lambda_{em}} \tag{1.21}$$

$$1 + z = \frac{\lambda_{obs}}{\lambda_{em}} = \frac{a(t_{obs})}{a(t_{em})} \tag{1.22}$$

where λ_{obs} is the observed and λ_{em} is the rest-frame/emitted wavelength of light, respectively. If t_{obs} is present time then conventionally $a(t_{obs}) = 1$. The relation between scale factor at the time of light emission and redshift is,

$$a(t_{em}) = \frac{1}{1 + z} \tag{1.23}$$

This equation implies that, when we observe a galaxy at redshift $z = 1$, the corresponding scale factor of the Universe at the time the light was emitted is $a(t_{em}) = 0.5$. This means that the distances between fundamental observers (or galaxies) were half of their current values [81].

We can also define the comoving distance $\chi(z)$ as the radial distance travelled by light from redshift z to the observer, accounting for the expansion of the Universe, and is given by

$$\chi(z) = \int_0^z \frac{cdz'}{H(z')}. \tag{1.24}$$

In summary, redshift is central to observational cosmology, as it provides a direct mapping between the scale factor of the Universe and the wavelength of light we observe today. Precise redshift measurements are therefore indispensable for calibration and precision cosmology, while photometric redshift estimates enable redshift information for the much larger samples delivered by wide-field imaging surveys. The following Chapters (3 & 4) discuss the redshift estimation methodology and its importance in advancing cosmological studies.

1.3.1 Hubble-Lemaître's law

Hubble [59] showed that the apparent recession velocity for $z \ll 1$ ($v \approx cz$) is proportional to the distance of that object from us (d). This is known as Hubble-Lemaître's law.

$$cz \approx H_0 d \tag{1.25}$$

This linear form is valid only at low redshift ($z \ll 1$), where different cosmological distance measures become approximately equal; here d denotes the (approximately) proper

distance to the source. We can write the empirical formula for low-redshift as,

$$z = \frac{d}{D_H} \tag{1.26}$$

If the recession velocity units are km s^{-1} and the distance units are Megaparsec (Mpc) then the H_0 unit is km s^{-1} Mpc^{-1}. The dimension of H_0 is $[s^{-1}]$. Therefore, the inverse of H_0 is called the Hubble time, $t_H \equiv 1/H_0$ that gives an estimate of the age of the Universe. Hubble distance is defined as

$$D_H = \frac{c}{H_0} \tag{1.27}$$

The dimensionless form of the Hubble constant is usually denoted as h.

$$h = \frac{H_0}{100 \text{ km s}^{-1} \text{ Mpc}^{-1}} \tag{1.28}$$

The Planck [99] results based on CMB anisotropies within the ΛCDM model yield $H_0 = 67.4 \pm 0.5$ km s^{-1} Mpc^{-1} whereas the *SH0ES* collaboration, using Type Ia supernovae calibrated with Cepheid variables, reports a significantly higher value of $H_0 = 73.04 \pm 1.04$ km s^{-1} Mpc^{-1} [101]. Other independent late-time probes (e.g. TRGB [45], time-delay lenses [14]) infer different estimates for H_0. This significant discrepancy between measurements of the Hubble constant obtained from observations of the early- and late-Universe is known as Hubble tension.

1.4 Theory of structure formation

This section briefly summarises the standard linear theory of density perturbation growth in an expanding background, which provides the conceptual basis for the later methodology. The diversity of structures such as stars, galaxies, and galaxy clusters that we observe in the Universe today is absent in the Friedmann model. To develop more realistic models of the Universe, we include small density perturbations into the homogeneous, isotropic models and study their development under gravity [81]. The standard paradigm is that LSS grows by gravitational instability from small primordial density perturbations [97, 36]. Two key aspects of the understanding of structure formation are (i) the origin and statistical properties of the initial density fluctuations, and (ii) the time evolution of cosmological density perturbations. As long as these perturbations remain relatively small**,** we can treat them in perturbation theory.

Although GR provides the fundamental description of cosmological perturbations, the Newtonian approach is adequate for the regime relevant to structure formation in the late Universe. In the GR framework, perturbations are introduced by expanding the EFEs around the homogeneous and isotropic FLRW background, and linearising the Einstein tensor accordingly. This yields the relativistic perturbation equations, which are

the starting point for a fully consistent treatment of metric and matter fluctuations. For more details, see [36, 82]. In this thesis, the Newtonian formulation is used as an accurate approximation for sub-horizon modes in the late-time, non-relativistic matter-dominated regime, where gauge subtleties are not essential for the quantities considered.

1.4.1 Newtonian perturbation theory

The perturbed matter density field can be written as

$$\rho(\mathbf{x}, \tau) = \bar{\rho}(\tau) + \delta\rho(\mathbf{x}, \tau) \tag{1.29}$$

where τ is the conformal time, related to the cosmic time t by

$$d\tau = \frac{dt}{a} \tag{1.30}$$

Here, $\bar{\rho}(\tau)$ denotes the mean matter density of the homogeneous background, while $\delta\rho(\mathbf{x}, \tau)$ represents the perturbation at comoving position $\mathbf{x}$ and time τ. A dimensionless density field, known as the density contrast, is defined in real space as

$$\delta(\mathbf{x}, \tau) \equiv \frac{\delta\rho(\mathbf{x}, \tau)}{\bar{\rho}(\tau)} \tag{1.31}$$

Positive values of $\delta(\mathbf{x}, \tau)$ correspond to overdense regions associated with matter clustering, while $\delta(\mathbf{x}, \tau) < 0$ describes underdense regions, leading to the formation of cosmic voids.

Any general field can be represented as a superposition of modes. In flat comoving space, plane waves form the natural basis, allowing a Fourier transform representation. In curved spaces, the appropriate basis is given by the eigenfunctions of the Laplacian (wave equation). The Dirichlet conditions constitute sufficient, though not necessary, criteria for the existence of a well-defined Fourier transform. In cosmology, the density contrast field $\delta(\mathbf{x})$ does not strictly satisfy these conditions on an unbounded domain. To circumvent this, one introduces δ within a finite cubic volume subject to periodic boundary conditions and subsequently takes the limit as the box size tends to infinity. In this way, although δ does not fulfil the Dirichlet conditions, Fourier analysis of the field remains mathematically consistent and physically meaningful [96]. The Fourier transform of the density contrast field is,

$$\delta(\mathbf{k}, \tau) = \int \delta(\mathbf{x}, \tau) e^{i\mathbf{k}\cdot\mathbf{x}} d^3x \tag{1.32}$$

and

$$\delta(\mathbf{x}, \tau) = \frac{1}{(2\pi)^3} \int \delta(\mathbf{k}, \tau) e^{-i\mathbf{k}\cdot\mathbf{x}} d^3k$$

Therefore,

$$\delta(\mathbf{k},\tau) = \frac{\delta\rho(\mathbf{k},\tau)}{\bar{\rho}(\tau)} \tag{1.33}$$

On scales well inside the Hubble radius (wavenumber of Fourier mode, $k \gg aH/c$), and for non-relativistic matter (velocity $\ll c$) with weak gravitational potentials ($|\Phi| \ll 1$), the relativistic scalar perturbation equations (see Appendix A) reduce to their Newtonian counterparts. This Newtonian limit applies to sub-horizon modes relevant for late-time LSS, for non-relativistic matter and weak potentials. In this limit, the perturbed 00-component of the linearised Einstein equations for scalar perturbations gives the Poisson equation for the Newtonian potential. This relates the gravitational potential with the matter density perturbation by

$$k^2\Phi = 4\pi G a^2\, \delta\rho(\mathbf{k},\tau) \tag{1.34}$$

This is the Poisson equation in Fourier space. Using Eq. 1.33, it can be written in terms of density contrast as,

$$k^2\Phi = 4\pi G a^2\, \bar{\rho}(\tau)\, \delta(\mathbf{k},\tau) \tag{1.35}$$

The energy-momentum content of the Universe can be described using the fluid approximation. More precisely, the fluid approximation is an effective description after coarse-graining on sufficiently large-scales and in the single-stream regime (before significant shell-crossing), where collisionless CDM can be treated as a pressureless fluid. In the sub-horizon, matter-dominated regime, we neglect radiation and (on the large scales of interest) consider only non-relativistic matter, namely cold dark matter and baryons. The fluid description is valid for baryons and CDM; baryons have a mean free path that is much smaller than the scale of interest, while CDM is a pressureless matter [89]. The energy-momentum tensor for the perfect fluid is,

$$T^{\mu\nu} = (\rho + P)u^\mu u^\nu + P g^{\mu\nu} \tag{1.36}$$

Where pressure P and energy density ρ at a given point are defined to be the pressure and energy density measured by a comoving observer at rest with the fluid at the instant of measurements. u^μ is the four-velocity of the fluid [82].

While the Poisson equation relates the Newtonian gravitational potential to the density perturbation, the dynamical law for $\delta(\mathbf{k},\tau)$ comes from the local conservation of energy–momentum. From the Bianchi identity we will get,

$$\nabla_\mu G^{\mu\nu} = 0$$

Therefore,

$$\nabla_\mu T^{\mu\nu} = 0$$

This gives the local conservation of energy and momentum. We obtain two independent sets of equations. The temporal component ($\nu = 0$) corresponds to energy conservation and governs the evolution of the density perturbations. By analogy with fluid dynamics, this is referred to as the continuity equation. The spatial components ($\nu = i$) correspond to momentum conservation and describe the evolution of the velocity perturbations; these are commonly called the Euler equations in cosmology (Section 4 in [82] provides more details). In the Newtonian and matter-dominated regime, both reduce to the continuity and Euler equations of classical fluid mechanics. The continuity equation in Fourier space of spatial coordinates is,

$$\dot{\delta}(\mathbf{k}, \tau) = -\theta \tag{1.37}$$

Where the derivatives are taken with respect to conformal time τ and θ is the divergence of the flowing fluid with coordinate velocity $\mathbf{v} \equiv d\mathbf{x}/d\tau$.

$$\theta(\mathbf{k}, \tau) = i\,\mathbf{k} \cdot \mathbf{v}(\mathbf{k}, \tau)$$

The continuity equation expresses mass conservation by relating the time evolution of the density contrast $\delta(\mathbf{k}, \tau)$ to the velocity divergence θ. If $\theta < 0$, corresponding to a converging flow, the overdensity grows ($\dot{\delta}(\mathbf{k}, \tau) > 0$), whereas if $\theta > 0$, corresponding to an expanding flow, the overdensity decreases. In other words, the density changes because matter flows into or out of a given region. Euler equation in Fourier space is,

$$\dot{\theta} = -\frac{\dot{a}}{a}\theta + k^2 c_s^2 \delta(\mathbf{k}, \tau) - k^2 \Phi \tag{1.38}$$

c_s is the sound speed of the fluid [82]. This term comes from the definition of pressure perturbation,

$$\delta P \equiv c_s^2 \delta\rho(\mathbf{k}, \tau)$$

Combining Eq.1.35, 1.37, and 1.38, we will get the linear evolution equation for the density contrast, $\delta(\mathbf{k}, \tau)$, in the matter-dominated era. The pressure term becomes negligible and the relevant equation can be written without the sound speed term,

$$\ddot{\delta}(\mathbf{k}, \tau) + \frac{\dot{a}}{a}\dot{\delta}(\mathbf{k}, \tau) - 4\pi G a^2 \bar{\rho}\delta(\mathbf{k}, \tau) = 0 \tag{1.39}$$

This is called the growth equation. The equation is written in Fourier space over the spatial coordinates and all derivatives are taken with respect to the conformal time τ. The term $\frac{\dot{a}}{a}\dot{\delta}$ is from the expansion of the Universe and is known as Hubble drag or Hubble friction. In the pressureless linear regime, the growth equation has no explicit k-dependence, so the growth factor is scale-independent and each mode evolves with the same time dependence (up to its initial amplitude). In a flat, matter-dominated Universe,

this equation can be written as

$$\ddot{\delta}(\mathbf{k},\tau) + \frac{\dot{a}}{a}\dot{\delta}(\mathbf{k},\tau) - \frac{3}{2}\left(\frac{\dot{a}}{a}\right)^2 \delta(\mathbf{k},\tau) = 0$$

Now, the growth equation takes the form of a Cauchy-Euler ordinary differential equation in conformal time τ. Assuming a power-law ansatz $\delta(\tau) = \tau^\alpha$ yields the algebraic condition $\alpha = 2, -3$, giving two independent solutions $\delta \propto \tau^2$ (growing) and $\delta \propto \tau^{-3}$ (decaying). Transforming to cosmic time t using $a(t) \propto t^{2/3}$ and $\tau \propto t^{1/3}$. We can write the general solution as

$$\delta(\mathbf{k},t) = \delta_+(\mathbf{k})D_+(t) + \delta_-(\mathbf{k})D_-(t)$$

where $\delta_\pm(\mathbf{k})$ are the Fourier amplitudes of the initial density field and $D_\pm(t)$ are $\mathbf{k}$-independent functions. The functions $D_+(t)$ and $D_-(t)$ correspond to the growing and decaying modes, respectively, with $D_+(t)$ known as the *linear growth factor* since it governs the growth of structure in the linear regime ($\delta < 1$).

In a Universe with both matter and dark energy, the expansion of the Universe at late times is accelerated by the dark energy. As a result, the gravitational pull from matter is effectively weaker relative to the expansion, and the growth of density perturbations slows down compared to the matter-dominated case. While the general behaviour with a growing and a decaying mode still exists, the growing mode no longer increases proportionally to the scale factor; instead, its growth gradually slows when dark energy dominates. The decaying mode remains negligible, so the formation of structure in the late Universe is primarily governed by the suppressed growing mode. Despite this slowdown, the growth remains approximately the same on all scales in the linear regime [8].

1.4.2 Linear matter power spectrum

To uniquely specify the solution for the growth equation, initial conditions must be set at an early time t_i when perturbations are super-horizon. These are set by the primordial curvature perturbation $\mathcal{R}(\mathbf{k})$ generated during inflation. On these scales,

$$\delta(\mathbf{k},t_i) = \delta_i(\mathbf{k}), \tag{1.40}$$

$$\dot{\delta}(\mathbf{k},t_i) \approx 0. \tag{1.41}$$

Inflation predicts that the initial fluctuations $\delta_i(\mathbf{k})$ form a Gaussian random field, fully characterised by its power spectrum. The primordial power spectrum of density perturbation, $P_\mathcal{R}(k)$ is

$$\langle \delta_i(\mathbf{k})\delta_i^*(\mathbf{k}')\rangle = (2\pi)^3\delta_D^{(3)}(\mathbf{k}-\mathbf{k}')P_\mathcal{R}(k) \tag{1.42}$$

$\delta_D(\mathbf{k} - \mathbf{k}')$ is the Dirac-delta function. A nearly scale-invariant spectrum, a generic prediction of simple inflationary models, is parameterised as a power law:

$$P_{\mathcal{R}}(k) = A_s \left(\frac{k}{k_*}\right)^{n_s - 1} \tag{1.43}$$

where A_s is the amplitude at the pivot scale k_*, and n_s is the spectral index. Measurements from the CMB [99] find $n_s \approx 0.96$, close to, but not exactly, scale-invariance ($n_s = 1$).

The late-time linear matter power spectrum, $P_{\rm lin}(k, t)$, is then the product of this primordial initial condition and the square of the transfer function and the growth factor, which encode the subsequent linear evolution.

$$P_{\rm lin}(k, t) = P_{\mathcal{R}}(k)\, T(k)^2\, D_+(t)^2 \tag{1.44}$$

$T(k)$ captures the scale-dependent processing of primordial perturbations through horizon entry and the radiation-to-matter transition (and possibly baryon features), while $D_+(t)$ is the scale-independent growth factor in the linear, pressureless regime, usually normalised to $D_+(t_0) = 1$ today. $T(k)$ is defined such that $T(k \to 0) = 1$ at late times

In summary, the initial conditions used here for solving the growth equation are the primordial density fluctuations laid down by inflation. The adiabatic nature of these fluctuations motivate a convenient super-horizon initial condition $\dot{\delta} \approx 0$, which selectively picks out the growing mode as the physically relevant solution. The statistical properties of these initial seeds, embodied in the primordial power spectrum $P_{\mathcal{R}}(k)$, are the source of structures in the Universe [36].

1.4.3 Non-linear matter power spectrum

While linear perturbation theory accurately describes the evolution of the matter power spectrum on large-scales, in practice, this corresponds roughly to wavenumbers $k \lesssim 0.1\, h\,\mathrm{Mpc}^{-1}$ at low redshift, with the precise boundary being mildly redshift dependent [21]. Linear theory breaks down once density fluctuations grow to order unity. In this non-linear regime, mode coupling transfers power from large-scales to small scales, producing enhanced clustering within structures such as dark matter (DM) halos. DM halos are gravitationally bound, virialised structures that form through the nonlinear gravitational collapse of DM overdensities. Because density perturbations grow larger on smaller scales, collapse occurs first on small scales, with the resulting structures subsequently merging and assembling hierarchically into more massive halos [36]. As a result, the matter power spectrum acquires non-linear corrections that cannot be captured by the linear growth factor alone.

A widely used prescription for modelling these effects is the `Halofit` model [111, 116]. In our analysis, we compute the `Halofit` corrections using both CLASS [21] and

the Core Cosmology Library (CCL; [22]) in Chapter 4. It is an empirical fitting formula inspired by halo-model behaviour. The halo model is a framework in which all matter in the Universe is assumed to reside in DM halos, with each particle belonging to a single halo. Under this assumption, the full matter density field can be constructed from the distribution of halos and a model for their internal structure [36]. However, `Halofit` does not constitute a first-principles halo-model calculation in which the large- and small-scale contributions are derived from physically computed components. Here, the two-halo term P_{2h}, representing correlations between separate halos, and a one-halo term P_{1h}, describing correlations of particles within the same halo. Therefore, non-linear power spectrum can be written as

$$P_{\rm NL}(k) = P_{2h}(k) + P_{1h}(k). \tag{1.45}$$

On large-scales, $P_{2h}(k)$ dominates and reduces to the linear power spectrum, while on small-scales the one-halo term becomes the leading contribution, capturing the internal halo structure.

Although `Halofit` is not a first-principles physical model, its empirical calibration to N-body simulations yields percent-level accuracy for ΛCDM cosmologies. The revised fitting formula provides 5%-level accuracy in the non-linear matter power spectrum over the range $k \lesssim 1h\text{Mpc}^{-1}$ at redshifts $0 \leq z \leq 10$ and 10% accuracy for $1 \leq k \leq 10h\text{Mpc}^{-1}$ at $0 \leq z \leq 3$ [116]. The accuracy degrades at higher k, higher redshift, and for non-standard cosmologies (e.g., models with massive neutrinos, evolving dark energy, or modified gravity). More recent extensions incorporate baryonic feedback, massive neutrinos, and modified gravity effects; however, `Halofit` remains the foundational baseline for nonlinear power spectrum modelling used in contemporary cosmological surveys.

1.5 Clustering statistics

This section introduces the statistical framework used to quantify the LSS. This section begins with the two-point statistics of the matter overdensity field and presents the correlation function. It will then extend the discussion to discrete tracers, introducing galaxy correlation functions, estimators and linear bias. Finally, the section addresses the angular two-point correlation function and its Limber projection. It concludes with Count-in-Cells statistics and higher-order moments, which are used in subsequent chapters.

The linear growth equation (Equation 1.39) describes the evolution of the density contrast field for each Fourier mode in the matter-dominated era, providing a theoretical framework for structure formation. However, cosmological observations do not give direct access to the density contrast. The Universe's matter content is dominated by dark matter, with baryonic matter contributing as well, which cannot be fully mapped directly. Instead, observations probe the large-scale matter distribution indirectly through tracers

such as galaxies and quasars [36]. Quasars (quasi-stellar objects, QSOs) are among the most luminous and energetic objects in the Universe. A more detailed discussion of quasars is provided in Chapter 4.

This section introduces how observations of galaxy clustering can be used to test models of structure formation. Different statistical descriptors probe distinct aspects of large-scale structure formation. Measures such as the two-point correlation function [97], cell-count variance, and the power spectrum are directly related to the fluctuation spectrum, allowing constraints on its amplitude and shape. Other approaches, including percolation analysis [126, 41] and topological measures such as Minkowski functionals [86], focus on the morphology of the LSS distribution, revealing features of the cosmic web such as sheets, filaments, or bubbles that may arise from nonlinear evolution or non-Gaussian perturbations. Higher-order correlation functions [97] provide insight into the role of self-similarity in structure formation. Together, these complementary methods constrain different aspects of cosmological structure-formation models [24]. Redshift surveys measure the tracer sky positions and precise redshifts. However, large imaging surveys identify the sky positions of many millions of galaxies without the precise redshift information. From these data, one can construct tracer density fields and estimate corresponding power spectra [36], allowing a comparison between theoretical predictions and observational measurements of matter clustering.

1.5.1 Clustering of matter overdensity field

In the early Universe, density fluctuations arose from quantum processes and evolved under stochastic influences. These microscopic quantum fluctuations, stretched to macroscopic scales during inflation, generated the primordial density field, which provides the initial conditions for the subsequent evolution of $\delta(\mathbf{x})$. These fluctuations result from the superposition of many independent processes. Primordial fluctuations are commonly modelled as a Gaussian random field and are consistent with the primordial non-Gaussianity constraints from CMB observations [98].

In cosmology, $\delta(\mathbf{x})$ of matter distribution is modelled as a continuous random variable. Consequently, the probability of observing any exact value at a single point is zero. Meaningful probabilistic statements can therefore only be made over intervals, using the probability density function (PDF). Within this framework, statistical descriptors such as the mean, variance, and two-point correlation function are defined as integrals over the PDF or joint PDF. The mean and variance correspond to integrals over the full distribution of δ, while the two-point correlation function (2PCF) is computed as the expectation value of the product of δ at two points. In this way, all relevant statistical measures of the density field arise from sets of non-zero measure rather than from exact point values, providing a rigorous foundation for describing the stochastic nature of cosmic structure.

1.5.1.1 Two-point correlation function of matter density field

Density perturbations in a statistically homogeneous and isotropic Universe are described by the statistical moments of the density contrast field, such as the variance of the over-density field. These moments do not depend on location or orientation. Since the density contrast is a random variable drawn from an underlying probability distribution, the mean of the distribution is the expectation value of this random variable. Statistical homogeneity implies that these means at each location are identical.

$$\langle \delta_i \rangle = \bar{\delta}, \quad \forall i \in \{1, \ldots, N\}.$$

Using the ergodic hypothesis, we can say that the mean value of the distribution can be computed by the ensemble average of the values of δ across the spatial field. If we take the ensemble (or spatial) average of $\delta(\mathbf{x})$:

$$\langle \delta(\mathbf{x}) \rangle = \left\langle \frac{\rho(\mathbf{x}) - \bar{\rho}}{\bar{\rho}} \right\rangle$$

Since $\bar{\rho}$ is constant, it can be taken outside the average:

$$\langle \delta(\mathbf{x}) \rangle = \frac{\langle \rho(\mathbf{x}) \rangle - \bar{\rho}}{\bar{\rho}}$$

By the definition of the mean density:

$$\langle \rho(\mathbf{x}) \rangle = \bar{\rho}.$$

The density contrast δ has zero mean. Therefore, the amplitude of the cosmological perturbation is given by the variance of the integral over the PDF of δ. The larger variance allows the possibility of producing realisations with larger values of δ. The covariance between two random variables (here, δ) at $\mathbf{x}$ and at $\mathbf{x}'$ depends only on the distance between them because of statistical homogeneity and isotropy [24], and is given by,

$$\xi_m(\mathbf{r}) = \xi_m(r) = \langle \delta(\mathbf{x}) \delta(\mathbf{x}') \rangle \tag{1.46}$$

where $r = |\mathbf{x} - \mathbf{x}'|$ and the average is taken over all spatial locations. Equation 1.46 is called the two-point correlation function (2PCF). It quantifies how the overdensity at one location is statistically related to that at another. Observations and simulations show that $\xi_m(r)$ is positive at small scales ($r \lesssim$ 10-20 Mpc), reflecting the clustering of matter into halos and filaments. At intermediate scales ($r \sim$ 20-100 Mpc), the correlation decreases and can become slightly negative, while at large-scales ($r \gtrsim 100$ Mpc) $\xi_m(r) \to 0$, indicating that density fluctuations become uncorrelated. Therefore, the 2PCF of the density contrast is not constant and positive at all scales; instead, it decreases with

separation and approaches zero at sufficiently large distances, consistent with the observed homogeneity of the Universe.

Representing the two-point correlation function in Fourier space, through the power spectrum $P(k)$, offers several advantages. For a Gaussian random field, different Fourier modes are statistically independent, which simplifies theoretical calculations compared to the real-space 2PCF, where correlations at different separations are not independent. The power spectrum also provides a direct decomposition of fluctuations by spatial scale, allowing us to identify which scales dominate structure formation. Many cosmological models predict $P(k)$ directly, making Fourier-space comparisons with observations more straightforward. Additionally, computationally efficient FFT methods enable fast estimation of $P(k)$ from large datasets. Features such as the BAO are more clearly identified in Fourier space, highlighting the connection between observational data and underlying physical processes. Thus, Fourier-space representation complements the real-space 2PCF by providing both theoretical clarity and practical advantages. The covariance of two Fourier modes is,

$$\langle \delta(\mathbf{k})\delta^*(\mathbf{k}')\rangle = (2\pi)^3\delta_D(\mathbf{k}-\mathbf{k}')P(\mathbf{k})$$

$P(\mathbf{k})$ is the variance of each mode as a function of $\mathbf{k}$ known as the power spectrum. By using the last two expressions, 2PCF can be written as,

$$\xi_m(r) = \langle \delta(\mathbf{x})\delta(\mathbf{x}+\mathbf{r})\rangle = \int \frac{d^3k}{(2\pi)^3}P(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{r}} \tag{1.47}$$

In an isotropic Universe, $\xi_m(r)$ does not depend on the orientation of r. Similarly, the power spectrum depends only on the magnitude $\mathbf{k}$ [96]. Correlation is real and therefore,

$$\mathrm{R}e(e^{-ikr\cos\theta}) = \cos(\mathrm{krcos}\theta)$$

In spherical coordinates,

$$d^3k = k^2\sin\theta dk d\theta d\phi$$

Power spectrum depends only on k and the final expression for the 2PCF in Fourier space is,

$$\xi_m(r) = \frac{1}{2\pi^2}\int_0^\infty P(k)\frac{\sin(kr)}{kr}k^2dk$$

This is the analytical expression for the 2PCF in cosmological perturbation theory. The factor $\sin(kr)/kr$ is the spherical Bessel function of first kind $j_0(kr)$. The modes of the power spectrum are statistically independent and directly show the clustering amplitude as a function of scale.

In cosmological analyses, the choice of matter power spectrum $P(k)$ depends fundamentally on the physical scales probed. On ultra-large scales ($k \lesssim 0.01h\mathrm{Mpc}^{-1}$), where perturbation wavelengths approach the Hubble horizon, the relativistic power spectrum

must be employed to properly capture general relativistic effects such as gravitational potential evolution and light cone effects. For large-scales ($0.01 \lesssim k \lesssim 0.1h\text{Mpc}^{-1}$), the linear Newtonian power spectrum adequately describes primordial fluctuations and baryon acoustic oscillations. At intermediate scales ($0.1 \lesssim k \lesssim 1h\text{Mpc}^{-1}$) relevant to quasar clustering presented in Chapter 4, non-linear corrections become important, necessitating semi-analytic models like `Halofit` or emulators that account for gravitational collapse. On highly non-linear scales ($k \gtrsim 1h\text{Mpc}^{-1}$), the power spectrum is dominated by halo profiles and requires specialised modelling incorporating baryonic feedback effects. This hierarchical approach ensures that the appropriate physical description is applied across all cosmological scales, from horizon-sized perturbations to virialised structures.

In contrast, the N values of the correlation function are themselves correlated and do not isolate information from individual scales since it is the integration over all k. Despite this, the two-point correlation function is still valuable because it is defined in real space, which means it can be measured directly from observational data on the sky.

1.5.2 Clustering of tracer overdensity field

We cannot observe the continuous matter density contrast field $\delta(x)$ directly. Assume that a discrete set of galaxies or quasars traces the three-dimensional distribution of overdensities. Under the assumptions of statistical homogeneity and isotropy (i.e., the underlying distribution can be described as a stationary and isotropic point process) and in the infinitesimal volume limit $dV \to 0$, the probability of finding an object within a volume element dV is proportional to $n\,dV$, where n denotes the mean number density. In this limit, the two-point correlation function depends only on the magnitude of the separation between points, $r = |\mathbf{r}|$, rather than on absolute position or direction.

The density contrast field of these tracers is modelled as a discrete point process. Within a survey volume V, we observe N_g galaxies, yielding a mean number density

$$n = \frac{N_g}{V}.$$

In practice, real surveys are affected by a spatially varying selection function and survey mask, so the observed number density is not strictly uniform. These effects are well accounted for in the analysis through the use of random catalogues such as organised randoms [65, 123] in the Kilo-Degree Survey (KiDS) that trace the survey geometry and selection function. The construction of the random catalogues used in our work is described in Chapter 4.

Consider two events; event A is the detection of *a galaxy* (any galaxy from the population) within an infinitesimal volume element dV_1 located at $\mathbf{x}_1$, and event B is the detection of a galaxy within an infinitesimal volume element dV_2 at $\mathbf{x}_2$. The correlation function and joint probability are always defined over finite volumes, not exact points.

Since galaxies are clustered, these two events are statistically correlated. Let $\mathbf{r}_{12} = \mathbf{x}_1 - \mathbf{x}_2$ denote the separation vector between the two positions with magnitude $r_{12} = |\mathbf{r}_{12}|$. The probability of finding a galaxy within dV_1 at $\mathbf{x}_1$ is

$$d\mathcal{P}_1 = n\, dV_1.$$

The joint probability of finding a galaxy in both dV_1 at $\mathbf{x}_1$ and dV_2 at $\mathbf{x}_2$ is given by

$$d\mathcal{P}_{12} = \mathcal{P}(A \cap B) = \mathcal{P}(A)\, \mathcal{P}(B|A),$$

where $\mathcal{P}(B|A)$ denotes the conditional probability of finding a galaxy at $\mathbf{x}_2$ given that there is already a galaxy at $\mathbf{x}_1$. This conditional probability can be written as

$$\mathcal{P}(B|A) = n\, [1 + \xi_g(r_{12})]\, dV_2,$$

where $\xi_g(r_{12})$ is the two-point correlation function. Due to statistical homogeneity and isotropy, ξ_g depends only on the magnitude of the separation vector $\mathbf{r}_{12}$, and not on its direction. If there is no clustering, the galaxies are distributed randomly and the correlation function vanishes, $\xi_g(r_{12}) = 0$. In this uncorrelated case, the joint probability reduces to

$$d\mathcal{P}_{12}^{\text{uncorr}} = n^2\, dV_1\, dV_2.$$

Comparing $d\mathcal{P}_{12}$ with dP_{12}^{uncorr} reveals the excess probability, i.e., the fractional increase in the joint probability due to clustering above the Poisson random catalogue baseline.

$$\Delta\mathcal{P} = d\mathcal{P}_{12} - d\mathcal{P}_{12}^{uncorr} = n^2 dV_1 dV_2 \xi_g(r_{12}) \tag{1.48}$$

Thus, we can interpret that the two-point correlation, $\xi_g(r_{12})$, is the excess probability per unit $n^2 dV_1 dV_2$. It quantifies the degree of clustering of galaxies as a function of their separation. In this context, the two-point correlation function describes the clustering properties of a discrete set of points [24].

1.5.3 Correlation function estimators and bias

In practice, we cannot evaluate the ensemble probabilities directly, since we only observe a single realisation of the Universe within a finite survey volume. Under the ergodic hypothesis, we replace ensemble averages with spatial averages and estimate probabilities through pair counts of galaxies. The number of tracer-tracer pairs with separation in a bin around r, denoted $DD(r)$, serves as a Monte Carlo estimator of the joint probability distribution, while a random catalogue with the same geometry provides the expected pair counts in the absence of clustering, $RR(r)$. Comparing these quantities yields an estimate of the fractional excess probability, i.e. the two-point correlation function, with

refined estimators such as Landy-Szalay combining $DD(r)$, $RR(r)$, and data-random pairs $DR(r)$ to minimise variance and correct for survey boundaries. The terms $DD(r), DR(r)$, and $RR(r)$ represent normalised pair counts rather than raw pair counts. The raw data pair count is divided by $N_D(N_D-1)/2$, random-random pairs by $N_R(N_R-1)/2$, and cross-pairs by $N_D N_R$. Here, N_D is the number of observed data points and N_R is the number of points in the random catalogue. This normalisation transforms simple counts into a ratio of probabilities The Landy-Szalay estimator [74] is defined as

$$\xi_g(r) = \frac{DD(r) - 2DR(r) + RR(r)}{RR(r)}. \tag{1.49}$$

The 2PCF of tracers is often well approximated by a power law,

$$\xi_g(r_{12}) \approx \left(\frac{r}{r_0}\right)^{-\gamma}$$

For low-redshift galaxy samples (approximately $0 \lesssim z \lesssim 0.2$), $r_0 \approx 5\ h^{-1}\,\mathrm{Mpc}$ and $\gamma \approx 1.8$ in the in the range $0.1\,\mathrm{h}^{-1}\,\mathrm{Mpc} \leq r \leq 10\,\mathrm{h}^{-1}\,\mathrm{Mpc}$ [24]. In cosmology, a tracer is any observable population of objects or signal that is used to map the underlying matter distribution in the Universe. This power-law behaviour is regarded as an empirical approximation rather than a fundamental prediction, and it holds only over a limited scale range. The best-fit parameters (r_0, γ) depend on tracer selection, luminosity, colour, redshift, and survey characteristics, and can differ significantly for other populations such as quasars or high-redshift galaxies.

Luminous objects, such as galaxies and quasars, are not direct tracers of the total matter in the Universe. This indicates that there is a *bias* between the spatial distribution of tracer $\delta_g(x)$ and total matter $\delta_m(x)$. In the linear approximation,

$$\delta_g(x) = b\delta_m(x) \tag{1.50}$$

Where b is called tracer bias [67, 35]. In terms of the two-point correlation function, the linear bias relation becomes the following,

$$\xi_g(r) = b^2\xi_m(r) \tag{1.51}$$

This equation corresponds to a deterministic, linear, and scale-independent bias approximation that is expected to be valid only on sufficiently large-scales and for a specified smoothing scale. On smaller scales bias can be scale-dependent and stochastic, but linear bias is used as a first baseline.

1.5.4 Angular two-point correlation function

The two-point correlation function introduced in the previous sections is defined within the full three-dimensional cosmological volume. Its evaluation requires accurate knowledge of galaxy or quasar positions together with their redshifts. Spectroscopic surveys such as the Sloan Digital Sky Survey (SDSS; [103]) and the Dark Energy Spectroscopic Instrument (DESI; [32]) provide this information by delivering angular coordinates (right ascension and declination) in combination with spectroscopic redshifts. In contrast, wide-field photometric surveys, including KiDS, Legacy Survey of Space and Time (LSST; [72]), and the Dark Energy Survey (DES, [34]), are capable of detecting the angular positions of millions of galaxies and quasars. However, without costly and time-consuming spectroscopic follow-up, their radial distances cannot be measured with high precision. Instead, redshifts are estimated from multi-band photometry, yielding a statistical redshift distribution for the photometric survey data. This information can still be used to probe the large-scale density field through correlation analyses, specifically, one can measure the angular two-point correlation function, $\omega(\vartheta)$ [97, 121]. This statistic quantifies the excess probability of finding galaxy pairs separated by an angle ϑ on the sky. Mathematically, the angular 2PCF is a two-dimensional correlation function obtained by projecting the three-dimensional correlation function onto the celestial sphere, typically by fixing angular coordinates and integrating along the line of sight.

Consider, two objects having 3D coordinates $\mathbf{x}$ and $\mathbf{x}'$ and $\mathbf{r}$ is the distance between them. $\xi_m(\mathbf{r})$ is the 3D correlation function. Let $\hat{\mathbf{n}}$ and $\hat{\mathbf{n}}'$ denote two directions on the sky with angular separation

$$\vartheta = \arccos\left(\hat{\mathbf{n}} \cdot \hat{\mathbf{n}}'\right)$$

The angular 2PCF of matter can be written as,

$$\omega_m(\vartheta) = \int d\chi W(\chi) \int d\chi' W(\chi') \xi_m(\mathbf{r})$$

χ is the comoving radial coordinate. In this context of clustering, the weight function $W(\chi)$ in the projection corresponds to the normalised redshift distribution $p(\chi)$ of density tracers such as galaxies or quasars.

$$p(\chi) \equiv \frac{1}{N} \frac{dN}{d\chi}$$

with

$$\int d\chi p(\chi) = 1$$

Here N is the total number of tracer objects. A higher value of the weight function at a given redshift z (or comoving distance χ) means that tracers at that redshift contribute more strongly to the integral. The 3D correlation function is a decreasing function of

separation, and the pairs that are physically close contribute. We can consider the pair with $\chi \sim \chi'$. In the Limber (large multipole moments ℓ) limit, the oscillatory projection kernel suppresses contributions from widely separated radial distances, so that the dominant contribution arises from $\chi \simeq \chi'$. The window function has a slow variation in redshift; $W(\chi) \sim W(\chi')$.

The 3D vector scale can be decomposed into 2D transversal and 1D longitudinal components. By using Equation 1.47, angular 2PCF can be written as,

$$\omega(\vartheta) = \int d\chi\, W^2(\chi) \int d\chi' \int \frac{dk_3\, d^2k_\perp}{(2\pi)^3}\, P(k_\perp, k_3; z)\, e^{-ik_3(\chi-\chi')} e^{-i\vec{k}_\perp \cdot \boldsymbol{\vartheta}\, S_k(\chi)}$$

And,

$$\int d\chi'\, e^{-ik_3(\chi-\chi')} = 2\pi\delta_D(k_3)$$

where, $\delta_D(k_3)$ is the Dirac delta function.

$$\omega_m(\vartheta) = \int d\chi\, W^2(\chi) \int \frac{d^2k_\perp}{(2\pi)^2} P(\vec{k}_\perp, z)\, e^{-i\vec{k}_\perp \cdot \boldsymbol{\vartheta}\, S_k(\chi)} \tag{1.52}$$

This is the relation between angular 2PCF and power spectrum, and is called the Limber equation. $S_K(\chi)$ is defined in the Eq.1.6. $S_K(\chi)\vartheta$ is the comoving physical separation.

In three dimensions, the correlation function $\xi_m(r)$ is naturally expanded in plane waves (see Equation 1.47), reflecting the translational invariance. By contrast, the angular 2PCF is defined on the surface of the celestial sphere, where the relevant symmetry is rotational rather than translational. Consequently, the appropriate basis functions are spherical harmonics, which form a complete, orthonormal set on the sphere and respect its rotational invariance. In the flat-sky approximation, the sphere is locally Euclidean, and the spherical harmonics reduce to 2D Fourier modes. Angular power spectrum $C(\ell)$ is the transformation of the angular 2PCF in harmonic space.

$$C(\ell) = \int d^2\vartheta e^{i\vec{\ell}.\vec{\vartheta}} \omega_m(\vartheta)$$

This becomes,

$$C(\ell) = \int d\chi \frac{W^2(\chi)}{S_k^2(\chi)} P\left(\frac{\ell}{S_k(\chi)}, z\right) \tag{1.53}$$

The angular power spectrum $C(\ell)$ at a given multipole ℓ can be viewed as a weighted projection of the three-dimensional matter power spectrum, evaluated at wavenumber $k = \ell/S_k(\chi)$, with the weight set by the radial selection or window function. In this sense, $C(\ell)$ encodes the cumulative contribution of fluctuations at different comoving distances to the clustering signal observed on the sky [36]. Due to statistical isotropy, the angular power spectrum depends only on the magnitude of the multipole ℓ, in complete analogy to the three-dimensional matter power spectrum $P(k, z)$, which depends only on

the modulus of the wavevector $\mathbf{k}$. Physically, low multipoles (ℓ) correspond to correlations on large angular scales, while high multipoles probe smaller angular scales, just as small k in $P(k)$ describes large-scale structure in three dimensions and large k corresponds to small-scale clustering.

For a flat Universe, $S_k(\chi) = \chi$, and the line-of-sight integral can be rewritten in terms of redshift using

$$d\chi = \frac{cdz}{H(z)} \tag{1.54}$$

The window function describes the radial distribution of sources along the line of sight and is defined per unit comoving distance.

$$W(\chi) = \frac{dN}{d\chi}$$

Expressing this distribution in terms of redshift introduces the Jacobian $dz/d\chi = H(z)$, such that

$$W(z) = \frac{H(z)}{c}\frac{dN}{dz} \tag{1.55}$$

Substituting this into Equation 1.53 yields

$$C_\ell = \int dz \, \frac{H(z)}{\chi^2(z)} \left(\frac{dN}{dz}\right)^2 P\left(\frac{\ell}{\chi(z)}, z\right) \tag{1.56}$$

Finally, applying the extended Limber approximation [79], $\ell \to \ell + 1/2$, the angular power spectrum becomes [36],

$$C_\ell = \int dz \, \frac{H(z)}{\chi^2(z)} \left(\frac{dN}{dz}\right)^2 P\left(\frac{\ell + 0.5}{\chi(z)}, z\right) \tag{1.57}$$

The angular 2PCF of tracers can be estimated using the Landy–Szalay estimator,

$$\omega_g(\vartheta) = \frac{DD(\vartheta) - 2DR(\vartheta) + RR(\vartheta)}{RR(\vartheta)} \tag{1.58}$$

Here, $DD(\vartheta)$, $DR(\vartheta)$, and $RR(\vartheta)$ represent the pair counts normalised by number density within an angular bin centred at ϑ, for the data-data, data-random, and random-random pairs, respectively. Angular 2PCF of matter and tracer can also be related by scale-independent tracer bias,

$$\omega_g(\vartheta) = b^2 \omega_m(\vartheta) \tag{1.59}$$

The framework presented in this section underpins the analysis in Chapter 4, where it is applied to the angular clustering of quasars. In that chapter, systematic effects and statistical uncertainties associated with angular 2PCF measurements are quantified. The resulting clustering measurements enable a detailed study of the redshift evolution of

quasar bias and provide estimates of the characteristic masses of quasar host dark matter halos.

1.5.5 Count-in-Cells statistics

At early epochs (redshift ≥ 10), the growth of density perturbations can be accurately described by linear perturbation theory. The Fourier modes of the perturbation evolve independently, thereby preserving the statistical properties of the primordial fluctuations. In particular, if the initial density field is a Gaussian random field, it remains Gaussian under linear evolution. As a result, the statistical properties of the density field are completely specified by the two-point correlation function or power spectrum[108].

As density fluctuations grow through gravitational instability, linear theory eventually breaks down, and non-linear effects become important. In this non-linear regime, the evolution of density perturbations is governed by mode coupling between different Fourier modes, arising from the intrinsic non-linearity of gravitational instability. The theory is derived using higher-order perturbation theory and provides a systematic expansion of the non-linear evolution equations for the density contrast. Gaussian initial conditions for this non-linear evolution generate non-Gaussianity only through gravitational mode coupling, whereas non-Gaussian initial conditions additionally imprint primordial higher-order correlations (check [10] for more details).

A complete statistical description of the evolved matter distribution, therefore, requires measurements beyond second order. In particular, the bispectrum and trispectrum, corresponding to the connected three- and four- point correlation functions (3PCF and 4PCF) are non-zero [97, 108]. They contain the information that is not captured by the power spectrum alone. Estimation of 3PCF and 4PCF of point-wise density field for galaxy surveys require triplet and quadruplet counting, respectively [114]. Such higher-order measurements are computationally expensive, present a significant challenge for accurate covariance estimation, and require careful treatment for shot-noise of galaxy data. Furthermore, the precision of the correlation function is largely affected by the observational systematic effects, such as redshift uncertainties.

These motivate us to use a different estimator for the higher-order correlation function. In addition to higher-order correlation functions, there are other alternative statistical descriptions exist in cosmology to characterise the non-Gaussian features of density distribution, including Minkowski functionals [86], void statistics [120]. Each of these approaches has distinct advantages and limitations in terms of physical interpretation, computational cost, and sensitivity to observational systematics.

This thesis employs count-in-cells (CiC) statistics to study non-Gaussian features through volume- or angular-averaged correlation functions, rather than the pointwise n-point correlation functions [97, 17, 66, 47, 54]. It avoids explicit n-tuple counting. These quantities are derived by convolving the density field with a window function,

yielding a smoothed density field. Theoretical predictions for these higher-order statistics are obtained from cosmological perturbation theory up to fourth order. Higher-order from galaxy data compared with results from N-body simulations where perturbative predictions become unreliable. CiC statistics are estimated from observed galaxy data by counting the objects in a randomly placed cell over the survey footprint. The following sections discuss the theoretical prediction for these volume-averaged correlation functions of the matter overdensity field up to fourth order, and describe how they are estimated from photometric galaxy samples.

1.5.5.1 Count-in-Cells statistics of matter over density field

The smoothed density contrast field $\delta_W(\mathbf{x})$ is obtained by convolving the density contrast $\delta(\mathbf{x})$ with a window function $W(\mathbf{x}-\mathbf{x}')$. This operation replaces the local value of the field with a weighted average over neighbouring points, where the weighting is determined by the chosen window. The effective volume associated with the window sets the spatial scale of smoothing. Top-hat or Gaussian window functions are the two most popular smoothing functions [47, 10]. In the case of the former, all points within the volume contribute uniformly, while for the latter, points closer to the centre of the window contribute greater weight. The smoothed field thus encodes the density contrast averaged over a finite spatial scale. This thesis adopts the top-hat filter because of its mathematical simplicity, and it is straightforward to calculate the smoothed density variance and higher-order moments for arbitrary power spectra [10]. Smoothed density contrast field is defined as,

$$\delta_W(\mathbf{x}) = \frac{1}{V_W}\int d^3\mathbf{x}'\delta(\mathbf{x}')W(|\mathbf{x}-\mathbf{x}'|) \tag{1.60}$$

Where,

$$W(\mathbf{x}-\mathbf{x}') = \begin{cases} 1, & |\mathbf{x}-\mathbf{x}'| \leq R \\ 0, & \text{otherwise} \end{cases}$$

R is the radius of the sphere and V_W is the effective volume of the tophat window function.

$$V_W = \int d^3\mathbf{x}W(\mathbf{x}) = \frac{4\pi R^3}{3}$$

For top-hat window, $\delta_W(\mathbf{x})$ is the volume average of $\delta(\mathbf{x})$. The correlation function for δ_W is called the averaged correlation function [47]. Averaged 2PCF, $\bar{\xi}_2(\mathbf{r})$ is the joint ensemble average of the density in an arbitrary two locations [10]. By the ergodic hypothesis, this ensemble average can be treated as spatial average. We can define the volume-averaged J-point correlation function as follows.

$$\bar{\xi}_J(R) \equiv \left\langle \delta_W^J(\mathbf{x}) \right\rangle_c$$

Here, the subscript c stands for connected moments, also called cumulants (see Section 3.2.1 in [10]).

$$\bar{\xi}_J(\mathbf{r_1}, \mathbf{r_2}, ..., \mathbf{r_J}) = \frac{1}{V_W^J} \int d\mathbf{r_1}, ..., d\mathbf{r_J} W(\mathbf{r_1}), ..., W(\mathbf{r_J}) \xi(\mathbf{r_1}, .., \mathbf{r_J})$$

This can be set down in Fourier space. Statistical homogeneity implies that the $\langle \delta_W(\mathbf{k_1}), \delta_W(\mathbf{k_2}), .., \delta_W(\mathbf{k_J}) \rangle_c$ is proportional to $\delta_D(\mathbf{k_1} + \mathbf{k_2} + ... + \mathbf{k_J})$, where δ_D is the Dirac delta function.

$$\langle \delta_W(\mathbf{k_1}), \delta_W(\mathbf{k_2}), .., \delta_W(\mathbf{k_J}) \rangle_c = \delta_D(\mathbf{k_1} + \mathbf{k_2} + ... + \mathbf{k_J}) P_J(\mathbf{k_1}, \mathbf{k_2}, ..., \mathbf{k_J})$$

For J=3, $P_3(\mathbf{k_1}, \mathbf{k_2}, \mathbf{k_3})$=$B(\mathbf{k_1}, \mathbf{k_2}, \mathbf{k_3})$ called the Bispectrum. For J=4, ($P_4(\mathbf{k_1}, \ldots, \mathbf{k_4}) = T(\mathbf{k_1}, \ldots, \mathbf{k_4})$ called Trispectrum. These are obtained from higher-order perturbation theory.

We can define the angular projection of the smoothed density contrast. The significance of angular projections is explained in Section 1.5.4. The comoving 3D coordinates $\mathbf{x}$ can be defined as $[\chi, \vec{\Omega}]$ where χ is the comoving radial coordinate (see, Equation 1.4) and $\vec{\Omega}$ is the sky-pointing angular vector [37].The projected overdensity weighted by the (normalised) radial selection function, $\delta_{\rm 2D}(\Omega)$, is

$$\delta_{\rm 2D}(\vec{\Omega}) = \int_{\chi_{\rm min}}^{\chi_{\rm max}} d\chi W(\chi) \chi^2 \delta([\chi, \vec{\Omega}])$$

$\chi_{\rm min}$ and $\chi_{\rm max}$ are the distance ranges covered by the survey, $W(\chi)$ is the normalised radial selection function. $W(\chi)$ is the normalised probability that an object at a distance χ is included in the catalogue [10]. This local projected density contrast is smoothed over an angular scale θ. For a top-hat window in the angular plane, θ is the radius of a circle centred on an angular direction $\vec{\Omega}$. The moments of the smoothed projected density contrast $\delta^J_{2D,W}(\theta)$ give the J-point angular averaged correlation function.

$$\bar{\omega}_J(\theta) \equiv \langle \delta^J_{2D,W}(\theta) \rangle_c \tag{1.61}$$

The second ($\bar{\omega}_2(\theta)$) and third ($\bar{\omega}_3(\theta)$) moments can be approximated using tree-level perturbation theory, which is valid in the weakly non-linear regime [48].

$$\bar{\omega}_2(\theta) = \frac{1}{2\pi} \int_{\chi_{min}}^{\chi_{max}} \chi^4 F^2(\chi) d\chi \int_0^\infty k P(k) W^2_{\rm 2D}(k\theta\chi) dk \tag{1.62}$$

$$\bar{\omega}_3(\theta) = \frac{6\theta^{-4}}{(2\pi)^2} \int_{\chi_{min}}^{\chi_{max}} \chi^2 F^3(\chi) d\chi \left[\frac{5}{14} \int_0^\infty q W^2_{\rm 2D}(q) P(k) dq - \frac{1}{4} \int_0^\infty q^2 W^2_{\rm 2D}(q) \frac{dP(k)}{dq} dq \right] \tag{1.63}$$

Here, $q = k\theta\chi$, $W_{2D}(k\theta\chi)$ is the window function in Fourier space. For the top-hat window,

$$W_{2D}(k\theta\chi) = 2\,\frac{J_1(k\theta\chi)}{k\theta\chi}$$

J_1 is the first-order Bessel function. Re-scaled connected moments, also known as reduced cumulants or hierarchical amplitudes, quantify the non-Gaussianity of the distribution. It is defined as

$$S_J(\theta) = \frac{\bar{\omega}_J(\theta)}{[\bar{\omega}_2(\theta)]^{J-1}} \tag{1.64}$$

$S_3(\theta)$ and $S_4(\theta)$ are referred to as the skewness and kurtosis, respectively, while for $J > 4$ they are known as higher-order reduced cumulants. Beyond the weakly non-linear regime, where tree-level perturbation theory becomes inaccurate, one-loop perturbative corrections can be used to extend the theoretical description [10]. On even smaller scales, where higher-order perturbative corrections are no longer sufficient, the fully non-linear evolution of the density field is best described using N-body simulations (see [37] for a comparative study).

1.5.5.2 Count-in-Cells estimator for photometric galaxy sample

The smoothed density contrast field is constructed directly from galaxy counts by the following procedure. This section explains how to estimate the J-point angular averaged correlation function, $\bar{\omega}_J(\theta)$ for the photometric galaxy sample.

Under the ergodic hypothesis, the moments of the smoothed density contrast can be replaced by spatial averages over the sky. We implemented a Monte Carlo estimator in which a large number of circles,$N_C \to \infty$, with a fixed angular smoothing scale θ are randomly placed over the survey footprint (see [37, 38] for more details). For each placement, the number of objects N enclosed by the window is counted, yielding N_C measurements of counts at a given θ. N is a discrete random variable and its J-th central moment is given by,

$$m_J(\theta) \equiv \frac{1}{N_C}\sum_{i=0}^{N_C}(N_i - \langle N\rangle)^J \tag{1.65}$$

where N_i is the number of galaxies in i-th cell, $\langle N\rangle$ is the mean count over all circles with a given smoothing scale θ.

Central moments are the moments about the average count and they contain both Gaussian and non-Gaussian parts. The non-Gaussian information is isolated by considering connected moments (cumulants) of the counts-in-cells distribution, which subtract the disconnected (Gaussian) contributions [10, 37]. The first few connected moments,μ_J, are follows.

$$\begin{aligned}\mu_2 &= m_2,\\ \mu_3 &= m_3,\\ \mu_4 &= m_4 - 3m_2^2.\end{aligned} \tag{1.66}$$

The shot-noise contribution of the sparse galaxy sample is estimated by modelling the discrete random variable N as a Poisson distributed with mean $\langle N \rangle$ [48, 37]. Shot-noise corrected connected moments of the random variable N are,

$$\begin{aligned} k_2 &= \mu_2 - \langle N \rangle, \\ k_3 &= \mu_3 - 3k_2 - \langle N \rangle, \\ k_4 &= \mu_4 - 7k_2 - 6k_3 - \langle N \rangle. \end{aligned} \tag{1.67}$$

In summary, m_J are central moments of the counts-in-cells random variable N, μ_J are the corresponding connected moments (cumulants) of N, and k_J are the Poisson shot-noise corrected cumulants (often equivalent to factorial cumulants) used to estimate the connected moments of the underlying projected overdensity field after normalising by $\langle N \rangle^J$. The connected moments of the density contrast therefore provide estimators of the J-point angular-averaged correlation functions for our galaxy sample [97].

$$\langle \delta^J_{\mathrm{2D},W}(\theta) \rangle_c = \frac{k_J}{\langle N \rangle^J} = \bar{\omega}_J(\theta) \tag{1.68}$$

Thus we can get the reduced cumulants by employing Equation 1.64.

Importantly, the CiC estimator is sensitive to both gravitationally induced non-Gaussianity and tracer bias. Because galaxy bias affects higher-order moments differently than the two-point function. Reduced cumulants like S_3 and S_4 depend on non-linear bias parameters (quadratic/cubic bias) in addition to gravitational evolution. Therefore, they can be used to constrain non-linear bias parameters beyond the linear bias approximation.

1.6 Summary

In this chapter, the theoretical background necessary for studying the large-scale structure of the Universe was introduced. The discussion began with the basic principles of cosmology within the framework of General Relativity, where gravity is described by Einstein's field equations. Under the assumptions of large-scale homogeneity and isotropy, these equations lead to the Friedmann-Lemaître-Robertson-Walker (FLRW) metric and the Friedmann equations, which govern the expansion of the Universe. The standard cosmological model, ΛCDM, was then introduced, describing the Universe in terms of matter, radiation, and dark energy components.

The concept of cosmological redshift and its relation to the expansion of the Universe was reviewed, together with its importance for measuring distances and mapping the large-scale distribution of galaxies. The Hubble-Lemaître law was discussed as the observational relation between redshift and distance at low redshift, highlighting its role in observational cosmology.

The chapter then introduced the theoretical framework of cosmic structure formation. Small primordial density perturbations, generated during inflation, grow through

gravitational instability in an expanding Universe. In the linear regime, the evolution of these perturbations can be described using Newtonian perturbation theory, leading to the linear growth equation for the matter density contrast. The statistical properties of these fluctuations were characterised using the matter power spectrum, which encodes the distribution of density fluctuations across spatial scales. The transition from linear to non-linear structure formation and the use of empirical models such as `Halofit` to describe the nonlinear matter power spectrum were also discussed.

Finally, the statistical tools used to quantify the clustering of matter and its observable tracers were introduced. The two-point correlation function and its Fourier-space counterpart, the power spectrum, provide fundamental measures of the clustering of matter in the Universe. Since observations detect discrete tracers such as galaxies and quasars rather than the underlying matter field, the concept of tracer bias was introduced to relate the clustering of tracers to that of dark matter. For photometric surveys lacking precise distance measurements, the angular two-point correlation function was presented as a key observable obtained by projecting the three-dimensional clustering signal onto the sky.

In addition to two-point statistics, the chapter introduced Count-in-Cells statistics as an alternative approach to characterising the statistical properties of galaxy distributions. In this method, the survey volume is divided into spatial cells and the number of galaxies within each cell is counted to study the probability distribution and higher-order moments of the density field. Count-in-Cells statistics provide complementary information to correlation functions by probing the variance and non-Gaussian properties of the density field.

The theoretical framework developed in this chapter provides the cosmological and statistical foundations for the analyses presented in the remainder of this thesis. In particular, the concepts of redshift, clustering statistics, tracer bias, angular correlations, and Count-in-Cells measurements are central to the study of galaxy and quasar distributions in wide-field surveys. The following chapter introduces photometric redshift estimation methods, which enable redshift measurements for large photometric samples. These redshift estimates are subsequently used in later chapters to investigate the clustering of quasars and to analyse the statistical properties of galaxy distributions using Count-in-Cells statistics in the KiDS survey.

Part II

Photometric Redshift Estimation

This chapter presents the importance of redshift estimation and reviews the methods used to obtain it. It further introduces the key concepts of machine learning and discusses their application to photometric redshift estimation, which is one of the main focuses of this thesis.

2.1 Redshift and luminosity

The fundamental measurable quantity for celestial objects beyond our solar system is the energy received in the form of electromagnetic radiation, commonly referred to as *flux*. The measurement of flux, known as *photometry*, involves observing the intensity of radiation over wavelength bands. Photometric analysis provides crucial insights into an object's total energy output (*luminosity*), temperature, size, and other physical characteristics, particularly when combined with distance estimations. Intrinsic (rest-frame) luminosity is the total energy emitted per unit time in the source's rest frame that is independent of the observer's location. In contrast, observed flux is the energy received per unit time per unit area by an observer. Observed flux depends on redshift.

When observing a distant source at redshift z, the radiation we detect has been affected by the expansion of the Universe. If a source emits radiation with intrinsic spectral luminosity $L(\nu_1)$ at frequency ν_1, the observed frequency is redshifted as described in Section 1.3. The observed flux density $F(\nu_0)$ (energy received per unit time, area, and frequency) is related to the intrinsic luminosity by

$$F(\nu_0) = \frac{L(\nu_1)}{4\pi D_M^2 (1+z)}, \tag{2.1}$$

where D_M is the transverse comoving distance to the source. For bolometric quantities (integrated over frequency), the bolometric flux density F_{bol} is,

$$F_{\text{bol}} = \frac{L_{\text{bol}}}{4\pi D_L^2}, \tag{2.2}$$

where

$$D_L = (1+z)\, D_M \tag{2.3}$$

D_L is the luminosity distance. The luminosity distance is defined so that the flux-luminosity relation retains the familiar inverse-square form.

Due to redshift, the emitted frequency is shifted to the observed frequency, the overall flux amplitude is reduced through the combined effects of the $(1+z)$ factors and the luminosity distance, and intrinsic spectral features such as breaks and emission lines are systematically displaced across photometric filter bands. Together, these effects imprint a characteristic redshift-dependent signature on the observed multi-band fluxes. The following sections outline the methodology used to estimate the redshift from both spectroscopic and photometric surveys.

2.2 Spectroscopic and photometric redshifts

In spectroscopy, incident light is dispersed by a spectrograph into many narrow wavelength bins, allowing measurement of the flux density of an astronomical source as a function of wavelength. The ability of a spectrograph to distinguish between two nearby wavelengths is characterised by its *spectral resolution*, defined as

$$R = \frac{\lambda}{\Delta\lambda}, \tag{2.4}$$

where λ is the central wavelength and $\Delta\lambda$ represents the smallest wavelength difference that can be resolved, which is typically the Full Width at Half Maximum (FWHM) of the instrumental profile. Higher values of R correspond to finer wavelength discrimination. In practice, spectroscopic instruments cover a wide range of resolutions, from $R \sim 100$ for low-dispersion multi-object spectrographs (MOS) to $R > 30{,}000$ for high-resolution echelle instruments.

Redshifts are determined by identifying and measuring the observed wavelengths of spectral features (emission or absorption lines) and comparing them to their corresponding rest-frame values. When several distinct or blended features are detected, *spectroscopic redshifts* (spec-zs) can be measured with a precision better than 10^{-3}. The uncertainty of the redshift is fundamentally limited by the instrumental resolution and the signal-to-noise ratio (SNR) [92]. Under ideal conditions, if the spectral features are narrower than the instrumental response and the SNR is high, the redshift precision is of order the fractional wavelength precision. In reality, the widths of the spectral lines are finite, influenced both by physical broadening mechanisms (such as thermal or turbulent motions) and by instrumental resolution [53]. These effects collectively limit the achievable precision in redshift measurement.

For faint sources ($m_{\rm AB} \gtrsim 24$), the determination of spectroscopic redshifts become increasingly challenging due to the reduced SNR and the finite wavelength coverage of

spectrographs. Key spectral features may be too weak to detect, or they may fall outside the accessible wavelength range of the instrument for a given observational setup. To achieve a high SNR longer integration time is required. Consequently, the success rate for obtaining reliable spec-zs may fall below 50-70% [106]. Despite the high accuracy of spectroscopic measurements, their application to very large samples is constrained by the significant observation time required. The Stage-IV spectroscopic survey DESI observes up to 5000 targets simultaneously using its highly multiplexed instrumentation, thereby producing spec-z catalogue comprising millions of objects. Despite this capability, photometric surveys continue to play a crucial role, as discussed below.

Photometry measures the total integrated flux of an object through a set of filters rather than dispersing the light. Each filter transmits light over a finite wavelength range $\Delta\lambda$, centred at λ. The spectral resolution of the photometric observations can therefore be expressed analogously as $R = \lambda/\Delta\lambda$. Since each filter covers a broad wavelength interval (typically 1000Å wide), it requires a short exposure time. Therefore, photometric instruments have low effective resolutions, typically $R \sim 5$–50. In addition to the short exposure time, photometry covers a larger area of the sky compared to that of the MOS and allows us to measure the redshift of millions of objects simultaneously.

Large imaging surveys such as KiDS [127], the DES, and LSST employ broadband filters with $R < 10$. At such a low resolution, discrete spectral features are heavily integrated over, making it challenging to localise emission or absorption lines. As a result, *photometric redshifts* (photo-zs, explained in the next section) are inherently less precise than spectroscopic measurements. For high-quality, uniform photometry, obtained with numerous narrowband filters or for sources with consistent intrinsic spectra showing strong broad features, photo-z uncertainties of less than $0.01(1+z)$ (where z is the redshift of an object) have been achieved, for example, in PAUS [3], J-PAS [75] and COSMOS [109]. In contrast, for typical broadband photometric surveys, achieving photo-z uncertainties below $\sim 0.015(1+z)$ remains very challenging. An additional limiting factor is the filter wavelength coverage: the absence of key bands, such as the u-band or near-infrared data, can significantly degrade photo-z accuracy, and not all surveys provide such coverage.

Although less accurate, photometric redshifts are indispensable for large-scale cosmological studies, where spectroscopy is impractical for the billions of faint galaxies observed because of the difference between exposure times of spectroscopy and broadband imaging. For example, obtaining spec-zs for all the objects in LSST "gold sample" would be impractical, even with highly multiplexed instruments such as DESI. Moreover, a substantial fraction of faint galaxies fail to yield reliable spec-zs, whereas the catastrophic outlier rate of modern photo-z techniques can be lower than the spectroscopic failure rate [92]. Consequently, photo-z estimation remains the only feasible approach for characterising the redshift distribution of the largest extragalactic samples with the current available technologies. It enables estimating the redshift distribution and cosmological studies such

as cosmic shear [122], which then constrain tracer bias, S_8, Ω_m, and the growth rate of structures. In Chapter 4, the distribution of quasar photo-zs is used to determine the quasar bias and host halo mass from their angular clustering signal.

2.3 Photometric redshift estimation methods

Photometric redshift estimation relies on the empirical mapping between photometric observables (e.g., colours and magnitudes) and redshift, first introduced six decades ago [6, 7] and initially applied to elliptical galaxies in distant clusters. It was further developed during the 1980s [69, 80]. In the following decades, several significant developments improved both the accuracy and applicability of photo-z methods [27, 4, 16, 25, 12]. These developments collectively form the foundation of modern photo-z estimation, enabling its essential role in current and upcoming large-scale surveys.

Photo-z estimation relies on identifying how characteristic spectral features of galaxies such as the Balmer (4000 Å) and Lyman (912-1216 Å) breaks shift to longer wavelengths with increasing redshift. These breaks cause a sharp change in flux between the blue and red sides of the spectrum. In addition to this redshift of spectral features, the observed flux decreases approximately with the square of the luminosity distance (see Equation 2.1). Apparent magnitude carries redshift information but is not a direct proxy for distance without assumptions about the source's SED and the associated K-corrections due to bandpass shifting. Consequently, apparent magnitudes act only as indirect distance indicators.

In broadband photometry, where each filter measures the total flux over a wide wavelength range, the precise wavelength of the break cannot be directly measured. Instead, its presence is inferred from the colour differences between adjacent filters. A large change in flux between two bands indicates that a break is likely to lie between them. It is shown in Figure 2.1. However, strong colour gradients can also arise from other effects, including dust extinction, differences in galaxy type, strong emission lines, or photometric noise, all of which can mimic or obscure the true spectral break. Consequently, using a single colour (e.g., $r - i$) may lead to degeneracies, where multiple redshift or galaxy-type combinations produce similar colours. These degeneracies are reduced by combining many filters that cover a broad wavelength range, allowing the overall shape of the spectral energy distribution (SED) of galaxy to be more accurately constrained. The example of photo-z degeneracy is illustrated in Figure 2.2. When all colours are considered together, the pattern of flux variations reveals the most consistent redshift solution, enabling reliable photo-z estimation even without full spectroscopic resolution [106]. Photo-z estimation can be broadly categorised into two: template-based and data-driven methods. Template-based methods break degeneracies using physical SED models plus priors, while data-driven methods learn the mapping from representative spectroscopic training sets.

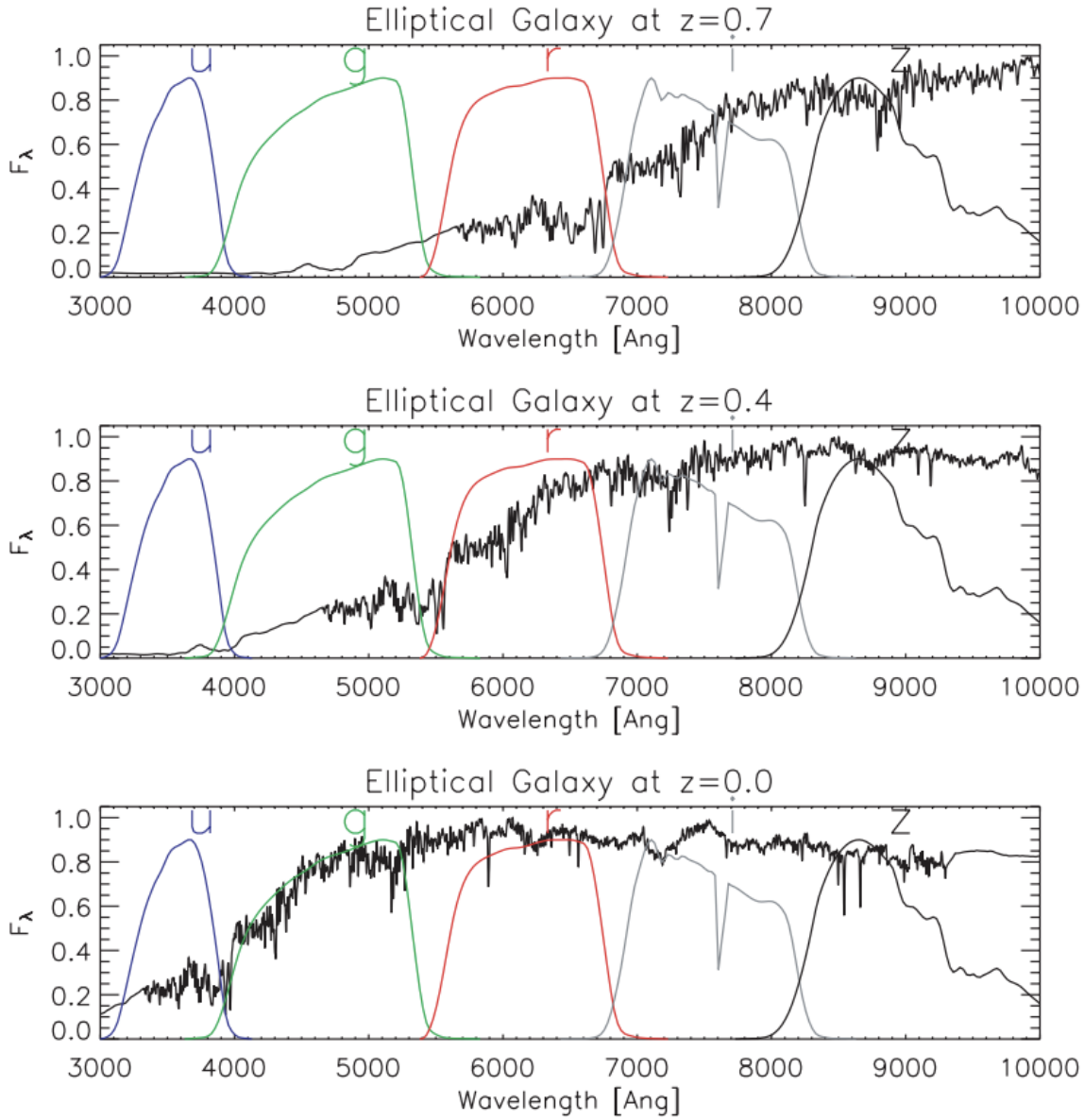


Figure 2.1: The figure is from [94]. Model spectrum of an elliptical galaxy at three redshifts (z = 0.0, 0.4, and 0.7) with SDSS u, g, r, i, and Z filter response functions, F_λ, overplotted. As redshift increases, the 4000 Å break shifts through the filters, producing a sharp drop in flux between u–g at $z = 0$, between g–r at $z = 0.4$, and between r–i at z= 0.7.

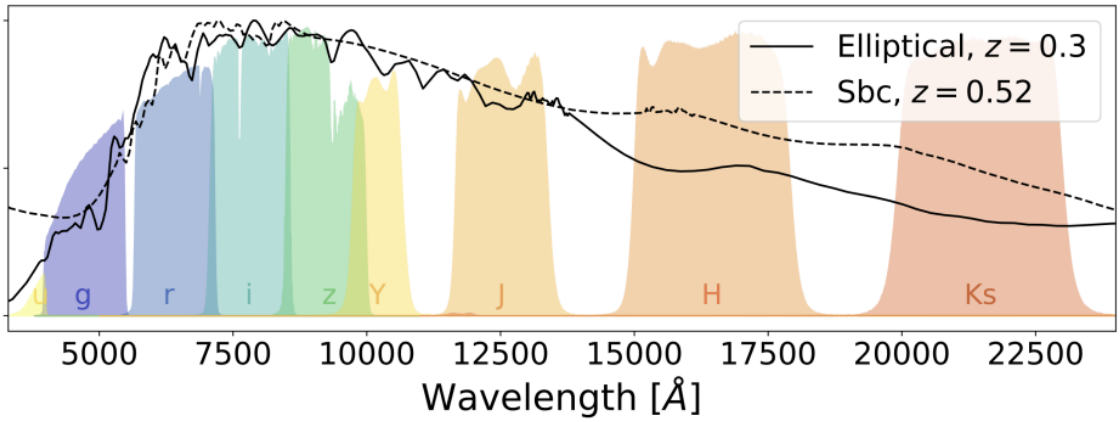


Figure 2.2: This plot is from [92]. It is an illustration of photo-z degeneracy. The y-axis is the normalised flux. Using only g, r, i band photometry leads to strong degeneracies between galaxy type (elliptical and Sbc) and redshift, preventing reliable photo-z estimates for individual galaxies. Additional photometric bands are required to break these degeneracies.

2.3.1 Template-based methods

Template-fitting methods are parametric techniques which estimate redshifts by comparing observed photometric fluxes with model predictions derived from a library of SED templates. These templates can be generated from either the theory using stellar population synthesis models [19, 83] or observed spectra. Each template is redshifted over a range of z values and corrected by including key astrophysical and observational effects such as nebular emission, dust attenuation, intergalactic medium absorption, and galactic extinction [106]. The resulting model spectra are integrated through the survey filter transmission curves to produce predicted fluxes that can be directly compared with the observations. For each galaxy, the best-fitting redshift and template combination is determined by minimising the χ^2 difference between observed flux ($f_{\mathrm{obs,i}}$) and predicted flux ($f_{pred(z,t,O),i}$) for a given set of parameters z, template type t, and any additional parameters O such as dust extinction/reddening, emission line strength.

$$\chi^2(z,t) = \sum_i \frac{\left(f_{obs,i} - f_{pred(z,t,O),i}\right)^2}{\sigma_i^2}$$

Where, σ_i is the uncertainty in the i-th observed flux. This approach often led to degeneracies and catastrophic outliers because it lacked prior information on galaxy distributions. The major advance was the Bayesian formulation introduced by Benítez [9], which incorporated priors such as absolute magnitude, colours, and computed the full posterior probability for the redshift given a set of observed fluxes $p(z \mid f_{\mathrm{obs}})$.

$$p(z \mid f_{\mathrm{obs}}) = \frac{\int p(f_{\mathrm{obs}} \mid z,t,O)p(z,t,O)\,dt\,dO}{p(f_{\mathrm{obs}})} \tag{2.5}$$

Here, the likelihood $p(f_{\rm obs} \mid z, t, O)$ is the probability of obtaining the particular values of the observed fluxes in each band for the object, assuming a set of values z, t,and O. For Gaussian errors, it is proportional to $e^{-\chi^2/2}$.

Modern implementations such as BPZ [9], LePhare [4, 61] and EAZY [18] compute the full probability distribution of the redshift and its uncertainties. In KiDS, the default photo-zs are obtained via BPZ. Despite their interpretability and solid physical foundation, template-fitting methods face intrinsic limitations arising from incomplete template libraries, simplistic dust and emission models, and high computational cost for large surveys. The template set is inherently incomplete since no two galaxies have identical SEDs, any finite library can only approximate the diversity of real galaxies. Expanding the template set introduces additional degrees of freedom, often leading to unphysical or biased solutions. Second, the priors may be inaccurate, even with flexible templates, and uncertainties in the assumed distributions of galaxy types, luminosities, and redshifts can strongly influence the results, especially when photometry is noisy or limited to a few bands. Finally, data rarely inform the model directly since most template-fitting frameworks treat the templates and priors as fixed, although real photometric data could, in principle, be used to refine them [92].

2.3.2 Data-driven methods

Data-driven photo-z methods encompass all techniques that derive the mapping between photometric observables and redshift directly from observational data rather than from physical SED models. Early examples include empirical colour-redshift relations and polynomial regressions calibrated on spectroscopic samples [e.g., 27], as well as nearest-neighbour interpolation methods [e.g., 30]. Modern machine learning (ML) approaches represent a natural extension of these data-driven techniques, employing more flexible and non-linear models to capture complex correlations in high-dimensional photometric data. Machine learning based photo-z methods offer several advantages over template-fitting approaches. They are computationally efficient and scalable to large datasets. They are also well-suited to modelling non-linear dependencies between photometric colours and redshift. Moreover, ensemble and probabilistic frameworks can provide empirical estimates of redshift uncertainties, which are essential for cosmological analyses. ML based photo-z uncertainties are reliable primarily within the training data distribution, whereas SED methods retain interpretability in extrapolation regimes. In addition, ML models suffer from *training set bias*: since they rely on spectroscopic samples for training, their accuracy decreases when applied to regions of colour–magnitude space that are poorly represented in the training data, such as very faint or high-redshift galaxies. Furthermore, it is difficult to validate photo-z performance beyond the training/calibration set. Standard mitigation and validation strategies include cross-validation within the spectroscopic sample, reweighting the training set to better match the photometric sample distributions,

and external consistency checks such as clustering redshifts or self-organizing maps to diagnose coverage in feature space [56, 63]. Recent developments in domain adaptation, self-supervised learning, and hybrid template–ML approaches aim to mitigate these limitations by improving generalisation across different datasets and extending applicability to the full photometric sample.

In summary, ML photo-z methods demonstrate strong performance when interpolating within the training set, offering computational efficiency and scalability. However, their performance degrades for out-of-distribution objects due to spectroscopic selection effects ("training-set bias"), as the spectroscopic sample is typically not representative across the full colour–magnitude–z space. In contrast, template-fitting methods can extrapolate based on physical spectral models, but their accuracy depends on the fidelity of the adopted templates and priors.

Building on the motivation outlined above, the next section introduces the fundamental concepts of machine learning. This is followed by chapters detailing its application to photo-z estimation for bright galaxies and quasars in the KiDS-DR4 sample [13, 91].

2.3.3 Machine learning

Astronomy has entered the era of big data in both volume and complexity, motivating the adoption of data-driven approaches that complement traditional model-based analyses. In this work, we focus specifically on machine learning methods for photo-z estimation. In particular, we consider supervised learning techniques, where models are trained on spectroscopic samples and evaluated using standard training, validation, and test splits. Key methodological aspects include the choice of loss function, mitigation of overfitting, the bias–variance trade-off, and the estimation of predictive uncertainties.

2.3.4 Supervised learning

Machine learning encompasses a variety of paradigms for extracting patterns and making predictions from data. In our work, we focus on supervised learning and regression task, which finds a mapping from inputs to outputs, $f : X \to Y$, given a labeled set of input-output pairs,

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}.$$

$\mathcal{D}$ is called training data and N is called the number of training samples. X is called the feature space, and its elements $\mathbf{x}_i$ are referred to as feature vectors. Y is called the label space, and its elements y_i are referred to as labels. In our work, $\mathbf{x}_i$ consists of galaxy image pixel intensities and galaxy magnitudes, while y_i will be the spectroscopic redshift.

Most supervised learning techniques categorise outputs, known as response/target variable, as either discrete (categorical or nominal) variables belonging to a finite set, $y_{\text{predicted}} \in \{1, \ldots, C\}$ (e.g., classification of galaxies, quasars, and stars in a survey catalogue), or continuous real-valued variables (e.g., photo-zs, stellar mass prediction of

individual galaxies). When the response is categorical, the learning task is called a classification or pattern recognition problem. Conversely, when the response variable is continuous, the task is defined as a regression problem. Another specialised form, ordinal regression, occurs when the output label space follows a meaningful ordering (such as grades from A to F) [90].

2.3.5 Galaxy images and photometric redshift estimation

The selection of input features for machine learning models in photo-z estimation is generally guided by a combination of domain knowledge and empirical validation. In practice, candidate features (e.g., colours, magnitudes, morphology indicators) are chosen based on their expected physical relevance and are assessed through ablation studies or performance on validation data. Feature-importance analyses can provide complementary diagnostics, particularly for tabular models, by quantifying the relative contribution of each input to the predictive performance and highlighting their relevance for the task. An example of such an analysis in the context of photo-z estimation from tabular data within the KiDS is presented by [91]. Photometric magnitudes and colours constitute the primary features used to train most photo-z models [104]. These quantities are engineered features derived from astronomical imaging data rather than raw pixel information. However, the measured magnitudes and colours are not direct physical observables; instead, they depend on several observational and methodological factors. These include the choice of aperture size, spatial and temporal variations of the Point Spread Function (PSF), and blending effects arising from nearby or overlapping sources. Even model-based magnitudes, which aim to correct for PSF effects and to account for galaxy light profiles through parametric fitting, capture only a limited fraction of the total information content available in the imaging data [95]. Consequently, conventional photometric features represent a compressed and potentially information losing summary of the underlying images, which may limit the performance of ML based photo-z estimators.

The success of image recognition in AI research [76] has influenced its applications in cosmology. It allows us to use images (raw pixel intensities) as input features, enabling the extraction of information that is often missing from engineered image features. The missing information carries redshift information, includes galaxy surface brightness distributions, detailed morphological structures and colour gradients, can improve the photo-z estimates [58, 112].

Although photometric quantities such as magnitudes and colours are ultimately derived from images and therefore do not contain fundamentally new information beyond the pixels, they are still often included as input features because they provide robust, noise-reduced, and physically meaningful summaries of the data. These engineered features are typically calibrated and corrected for instrumental effects, making them more stable than quantities learned directly from noisy images, especially in cases of limited sample size

or low signal-to-noise ratios. Moreover, learning global properties such as total flux or colour indices directly from pixels can be computationally demanding and data intensive, whereas supplying them explicitly can improve training efficiency and convergence of the model. Given the finite capacity of neural networks, combining images with photometric features also helps mitigate limitations in model architecture and training data. In this sense, pixel-level data capture detailed spatial and morphological information, while derived photometric quantities encode global astrophysical summaries; together, they provide complementary representations that often lead to improved empirical performance [33, 95].

In our work (Chapter 3 & 4), Input features to the model are pixel intensities of 2D galaxy images and galaxy magnitudes, while labels correspond to the spec-zs. Since the goal is to predict continuous photo-z values for individual objects, the task is formulated as a regression problem. Several other works [95, 1] formulate photo-z estimation as a classification problem by discretising the redshift range into narrow bins and assigning each galaxy to a corresponding bin. Each formulation has its own advantages and limitations. Regression-based approaches preserve the continuous nature of redshift and avoid discretisation effects, enabling per-object photo-z predictions that can be directly compared with spectroscopic measurements. However, regression models do not intrinsically provide redshift probability distributions and typically require additional techniques to quantify predictive uncertainties. In contrast, classification-based approaches naturally yield binned redshift probability distributions, which are valuable for uncertainty characterisation. Their performance, however, depends on the choice of redshift binning, and discretisation can introduce boundary effects or limit redshift resolution. Recent work within LSST [87] extends regression-based photo-z estimation by predicting continuous per-object redshift probability distributions and provides uncertainty information without relying on discretised bins.

The subsequent sections outline the ML approaches adopted to learn representations from both galaxy images (pixel-level data) and photometric magnitudes, and to integrate these for photo-z estimation.

2.3.6 Artificial Neural Networks

Linear, probabilistic, and tree-based models [15] remain useful for many applications. Deep learning models, particularly Artificial Neural Networks (ANNs, [85]) and Convolutional Neural Networks (CNNs,[77]) can learn representations from data, supporting high-dimensional, structured, and hierarchical pattern recognition across domains like vision, speech, and natural language. The term *deep* refers to the composition of multiple layers of nonlinear transformations. The presence of many layers of computational units (*neurons*) enables hierarchical feature representations [76]. ANNs, inspired by early models of biological neurons, can learn non-linear mappings between input and output data.

They consist of interconnected neurons, organised into sequential layers: an *input layer*, one or more *hidden layers*, and an *output layer*. Each neuron performs two fundamental operations: a linear transformation of its inputs followed by a nonlinear activation. Mathematically, a single neuron is represented as

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{2.6}$$

where x_i are input features, w_i are learnable weights, b is a bias term, and $f(\cdot)$ is a nonlinear activation function (e.g., Rectified Linear Unit (ReLU), sigmoid, or tanh). The parameters w_i and b are initially randomised and optimised by many iterations to minimise prediction error.

For computational efficiency, neuron operations are vectorised across layers. A layer containing m neurons transforms an input vector $\mathbf{x} \in \mathbb{R}^n$ into an output vector $\mathbf{a} \in \mathbb{R}^m$ through:

$$\mathbf{a} = f(\mathbf{W}\mathbf{x} + \mathbf{b}) \tag{2.7}$$

where $\mathbf{W} \in \mathbb{R}^{m \times n}$ is the weight matrix, $\mathbf{b} \in \mathbb{R}^m$ is the bias vector, and f applies the activation function element-wise [50]. This operation is visually represented in Figure 2.3, which illustrates a 2-3-1 network architecture where input $\mathbf{x} = [x_1, x_2]^T$ is processed by three hidden neurons to produce activations $\mathbf{a} = [a_1, a_2, a_3]^T$ and is called activation vector, demonstrating the matrix multiplication by $\mathbf{W}$. The purpose of the hidden units is to learn non-linear combinations of the original inputs and is called feature extraction or feature construction [90]. For multi-dimensional inputs such as images, the data is first *flattened* into a vector $\mathbf{x}_{\text{flat}} \in \mathbb{R}^{n_1 \cdot n_2 \cdot \cdots \cdot n_d}$ to conform to this computational structure. However, applying fully connected ANNs to images has practical drawbacks: flattening the image destroys spatial locality and translational structure, and results in a very large number of parameters in the fully connected layers. These issues are discussed in the following subsections.

2.3.7 Universal Approximation Theorem

ANNs were already applied experimentally before the formal theoretical results were established. The Universal Approximation Theorem (UAT) explains why the neural networks, with enough hidden units, can approximate almost any continuous function to any desired degree of accuracy, making them powerful tools for learning complex patterns. It states that for any unknown continuous function $f(x)$, we can construct another function $F(x)$, built as a weighted sum of a nonlinear function $\sigma(w^T x + b)$, that closely approximates $f(x)$ within any desired accuracy ε (see Figure 2.4). Here, x represents the input, while w (weight), b (bias). The function $\sigma : \mathbb{R} \to \mathbb{R}$ is a nonlinear activation function,

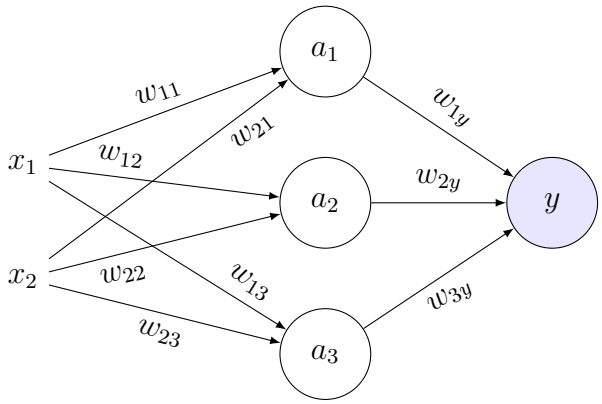


Figure 2.3: Feedforward neural network architecture with two input neurons (x_1, x_2), three hidden neurons (a_1, a_2, a_3), and one output neuron (y). The weights w_{ij} connect input x_i to hidden neuron a_j, while weights w_{jy} connect hidden neuron a_j to output y.

such as a sigmoid or rectified linear unit (ReLU), which introduces non-linearity.

$$\mathrm{ReLU}(x) = \begin{cases} x, & \text{if } x > 0, \\ 0, & \text{if } x \leq 0. \end{cases} \tag{2.8}$$

$$\mathrm{sigmoid}(x) = \frac{1}{1 + e^{-x}} \tag{2.9}$$

This theorem ensures that such functions are dense in the space of continuous functions $C(K)$ under the supremum norm, meaning they can approximate any function on a compact set K. [31] proved UAT for sigmoid activation. [57] generalised it to a broad class of bounded, nonconstant, continuous activation functions. Sigmoid function provides the bounded output between 0 and 1. Later, it was extended to ReLU and other unbounded activations, meaning the function class does not have to be a strict subalgebra. In our work (Chapter 3 & 4), we used both Sigmoid and ReLU to estimate photometric redshifts.

It is important to note, however, that the UAT is an existence result on approximation capacity over compact domains; it does not guarantee an efficient representation, successful optimisation with gradient-based methods, or good generalisation when trained on finite datasets. Practical success in deep learning depends on the architecture of layers, the optimisation properties of overparameterised networks trained with gradient descent, and both explicit and implicit regularisation mechanisms that promote generalisation from finite data.

2.3.8 Training of Neural Networks

A key distinction between traditional regression methods and artificial neural networks lies not in whether the function is predetermined, but in the flexibility and expressivity of

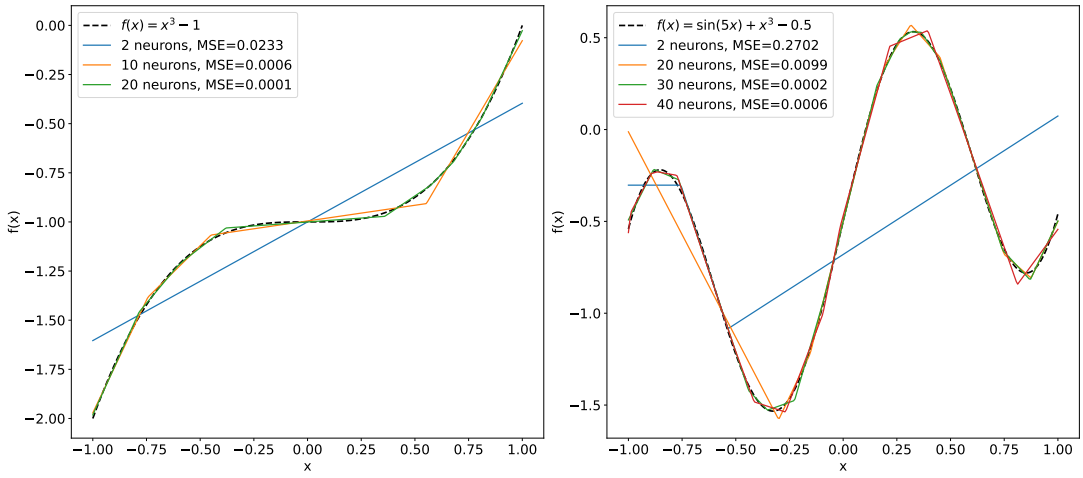


Figure 2.4: Comparison of neural network approximation performance for two different target functions using varying numbers of neurons arranged in a single layer. ReLU is used as an activation function. *left panel:* Approximation of $f(x) = x^3 - 1$ with 2, 10, and 20 neurons, showing progressively lower mean squared error (MSE) as neuron count increases. *right panel:* Approximation of $f(x) = \sin(5x) + x^3 - 0.5$ with 2, 20, 30, and 40 neurons, demonstrating improved accuracy with more neurons, though with a slight MSE increase at 40 neurons, due to overfitting. MSE values are provided for each configuration.

the function family being fitted. In conventional curve fitting, such as linear or polynomial regression, the functional form is fixed in advance (e.g., $y = mx + c$ or $y = ax^2 + bx + c$), and the task is to estimate a small number of coefficients. The hypothesis space is therefore low-dimensional and explicitly specified. ANNs, by contrast, also define a parameterised function family in advance, namely a composition of affine transformations and nonlinear activation functions. However, this family is highly expressive and capable of learning complex, hierarchical representations from data. Rather than fitting a predefined low-order analytic form, training an ANN involves optimising a large set of weights and biases so that the network can learn intermediate representations that enable accurate input-output mappings.

In a multi-layer network, the activation vector propagates sequentially through the network via forward propagation. Mathematically, forward propagation is the process of computing each layer's activations by applying a linear transformation followed by a non-linear activation function. For a network with L layers, the computation at each layer l is:

$$\mathbf{a}^{(l)} = f\left(\mathbf{W}^{(l)}\mathbf{a}^{(l-1)} + \mathbf{b}^{(l)}\right), \quad l = 1, 2, \ldots, L \tag{2.10}$$

where $\mathbf{a}^{(0)} = \mathbf{x}$ denotes the input, and $\mathbf{a}^{(L)} = \mathbf{y}_{\text{pred}}$ represents the final prediction. The weight matrices $\mathbf{W}^{(l)}$ and bias vectors $\mathbf{b}^{(l)}$ are shared parameters learned during training, while the activations $\mathbf{a}^{(l)}$ vary with each input.

Weight initialisation determines the initial values of $\mathbf{W}$ before training and plays a crucial role in effective learning. A suitable initialisation maintains a stable activation and gradient magnitudes across layers. Poor initialisation can result in vanishing or

exploding gradients, slow convergence, or premature saturation of nonlinear activation functions. In our work of photo-z estimation, we adopt the *He normal* initialisation, which is particularly well-suited for layers with ReLU-based activations (For more details, see [52]).

Training aims to optimise the weights and bias by minimising the loss function $\mathcal{L}(y_{\rm pred}, y_{\rm true})$ that quantifies the discrepancy between predictions and ground truth. Total loss is computed over a batch of samples. For regression tasks, the mean squared error (MSE) is commonly employed:

$$\mathcal{L}(y_{\rm true}, y_{\rm pred}) = \frac{1}{N}\sum_{i=1}^{N}\left(y_{{\rm true},i} - y_{{\rm pred},i}\right)^2 \tag{2.11}$$

While MSE strongly penalises large errors due to its quadratic form, it can be highly sensitive to outliers. An alternative is the Huber loss [60], which behaves quadratically for small errors but transitions to a linear form for large deviations. This makes it more robust to outliers while retaining sensitivity near the optimum. The Huber loss per sample is defined in terms of the residual $r = y_{\rm true} - y_{\rm pred}$ as

$$l_\delta(r) = \begin{cases} \frac{1}{2}r^2, & \text{if } |r| \le \delta, \\ \delta\left(|r| - \frac{\delta}{2}\right), & \text{otherwise.} \end{cases} \tag{2.12}$$

The total batch loss is then

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} l_\delta(y_{{\rm true},i} - y_{{\rm pred},i})\,. \tag{2.13}$$

where δ is a tunable threshold that determines the transition between the quadratic and linear regimes. For small residuals, the loss behaves like MSE; for large residuals, it acts like mean absolute error (MAE), reducing the influence of extreme outliers, which is a desirable property when measurement uncertainties or noise contamination can produce occasional large deviations. Huber loss was first introduced to photo-z estimation by [78], and was later adopted in our work [64]. It preserves the sensitivity of MSE near the optimum while adopting a linear penalty for large deviations, resulting in a more robust and stable regression performance.

Optimisation proceeds through iterative cycles of forward propagation to compute predictions, followed by *backpropagation* to calculate gradients of the loss with respect to weight and bias parameters using the multivariable chain rule of partial differentiation. These gradients drive parameter updation via gradient descent:

$$\mathbf{W}^{(l)} \leftarrow \mathbf{W}^{(l)} - \eta \frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(l)}} \tag{2.14}$$

$$\mathbf{b}^{(l)} \leftarrow \mathbf{b}^{(l)} - \eta \frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(l)}} \tag{2.15}$$

where η is the learning rate governing update magnitudes. Parameters are typically updated after processing *batches* of data samples, with one complete pass through the training dataset constituting an *epoch*. Multiple epochs are required to converge to an optimal parameter configuration [50].

2.3.9 Bias-variance tradeoff

The architecture of the network, including the number of layers (depth) and the number of neurons per layer (width), significantly affects learning and can lead to overfit/underfit. It is depicted in Figure 2.5. These can be understood through the bias-variance tradeoff [49], which describes the fundamental tension between model complexity and generalisation performance. In statistical learning theory, the expected prediction error of a model on unseen data can be decomposed into three components:

$$\mathbb{E}\left[(y - \hat{f}(x))^2\right] = \text{Bias}^2 + \text{Variance} + \text{Irreducible noise}. \tag{2.16}$$

Where, y is the label and $\hat{f}(x)$ is the model predicted output.

Bias refers to the error introduced by approximating a complex real-world function with a simplified model. In the context of neural networks, a shallow or narrow architecture may not have sufficient representational capacity to capture the underlying structure of the data. Consequently, such a model exhibits high bias and leads to underfitting, where both training and testing errors remain large.

Variance, on the other hand, measures the sensitivity of the model to fluctuations in the training data. Deep and wide neural networks possess high expressive power and can approximate highly complex functions. However, this flexibility may cause the network to model not only the true underlying signal but also random noise and outliers in the training data. This results in high variance, where training error is low but generalisation error on unseen data is high, leading to overfitting.

Therefore, selecting an appropriate network architecture involves finding a balance between bias and variance. A model that is too simple suffers from high bias, while a model that is too complex suffers from high variance. The optimal model achieves a tradeoff where the total generalisation error is minimised. In practice, techniques such as regularisation [117], dropout [113], and early stopping [100] are employed to control variance without excessively increasing bias. These methods help deep neural networks maintain high representational power while preserving generalisation capability.

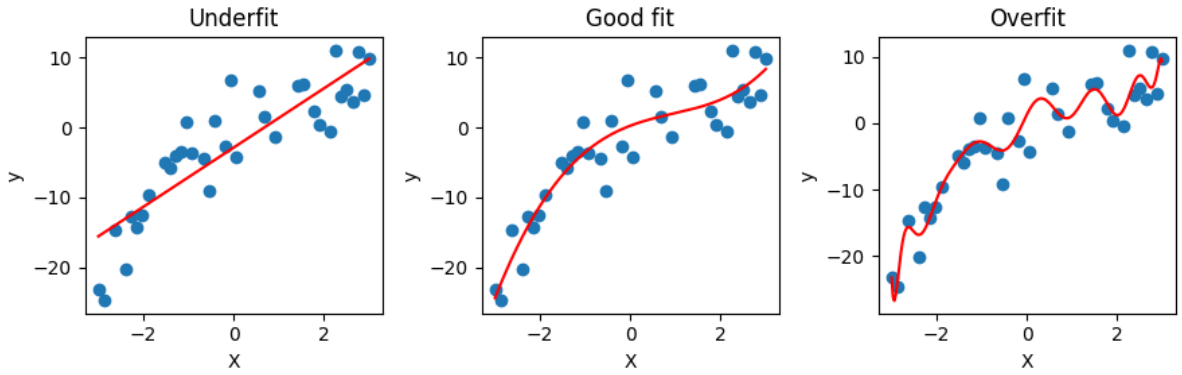


Figure 2.5: Illustration of the bias–variance tradeoff using polynomial regression models of increasing complexity. The left panel shows underfitting, where the model is too simple to capture the nonlinear structure of the data, resulting in high bias. The middle panel demonstrates an appropriate fit that captures the underlying trend while maintaining good generalisation. The right panel exhibits overfitting, where the model is excessively flexible and begins to model noise in the training data, leading to high variance. The figure uses polynomial regression. Increasing the polynomial degree in this example is analogous to increasing the capacity of a neural network (e.g., through greater depth or width), which similarly governs the transition between underfitting and overfitting.

2.3.10 Convolutional Neural Networks

In this PhD work, galaxy images are used as input data alongside tabular data. This section discusses how to learn patterns from images. ANNs can process images by flattening pixels into vectors; this approach becomes computationally and statistically inefficient for images of practical size. The flattening operation required for ANN processing discards spatial relationships inherent in image data. Fully connected networks may memorise patterns in training images but do not encode spatial locality and translation equivariance. Locality refers to the idea that meaningful image features arise from interactions among nearby pixels. For example, detecting an edge in an image only requires information from a small neighborhood of adjacent pixels; distant pixels are largely irrelevant for identifying that local structure. Convolutional layers exploit locality by applying filters over small receptive fields, rather than connecting every pixel to every other pixel as in a fully connected layer. Translation equivariance means that if the input image is shifted in space, the feature map produced by the network shifts in the same way. For instance, if a galaxy appears in the top-left corner of an image and is then shifted to the center, a convolutional network will detect the same feature at the new location. Fully connected networks lack this property because their weights depend on absolute pixel positions; shifting the input changes the pattern of activations in a way that is not simply a shifted version of the original.

For a 36×36 galaxy image with 4 optical bands, flattening results in an input vector of

$$36 \times 36 \times 4 = 5{,}184$$

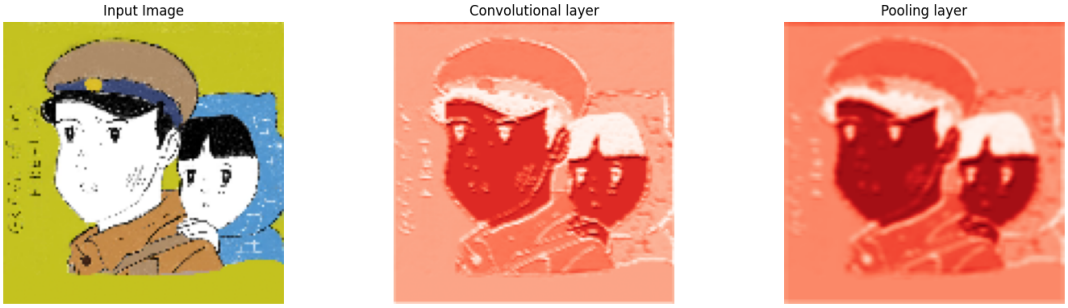


Figure 2.6: Visualisation of feature extraction in a single-layer convolutional neural network. The left panel shows the original input image of size $128 \times 128 \times 3$, where 3 corresponds to Red-Blue-Green (RGB) channels. The middle panel represents the activation map produced by the convolutional layer, highlighting learned spatial features such as edges and intensity transitions. The right panel shows the corresponding feature map after average pooling, where local activations are smoothed, and spatial variations are reduced while preserving the dominant structures. This illustrates how convolution extracts features, and pooling summarises them. The input image is shown in RGB to preserve its original colour information, while the convolution and pooling outputs are displayed in red colour because they represent single-channel (here R channel) activation maps rather than colour images.

dimensions. A subsequent fully-connected layer with 1,000 neurons would require

$$\text{Parameters} = 5{,}184 \times 1{,}000 + 1{,}000 = 5{,}185{,}000.$$

Therefore the main critical issues of using ANN for images are:

- Computational Intractability: The $\mathcal{O}(n^2)$ growth in parameters makes training and inference prohibitively expensive
- Overfitting: With millions of parameters, ANNs easily memorise training data rather than learning generalisable features
- Loss of Spatial Locality: Flattening destroys the natural grid structure and local correlations between adjacent pixels

Combining sparse matrices with weight sharing or parameterisation reduces the effective number of parameters, enabling efficient processing of high-dimensional data. Convolutional Neural Network (CNN) is a special structure which used the combination of sparsity and weight sharing suitable for data such as 2D-images [46, 77]. CNNs use discrete cross-correlation operation, it is called convolution due to historical convention. CNNs have following key mechanisms:

Local connectivity: Unlike fully connected ANNs, where every input unit connects to every neuron in the next layer, convolutional neural networks employ *local receptive fields*, in which small filters scan the input image. For simplicity, consider a single-channel input. The output feature map F_{ij} is given by

$$F_{ij} = \sum_{m=0}^{k_h-1} \sum_{n=0}^{k_w-1} W_{mn}\, X_{i+m,\, j+n} + b, \tag{2.17}$$

where W is a $k_h \times k_w$ kernel, X is the input image, and b is a bias term. Although this operation is commonly referred to as convolution, most CNN implementations perform cross-correlation, meaning that the kernel is not flipped prior to multiplication. [1]

Example of a 2×2 cross-correlation applied to a 3×3 single-channel input, producing a 2×2 output feature map. Each output element is computed as the sum of element-wise products between the kernel and the corresponding local receptive field.

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \star \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} a+2b+3d+4e & b+2c+3e+4f \\ d+2e+3g+4h & e+2f+3h+4i \end{bmatrix}$$

For a multi-channel input with C channels, the operation generalises to

$$F_{ij} = \sum_{c=1}^{C} \sum_{m=0}^{k_h-1} \sum_{n=0}^{k_w-1} W_{mn}^{(c)} X_{i+m,\,j+n}^{(c)} + b, \tag{2.18}$$

where c indexes the input channels.

Parameter sharing: CNNs exploit *weight sharing* across spatial positions, meaning the same filter detects features throughout the entire image. This provides *translation equivariance*, a fundamental property for vision tasks where object identity is independent of position. Weight sharing reduces parameter counts dramatically; a 5×5 filter has only 25 parameters plus bias, regardless of input image size.

Hierarchical feature learning: CNNs automatically learn feature hierarchies through stacked convolutional layers. Early layers detect simple patterns (edges, corners). It is illustrated in Figure 2.6. Intermediate layers combine these into complex features (textures, shapes). Deep layers represent high-level semantic concepts (objects, faces).

After convolution and pooling (see Appendix B), the output of a layer is a three-dimensional tensor of shape $(\text{height}, \text{width}, \text{channels})$, where each channel corresponds to a learned feature map. To produce a final prediction, the feature maps must be aggregated into a compact representation. This can be achieved, for example, by flattening the tensor and passing it to fully connected (dense) layers, or by applying global pooling before a regression head. The dense layers then learn the mapping from the aggregated feature representation to the target quantity. During inference, the trained network generates predictions using the optimised weights and biases learned during training.

$$\mathbf{y}_{\text{pred}} = f(\mathbf{W}\mathbf{x}_{\text{flat}} + \mathbf{b}), \tag{2.19}$$

[1] For the visualisation of the convolution operation, visit this page https://github.com/vdumoulin/conv_arithmetic.

where $\mathbf{x}_{\text{flat}}$ is the flattened feature representation from preceding layers and $f\cdot)$ denotes the output activation function. For classification tasks, f is typically the softmax function.

$$\text{Softmax}(x_i) = \frac{e^{x_i - \max(\mathbf{x})}}{\sum_{j=1}^{K} e^{x_j - \max(\mathbf{x})}}. \tag{2.20}$$

where x_i represents the i-th component of the input vector $\mathbf{x} = (x_1, x_2, \ldots, x_K)$, and K denotes the number of classes. For regression tasks, f is commonly the identity function (or occasionally a bounded linear or sigmoid mapping), producing continuous-valued predictions rather than probabilities distribution over output classes.

Convolution, pooling, and fully connected mappings are all differentiable; CNNs are trained end-to-end using gradient-based optimisation. While backpropagation was introduced in the ANN, here it is important to note that gradients flow not only through dense layers, but also through convolutional filters, bias terms, and even pooling operations (via gradient routing in max positions). Thus, the network jointly adapts both low-level and high-level features to optimally solve the task. In summary, these architectural principles laid the groundwork for increasingly sophisticated models such as Google's *Inception* networks [115], which extend CNN design through parallel multi-scale convolutions, enabling richer and more efficient feature extraction in deep vision systems.

2.4 Summary

In this chapter, the theoretical and methodological foundations of photometric redshift estimation were introduced. Redshift measurements are essential for reconstructing the three-dimensional distribution of galaxies and quasars and for studying the large-scale structure of the Universe. Although spectroscopic redshifts provide highly precise measurements, obtaining them for the extremely large samples produced by modern imaging surveys is observationally expensive. Photometric redshift techniques therefore offer a scalable alternative by estimating redshifts from broadband photometric observations.

The chapter began by reviewing the relationship between redshift, luminosity, and distance, emphasising how redshift measurements enable the reconstruction of cosmic structure. The distinction between spectroscopic and photometric redshifts was then discussed, highlighting the trade-off between precision and survey coverage that motivates the development of robust photo-z estimation methods.

Main classes of photo-z estimation techniques were presented. First, template-based methods estimate redshifts by comparing observed photometric measurements with theoretical or empirical galaxy spectral energy distribution (SED) templates. Second, data-driven methods rely on statistical relationships learned from spectroscopic training samples. Finally, the chapter focused on machine learning approaches, which are capable of modelling complex nonlinear relationships between photometric observables and redshift.

Within this context, the fundamental concepts of supervised learning, neural network training, the bias-variance tradeoff, and the universal approximation theorem were discussed. Particular attention was given to artificial neural networks and convolutional neural networks, which are well suited for extracting information from multi-band photometric data and galaxy images.

The concepts presented in this chapter provide the methodological foundation for the analyses developed in the subsequent chapters of this thesis. In Chapter 3 & 4, these principles are applied to develop and evaluate the `Hybrid-z` approach, which combines deep learning techniques with photometric data to improve redshift estimates for galaxies in the Kilo-Degree Survey (KiDS) bright galaxy sample. Accurate photo-zs obtained through this framework are then used to construct reliable galaxy and quasar samples for clustering analyses.

In particular, the improved redshift estimates are essential for the cosmological analyses presented later in this thesis. Chapter 4 investigates the angular clustering and bias of photometric quasars in KiDS-DR4, while Chapter 5 explores the statistical properties of galaxy distributions through Count-in-Cells measurements. Photo-z estimation therefore forms a crucial link between observational survey data and the statistical characterisation of large-scale structure.

Part III

Hybrid-z: Enhancing the Kilo-Degree Survey bright galaxy sample photometric redshifts with deep learning

The published work [64] is attached according to the copyright of the Astronomy & Astrophysics journal[1].

3.1 Introduction

As discussed in Chapter.2, we employed deep learning to estimate photometric redshifts for the KiDS-DR4 Bright Galaxy Sample using both galaxy images and magnitudes. A convolutional neural network extracts morphological and structural features from the four-band ($ugri$) KiDS image cutouts, while a fully connected neural network (ANN) processes nine-band ($ugriZYJHK$s) magnitudes, capturing complementary colour and SED information.

We introduce `Hybrid-z`, a framework that combines CNN-derived image features with ANN-based magnitude features, achieving a 20% reduction in scatter relative to `ANNz2` [104], which uses only magnitudes. The model is trained on spectroscopic redshifts from the Galaxy And Mass Assembly (GAMA) survey. This chapter details the network architecture, training and optimisation strategy, validation on independent datasets, and application to the full KiDS-DR4 Bright sample. The `Hybrid-z` package is publicly available and can be installed directly from GitHub[2], and the resulting photo-z catalog of KiDS-DR4 bright galaxy sample is accessible via the CDS repository[3].

[1] https://www.aanda.org/for-authors/author-information/copyright
[2] https://github.com/Anjithajm/Hybrid-z/tree/main
[3] https://cdsarc.cds.unistra.fr/viz-bin/cat/J/A+A/698/A276

A&A, 698, A276 (2025)
https://doi.org/10.1051/0004-6361/202453576

Astronomy & Astrophysics

Hybrid-z: Enhancing the Kilo-Degree Survey bright galaxy sample photometric redshifts with deep learning

Anjitha John William[1,⋆], Priyanka Jalan[1], Maciej Bilicki[1,⋆], Wojciech A. Hellwing[1], Hareesh Thuruthipilly[2], and Szymon J. Nakoneczny[3]

[1] Center for Theoretical Physics, Polish Academy of Sciences, al. Lotników 32/46, 02-668 Warsaw, Poland
[2] National Centre for Nuclear Research (NCBJ), ul. Pasteura 7, 02-093 Warsaw, Poland
[3] Division of Physics, Mathematics and Astronomy, California Institute of Technology, 1200 E California Blvd, Pasadena, CA 91125, USA

Received 21 December 2024 / Accepted 5 May 2025

ABSTRACT

We employed deep learning to improve the photometric redshifts (photo-zs) in the Kilo-Degree Survey Data Release 4 bright galaxy sample (KiDS-DR4 Bright). This dataset, used as foreground for KiDS lensing and clustering studies, is flux-limited to $r < 20$ mag with mean $z = 0.23$ and covers 1000 deg^2. Its photo-zs were previously derived with artificial neural networks from the ANNz2 package trained on the Galaxy And Mass Assembly (GAMA) spectroscopy. Here, we considerably improve on these previous redshift estimations by building a deep learning model, Hybrid-z, that combines an inception-based convolutional neural network operating on four-band KiDS images with an artificial neural network using nine-band magnitudes from KiDS+VIKING. The Hybrid-z framework provides state-of-the-art photo-zs for KiDS-Bright with negligible mean residuals of $O(10^{-4})$ and scatter at a level of $0.014(1 + z)$ – representing a reduction of 20% compared to the previous nine-band derivations with ANNz2. Our photo-zs are robust and stable independently of galaxy magnitude, redshift, and color. In fact, for blue galaxies, which typically have more pronounced morphological features, Hybrid-z provides a larger improvement over ANNz2 than for red galaxies. We checked our photo-z model performance on test data drawn from GAMA as well as from other KiDS-overlapping wide-angle spectroscopic surveys, namely SDSS, 2dFLenS, and 2dFGRS. We found stable behavior and consistent improvement over ANNz2 throughout. Finally, we applied Hybrid-z trained on GAMA to the entire KiDS-Bright DR4 sample of 1.2 million galaxies. For these final predictions, we designed a method of smoothing the input redshift distribution of the training set in order to avoid propagation of features present in GAMA related to its small sky area and large-scale structure imprint in its fields. Our work paves the way toward the best-possible photo-zs achievable with machine learning for any galaxy type for both the final KiDS-Bright DR5 data and for future deeper imaging, such as from the Legacy Survey of Space and Time.

Key words. techniques: miscellaneous – catalogs – surveys – galaxies: distances and redshifts – galaxies: photometry – cosmology: observations

1. Introduction

Redshift is a key quantity for cosmological analyses. As the basic proxy for galaxy distances, it allows one to map the large-scale structure of the Universe in time and three-dimensional space. Redshift can be measured to a sub-percent accuracy only via spectroscopy. For such spectroscopic redshifts (spec-zs), one first collects the spectrum of an object and then identifies the shift in spectral lines with respect to the rest frame. However, even in the current era of fast measurements, such as with the Dark Energy Spectroscopic Instrument (DESI Collaboration 2016), spec-zs can be obtained only for a small fraction of all detected galaxies.

On the other hand, redshifts can be estimated for a much larger sample from photometric measurements. Such photometry-based redshifts (photo-z) provide an alternative method to the spec-zs and are based on the correlation between redshift and apparent galaxy magnitudes. This approach was originally pointed out by Baum (1957) and first applied to obtain photoelectric magnitudes in nine passbands by Baum (1962). The two main attractive factors of photo-zs are that they allow one to obtain redshifts for galaxies fainter than generally possible with spectroscopy and that the number of objects with redshift estimates per unit telescope time is also much larger (Hildebrandt et al. 2010). Notably, photo-zs cannot provide redshift accuracy and precision as good as spec-zs. Nevertheless, they are indispensable in today's massive imaging surveys cataloging millions and billions of galaxies.

Photo-zs are based on a complicated mapping from photometry to redshift space, and they are difficult to handle analytically since this mapping depends on observational, computational, and statistical factors. Photo-z estimation methods can be generally categorized into template fitting and empirical approaches. In the latter case, which is our focus in this paper, the relation between photometric quantities and redshift is very commonly found by machine learning (ML) algorithms, although we note that simpler approaches of using functional fitting have also been proposed (e.g., Connolly et al. 1995; Krone-Martins et al. 2014).

In supervised ML techniques, the algorithm derives the empirical relation between observed quantities and labels from appropriate training on labeled data. Therefore, their main challenge and limitation lies in extrapolating the results beyond a representative training set. However, if such

⋆ Corresponding authors: anjithajm@cft.edu.pl, bilicki@cft.edu.pl

appropriate training data exist, ML methods can excel, and this has led to numerous approaches being proposed for photo-zs based on the newest developments in computer science. Some examples include support vector machines (Wadadekar 2004), random forest (Carliles et al. 2010; Li et al. 2021), artificial neural networks (ANNs; e.g., Tagliaferri et al. 2003; Collister & Lahav 2004; Oyaizu et al. 2008), ensemble learning (Cunha & Humphrey 2022), Gaussian processes (Way & Srivastava 2006; Bonfield et al. 2010), self-organizing maps (Way & Klose 2012), k-nearest neighbors (Graham et al. 2017), mixture density networks (Ansari et al. 2021), and finally deep neural networks (e.g., Hoyle 2016; D'Isanto & Polsterer 2018).

Among the various supervised ML techniques for photo-z derivation, deep learning (DL) has emerged as a particularly promising one. Deep learning makes it possible to entirely skip "higher-level" quantities such as galaxy magnitudes or sizes derived from photometric post-processing and build the ML model using multi-band imaging directly. In such frameworks, "deep" indicates that the models usually have compounded multi-layer architectures. Their usage for photo-zs was pioneered by Hoyle (2016) and then studied by, for example, D'Isanto & Polsterer (2018), Menou (2019), Pasquet et al. (2019), Dey et al. (2022), Henghes et al. (2022), Treyer et al. (2024) for Sloan Digital Sky Survey (SDSS) data; Schuldt et al. (2021) for Hyper-Suprime Cam (HSC); Li et al. (2022) for the Kilo-Degree Survey (KiDS); and by Roster et al. (2024) for DESI Imaging. These papers have demonstrated that using images directly, or in combination with magnitudes (e.g., Li et al. 2022; Jones et al. 2024; Roster et al. 2024), allows one to derive photo-zs of better performance than those based on tabular galaxy data such as magnitudes.

A particular realization of DL is convolutional neural networks (CNNs). They are a type of ANN suitable for computer vision problems, such as image feature detection or classification, and are inspired by the human vision system. Similar to real neurons, which receive input and pass electrochemical signals (McCulloch 1943), artificial neurons are used in CNNs. The significance of CNNs is evidenced by their vast use in image recognition tasks (e.g., LeCun et al. 1998). These networks are appropriate for processing data that have grid-like topology, such as images, because of their local connectivity, parameter sharing, and translational invariance. We chose CNNs to extract the galaxy image patterns by detecting features such as edges, textures, and shapes, which are expected to improve photo-z derivations compared to methods that do not employ such information.

In this paper, we present a photo-z estimation in which we incorporate both fluxes (magnitudes) and multichannel galaxy images by using DL techniques for KiDS (de Jong et al. 2013). KiDS is a multiband imaging survey covering about 1350 deg^2 of the sky, of which we employ ∼1000 deg^2 from its fourth data release (Kuijken et al. 2019). The DL photo-zs within KiDS have already been studied in detail by Li et al. (2022), where various setups using both images and magnitudes were compared for a general selection of galaxies spanning $0 < z \lesssim 3$. However, until now, such imaging-based photo-z approaches have not been employed in KiDS in the regime where they are expected to bring the most improvement over "shallow" ML, namely, for relatively bright and well-resolved galaxies. Here, we fill this gap and focus only on the bright end of the KiDS data.

We studied the performance of DL for photo-zs in the flux-limited "KiDS-Bright DR4 sample," which includes all the KiDS galaxies within the magnitude cut of $r < 20$ mag (Bilicki et al. 2018, 2021, hereafter B18, B21). This sample is particularly useful for such an analysis, as by design it is selected to mimic the spectroscopic Galaxy And Mass Assembly dataset (GAMA; Driver et al. 2011). As GAMA is flux limited and highly complete spectroscopically, it constitutes a very well matched training set for empirical photo-z models. In the DL context, this aspect has been taken advantage of in the recent work by Treyer et al. (2024), where photo-z derivation for a sample of SDSS galaxies at $r < 20$ mag was presented. The previous KiDS analysis by Li et al. (2022) at the low-redshift end using DL as well as the dedicated studies by B18, B21 with "shallow" ML and GAMA training gave state-of-the-art photo-z results for the respective selections in KiDS. Therefore, here we aim to build on and extend these previous successful endeavors. Among our goals is to check if the current KiDS-Bright DR4 sample redshift estimates could be further improved. This is relevant, for instance, in view of the forthcoming Legacy Survey of Space and Time (LSST Science Collaboration 2009), where high-resolution multi-wavelength imaging of a depth greater than in KiDS will be available for the entire southern sky. Improvements in photo-z accuracy and precision of foreground galaxies are important in this context, as they help minimize related systematics of photometric clustering and lensing analyses.

The default photo-zs in KiDS data releases are derived with the Bayesian photometric redshifts (BPZ; Benítez 2000) template-fitting tool. In particular, this tool is used to bin the weak lensing sources in redshift shells, and their true redshift distribution is then calibrated with self-organizing maps and via the clustering redshift technique (Hildebrandt et al. 2021). However, several studies have demonstrated that for bright low-redshift galaxies, empirical photo-z methods can outperform the default BPZ solution when selecting galaxy samples from KiDS (e.g., Cavuoti et al. 2015; Bilicki et al. 2018; Vakili et al. 2019; Li et al. 2022). This is possible if relevant training or calibration data are available to build a model mapping the photometric space to redshift using ML, but also other approaches such as red-sequence fitting (Rozo et al. 2016; Vakili et al. 2019, 2023) can be used as well. For the KiDS-DR4 Bright galaxy sample (B21) we are concerned with in this paper, the redshifts were estimated using ANNs from the public package ANNz2 (Sadeh et al. 2016). The ANNs used photometric quantities (magnitudes, colors) as input and were trained with spectroscopic redshifts from GAMA. A number of tests have shown that such photo-zs are statistically accurate and precise (i.e., have low mean bias and scatter; B18; B21) not only for the KiDS-DR4 Bright sample as a whole but also for the sub-populations such as red and blue galaxies. In particular, this was possible thanks to the already mentioned intentional very good match of the galaxy selection in the KiDS-Bright DR4 sample to the GAMA training set.

In this work, we extend the previous feature-based ML efforts to build a successful DL model that we call "Hybrid-z." The model integrates both images and features for KiDS-DR4 Bright photo-z estimations. We used the same GAMA training set as in B21 and constructed a photo-z model that is conceptually similar to one tested in Li et al. (2022). Namely, it combines a deep convolutional network, employing *ugri* imaging, with an ordinary ANN that is fed by nine-band galaxy magnitudes. An analogous configuration was also studied by Henghes et al. (2022) for SDSS, and inspired by their results, we also use "inception" as our basic architecture for the DL part.

This paper is organized in the following manner. In Sect. 2 we describe the data used. Next, in Sect. 3 we explain the basic concepts of CNNs and the special CNN architecture that we

used, called inception. In Sect. 4, we describe our Hybrid-z model to estimate the photo-zs and the statistics that we used to measure the performance of the network. Sect. 5 presents our results, and in Sect. 6, we conclude and discuss future prospects.

2. Data

In this section, we discuss the data used in this study. We employed the KiDS-Bright DR4 sample images supplemented with photometry (i.e., magnitudes). The required training and testing data are labeled using the spectroscopic redshifts from the GAMA survey.

2.1. KiDS images and photometry

Kilo-degree survey is an optical wide-field imaging survey of the European Southern Observatory (ESO) at the Very Large Telescope (VLT) Survey (VST, Capaccioli & Schipani 2011), having at the focal plane a 268 million pixel Charge-Coupled Device (CCD) mosaic camera called OmegaCAM (Kuijken 2008). VST is an alt-az mounted modified Ritchey-Cretien telescope located in the ESO Paranal Observatory, Chile. The images were taken in four broad bands (*ugri*), and the survey covers 1350 square degrees of the extragalactic sky. The final footprint of the survey is shown in Fig. 3 of Wright et al. (2024); here we use its publicly available subset.

KiDS-ESO Data Release 4 (KiDS DR4; (Kuijken et al. 2019)), is the fourth public release of KiDS. The ASTRO-WISE optical pipeline and data reduction environment (McFarland et al. 2013) is used to produce stacked (or co-added) composite images created by combining multiple individual exposures of the same sky area, one in each of the four bands. As a result, KiDS DR4 optical data are organized into 4×1006 one square-degree tiles.

KiDS DR4 products consist of astrometrically and photometrically calibrated co-added images with a uniform pixel scale of 0.2 arcsec. The pixel units are fluxes relative to the 0th magnitude (de Jong et al. 2015). We have downloaded the KiDS DR4 tiles[1] and made cutouts of galaxies with a size of $7.2'' \times 7.2''$ (36×36 pixels), since most of the objects we use are smaller than this. In particular, this cutout size is above the 99-percentile level of the half-light diameter (i.e., $2 \times$ `FLUX_RADIUS`) in KiDS-Bright. The largest galaxies, not fitting within our cutouts, are likely of little interest for our work anyway: they will be very nearby and will have had spec-z from wide-angle surveys or otherwise will not be useful for lensing studies.

We have also tried bigger cutouts such as $20.2'' \times 20.2''$ and $8'' \times 8''$, motivated receptively by Grespan et al. (2024) and Li et al. (2022), but our finally adopted size gave the best results. The smaller size reduces the noise in images without losing galaxy flux information. Too large cutouts could also lead to frequent situations when more than one galaxy appears in the image. This type of contamination could adversely affect the performance of the model, although it has been argued in the literature that CNNs for photo-z estimation could in fact benefit from physically close pairs in the images (e.g., Pasquet et al. 2019). In our case we also use magnitudes of individual galaxies, which should mitigate this effect, be it positive or negative. In any case, for our fiducial cutout size, more than one object is present in the image very rarely, in less than 1% cases.

Finally, we normalized the pixel values to the range $[0, 1]$ galaxy-wise, i.e. jointly for all the *ugri* bands for a given galaxy cutout. The normalization formula is

$$X_{\text{normalized}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}. \quad (1)$$

The minimum value of the pixels for the 4-band per-galaxy images (cutouts) is denoted as X_{min} and the maximum value as X_{max}. In occasional cases when the image has (a) saturated pixel(s), this normalization will not work properly. Dealing with this problem would be beyond the scope of our work as it would require building an extra framework to pre-analyze all the 4.8 million cutouts and cleaning them up of such artifacts. We note that in most cases, objects located in such corrupted areas will not be useful for science anyway, as they will bear an appropriate KiDS `MASK` value indicating that their photometry is not reliable (see next Section). According to our estimates, for the 'clean' data (i.e., those not affected by the mask), the presence of artifacts in the images is very infrequent, at a fraction of a percent level. Last but not least, the usage of magnitudes together with images in the model, will minimize their influence on the derived photo-zs.

KiDS DR4 photometric data consists of optical *ugri* and near-infrared (NIR) data from the VISTA Kilo-degree INfrared Galaxy survey (VIKING, Edge et al. 2013) with observations in five bands: $ZYJHK_s$. In KiDS the magnitudes are by default derived with the Gaussian Aperture And Point spread function (GAAP, Kuijken 2008) methodology. GAAP magnitudes are meant to provide robust galaxy colors irrespective of PSF differences in various bands, which makes them optimal for photo-z derivation. This is an important asset for general weak lensing tomographic studies (Kuijken et al. 2015). GAAP magnitudes were also shown to be optimal for low-redshift photo-z estimates in KiDS, as compared to other galaxy magnitude measurements available in this survey (B18). The GAAP magnitudes in KiDS are provided in the AB system and their zero-point calibration is achieved by using coadd overlaps and stellar locus regression. They are corrected for Galactic extinction (Schlegel et al. 1998) $E(B-V)$ map with (Schlafly & Finkbeiner 2011) coefficients.

We standardized these 9-band magnitude features using the StandardScaler class from the `scikit-learn` Python library (Pedregosa et al. 2011), which computes the mean ($\overline{m}$) and standard deviation (σ_m) for each band independently. The magnitude values (m) in each band are transformed as

$$m_{\text{standardized}} = \frac{m - \overline{m}}{\sigma_m}. \quad (2)$$

In comparison to the optical *ugri* images from KiDS, "coadds" are not readily available for the VIKING NIR, as these data are not processed by the ASTRO-WISE pipeline. Because of this, here we employ only the 4-band optical images, leaving the possible extension with NIR imaging to future work.

2.2. KiDS-DR4 Bright galaxy sample

In this work, we derive DL photo-zs for the KiDS-DR4 Bright galaxy sample, introduced in B21. This dataset contains galaxies selected from KiDS DR4 with the flux limit $r_{\text{auto}} < 20$ mag, where "auto" stands for `SExtractor`-derived estimate of the total flux via automatic aperture photometry (Bertin & Arnouts 1996). The selection of the sample is designed to have the best possible match with overlapping GAMA-equatorial spectroscopic data, which originally were selected using SDSS Petrosian magnitudes in the r band (Liske et al. 2015). As it was

[1] https://kids.strw.leidenuniv.nl/DR4

discussed in B21, among the various r-band magnitude measurements in KiDS, a cut in $r_{\rm auto}$ provides the best correspondence to the GAMA original $r_{\rm Petro} < 19.8$ limit.

In addition to the flux limit, the KiDS-Bright dataset uses various flags from the KiDS input catalog to remove point sources (stars, quasars) and artifacts. These in particular were `CLASS_STAR` < 0.5 & `SG2DPHOT` $= 0$ & `SG_FLAG` $= 1$ to select extended sources and `IMAFLAGS_ISO` $= 0$ & (`MASK`&28668) > 0 to remove imaging artifacts. However, we note that the published KiDS-Bright catalog also includes the objects without this final criterion applied, and in Sect. 5.1 we test how the masking affects our photo-z derivation.

2.3. GAMA spectroscopic data

In the training and testing phase, we used KiDS galaxies that have counterparts in the GAMA catalog (Driver et al. 2011), giving us the true redshift labels. GAMA is a multi-wavelength and spectroscopic survey in five fields[2] (three equatorial: G09, G12, and G15, and two southern ones: G02 and G23), with a total ∼286 deg^2 area. Spectra were collected using the AAOmega fiber-fed spectrograph facility on the 3.9-m Anglo-Australian Telescope. GAMA provides spectra, redshifts, their quality marks, and other ancillary information.

KiDS DR4 fully overlaps with four GAMA fields (all but G02). Of these, the equatorial ones present the largest flux-limited spectroscopic completeness, originally estimated as 98.5% at $r_{\rm SDSS} < 19.8$ mag (Liske et al. 2015) but subsequently revised to 98% at $r_{\rm KiDS} < 19.58$ (Driver et al. 2022) after ingestion of KiDS photometry into the GAMA database (Bellstedt et al. 2020). Following the results of B21, where the addition of the shallower and less complete G23 data did not lead to improvement in photo-z estimates, also here we used only the equatorial fields to train the model, and it is tested on both equatorial and G23 fields. Similarly, other datasets such as for instance SDSS or DESI Early Data Release, are not sufficiently complete at our flux limit of $r < 20$ mag to provide useful training sets for the overall galaxy sample (see, e.g., Jalan et al. 2024). Some of them, however, are useful for a posteriori tests of our photo-z model, and this is discussed in Sect. 5.2.

In this work, we employ the GAMA-II spectroscopic redshift catalog of galaxies from the final GAMA DR4 (Driver et al. 2022)[3]. For secure redshifts, we select galaxies with a normalized quality parameter $NQ \geq 3$ (see Liske et al. 2015, for details) and $z > 0.001$. We have identified galaxies within KiDS tiles that are shared between the GAMA and KiDS-Bright samples based on their right ascension and declination coordinates, assuming a 1″ matching radius between KiDS and GAMA. Then we made cutouts of size 7.2″×7.2″ from KiDS images with these galaxies positioned at the center.

3. Methodology

3.1. Convolutional neural networks

A CNN is a sequence of layers: convolutional, pooling, and fully connected layers. The convolution takes place between input images and a small matrix of weights called a kernel or filter. Initially, the weights are small random values. In a CNN, the learning is hierarchical. The first convolutional layers are responsible for extracting the lower-level features of images, such as edges, corners, and textures (Goodfellow et al. 2016). The kernels adjust their weights to extract these features. The filters in the following layers help integrate more complex features of input data. The network is trained by non-linear optimization of weights and biases through a gradient-descent algorithm. Input data is transformed by convolution operation, producing linear activation.

The relationship between input data and output labels is usually non-linear. The linear relations are limited for this complex mapping from input to output spaces. To introduce non-linearity, the convoluted output passes through a non-linear function in hidden and output layers. This decides which neurons should activate. Without using the activation function, the output would be linearly dependent on the input, since the convolution output is a linear combination of input pixels within the receptive field. Rectified Linear unit (ReLu) and its variants such as peaky ReLu and leaky ReLu, hyperbolic tangent, and sigmoid functions are the common activation functions (Jentzen et al. 2023). The output of CNN is multidimensional (a tensor) and referred to as a feature map; it is another representation of the input data.

The next stage is the pooling operation to modify the output. The pooling function replaces the output of the network at a certain location with a summary statistic of the nearby outputs (Goodfellow et al. 2016). We experimented with various pooling operations, including average pooling and max pooling (Gholamalinezhad & Khosravi 2020). Based on evaluation metrics, we found that average pooling enhances model performance, which might be thanks to the fact that it deals better with noise in the images than max pooling. Average pooling computes the average of a rectangular neighborhood.

3.2. Training

The training of a neural network involves several key steps and considerations. Initially, the dataset is divided into three subsets in a random manner using `scikit-learn` python library: training, validation, and testing set, in our case in a ratio of 70:15:15. The validation set serves to monitor the performance of the network during the training process. Also, there is no selection bias in magnitudes or redshifts in any subset. The final cutout catalog contains ∼173k galaxies in the equatorial fields and ∼125k of these are used to train the model. The model is validated on ∼26k galaxies in the equatorial fields, and the remaining data are used for testing. Finally, it is applied to the entire KiDS-DR4 Bright sample of about 1.2 million galaxies.

During training, the network updates its kernel values based on feedback signals. This iterative process aims to minimize the loss function, which represents the discrepancy between the predicted values and the true value. In the case of a CNN, the loss function is dependent on the weights. The weights are multiplied by the input values during the forward pass of the network to produce the output.

To minimize the loss function and approach the global minimum, the weights need to be updated in the opposite direction of the gradient. This optimization process involves gradient-based techniques. Thelearning rate, a crucial hyperparameter, determines the size of the step taken along the negative gradient direction. A too-small learning rate can result in slow convergence, while a too-large one can lead to divergent behavior of loss function.

The training process iterates over the dataset in batches, where the number of iterations is referred to as epochs, and the number of samples per batch is the batch size. At each epoch, the network computes the gradients of the weights

[2] See `https://www.astro.ljmu.ac.uk/~ikb/research/gama_fields/` for the GAMA field locations.
[3] `http://www.gama-survey.org/dr4`

with respect to the loss on the batch and updates the weights (Chollet 2017).

3.3. Data augmentation

Optimization and generalization play important roles in ML problems. Optimization is the process of adjusting the model parameters to get the best performance on training data, while generalization determines the model performance on unseen data. If the ML model is too simple to capture the underlying structure of the data, it performs poorly not only on the training data but also on unseen data. This is because it fails to learn relevant patterns from the training, leading to underfitting. On the other hand, overfitting happens when a model learns not only the underlying patterns but also the noise and random fluctuations present in the training (Xia 2024). As a result, it performs very well on the training data but poorly on unseen data because it has essentially memorized the training data instead of generalizing from it.

One of the solutions for over- and underfitting is generating a large training dataset using data augmentation. Data augmentation is commonly used in ML, particularly in the context of image classification and computer vision tasks (Shorten & Khoshgoftaar 2019). It involves generating additional training data by applying various transformations (rotation, width and height shift, flipping, etc.) to the existing training samples. After learning a certain pattern in an image, a CNN can recognize it anywhere due to its translational invariance.

We experimented with various data augmentation approaches in the training sample images and selected those that were most efficient for our case. We applied 5% shifts in both width and height. Also, we performed horizontal and vertical flipping on the images. By these four variations, we extended our training set to ∼484k objects and were able to considerably improve model performance. We also tried image rotations, but these did not lead to any improvements.

3.4. Inception module

In this section, we discuss a CNN architecture referred to as "inception" that we employ in photo-z prediction. Increasing the depth of the network improves the performance at a cost of high computation time. Inception, a deep CNN architecture developed for image detection and classification tasks. It was first used in GoogLeNet architecture (Szegedy et al. 2015). The smaller number of weights, and biases compared to its predecessors reduces the computation time and makes it appropriate for big-data handling. There are various versions of Inception available, differing by the use of regularization, reduction of overfitting when the training sample is limited, and the inclusion of additional DL architectures such as Resnet (He et al. 2015) in the Inception module. In our model, we used the Inception v1 module illustrated in Fig. 1. We also tested other versions and we found that Inception v1 gives the best performance for our dataset.

Choosing the right kernel is very important for feature extraction. On the one hand, a larger kernel is suitable for images where the information is distributed globally. However, a smaller-sized kernel is good for locally distributed information extraction. This selection becomes highly important as the galaxies in our sample have a wide range of apparent sizes, and hence our dataset contains both kinds of distributions. We used both larger and smaller-sized kernels in the Inception module. This is shown in Fig. 1.

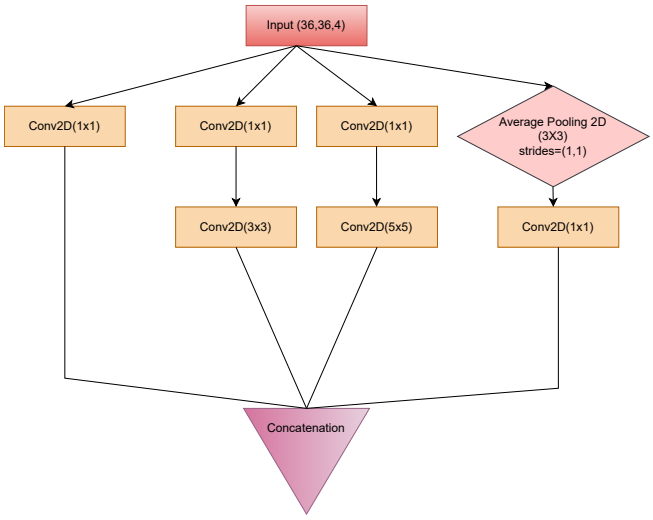


Fig. 1. Inception module used in this study to estimate the photo-z. The input layer has size (36,36,4). Conv2D is a two-dimensional convolution layer, with the kernel size specified in brackets. Average Pooling 2D is a two-dimensional pooling operation that uses a kernel of size 3×3. Each operation is represented using boxes of distinct colors. Concatenation is the combined feature maps from parallel convolutions.

The most straightforward way of improving the performance of deep neural networks is by increasing the complexity. This includes both increasing the depth (number of layers) of the network and its width: the number of units at each layer. The inception module uses parallel convolution operation with multiple filter sizes. Inception v1 uses 1×1, 3×3, and 5×5 spatial filters.

The concatenation process combines all the feature maps from the parallel convolutions. This merging takes place along the channel axis (depth-wise concatenation) which enables the network to efficiently process and extract information from complex visual data. Based on prior research in photo-z estimation, such as Henghes et al. (2022), Li et al. (2022), and the performance of our model, we decided to include Inception in our framework.

3.5. Metrics

In this Section, we discuss the metrics to evaluate the performance of our model and photo-zs. Here we discriminate the two types of metrics into those which are evaluated and optimized when building the model (during training and validation), and those computed a posteriori to quantify the behavior of the photo-zs. For the former, we mainly used three metrics: mean squared error (MSE), mean absolute error (MAE), and the R squared error (R^2). The MSE and MAE are defined as follows:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(z_i - \hat{z}_i)^2, \tag{3}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|z_i - \hat{z}_i|. \tag{4}$$

Here, n is the number of samples used for training, z_i is the predicted value and $\hat{z}_i$ is the true value. The R^2 is a statistical measure that shows how well a model predicts the outcome in a regression analysis. It is the proportion between the variance explained by the model and the total variance:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(z_i - \hat{z}_i)^2}{\sum_{i=1}^{n}(z_i - \bar{z})^2}. \tag{5}$$

Here, $\bar{z}$ is the mean of the spectroscopic redshift values from the GAMA×KiDS crossmatch. For an ideal case, the value of this metric would be unity.

The MSE is more sensitive to outliers than MAE. Due to the squaring part, MSE puts more weight on larger errors. MAE, on the other hand, treats all errors with equal importance, which can be advantageous when the dataset contains extreme values or noisy observations – which is common for galaxy images. Therefore, MSE is more appropriate for learning outliers, and MAE is better for ignoring them. To include the benefits of both these loss functions, we use the Huber (1964) loss function, similarly as in the previous KiDS analysis by Li et al. (2022). This function is quadratic when the absolute error is small, and linear when the absolute error exceeds a threshold δ. This makes it robust to outliers because the impact of large errors is reduced compared to using a purely quadratic loss function such as MSE. The Huber loss is defined as

$$L(z,\hat{z}) = \begin{cases} \frac{1}{2}(z-\hat{z})^2, & \text{if } |z-\hat{z}| \le \delta, \\ \delta(|z-\hat{z}| - \frac{\delta}{2}), & \text{otherwise.} \end{cases} \quad (6)$$

A very low δ value means that the transition region between the quadratic and linear parts of the loss function is extremely narrow. As a result, even points that are not true outliers but are slightly distant from the predicted values may significantly impact the loss. This can lead to overfitting to outliers. A high value ($\delta > 0.01$ for our case) makes the loss function less sensitive to outliers. These will affect the model generalization. We tried a range of values between 10^{-5} and 0.01, and found that $\delta = 0.001$ shows the lowest values for Huber loss, MSE, and MAE. Thus we chose this value of δ for our models.

We evaluated the resulting photo-z performance using the following standard statistics:

– Bias,

$$\delta z = z_{\rm phot} - z_{\rm spec}; \quad (7)$$

– Normalized (rescaled) bias,

$$\Delta z = \frac{\delta z}{1 + z_{\rm spec}}; \quad (8)$$

– Standard deviation of normalized bias, $\sigma_{\Delta z}$;
– Scaled median absolute deviation (SMAD) of Δz, where

$$\mathrm{SMAD}(x) = 1.4826 \times \mathrm{median}(|x - \mathrm{median}(x)|). \quad (9)$$

The first two of these metrics quantify the average residuals of the photo-zs from the true value, i.e. their statistical accuracy. The two others measure the scatter, i.e. statistical precision. The factor 1.4826 in the SMAD definition allows it to converge to one standard deviation for the Gaussian.

4. Photometric redshift model

Here, we explain the photometric redshift model, Hybrid-z, used in this study. It incorporates two types of input: galaxy images and magnitudes.

The model was trained using a dataset comprising $173k \times 4$ KiDS galaxy images from the GAMA equatorial fields and with data augmentation techniques applied, as outlined in Sect. 3.3. Together with the 4-band images, also 9-band magnitudes of the same galaxies were used. Then the model was validated with 26k galaxy samples and tested on an additional set of 26k galaxies also from GAMA equatorial × KiDS. Finally, we estimated photo-zs for all the KiDS-DR4 Bright sample galaxies.

We used the Rectified linear unit (Relu, Jentzen et al. 2023) as the activation function in all the layers except in the output one where the sigmoid function (i.e., logistic curve) is used, which enforces all the predictions to lie in the range $0 < z_{\rm phot} < 1$. For the layers other than the last (output) one, we also tried other activation functions such as leaky Relu (Jentzen et al. 2023), peaky Relu, softmax, swish, and tanh, but Relu gave the best performance.

For the output layer, the sigmoid function gives the best results in our redshift range where practically all the galaxies have $z < 1$ due to the $r < 20$ mag flux limit. For instance, in the entire GAMA spectroscopic sample, objects with $z > 1$ constitute less than 0.1% of the total and these are typically very bright and rare AGNs. For such sparse sources, empirical methods such as ours would not be able to render reliable redshift predictions unless some special approach to anomaly handling is taken. We therefore sacrificed a very small number of objects that lie at true $z > 1$ to have $z_{\rm phot} < 1$ in order to avoid a gross redshift overestimation for others, which could happen if the model had more freedom. The latter is for instance the case for ANNz2, where some of the photo-zs from B21 are predicted significantly above unity. Similar logic applies to photo-zs with non-physical predictions of $z < 0$ which are equally avoided in our model, while were present in the ANNz2 results. We note, however, that both cases were rare already in B21. In the clean KiDS-Bright sample, there were 82 galaxies with $z_{ANNz2} > 1$ and 444 with $z_{ANNz2} < 0$.

The architecture of the Hybrid-z model is shown in Fig. 2. The left-hand side is the CNN part, where images are processed, while to the right we have the ONN section using magnitudes. For CNNs, the input is 36×36 pixels (7.2″×7.2″) galaxy cutouts in four optical bands. When training the network, we used the Adam optimizer (Kingma & Ba 2014). One of its key features is the adaptive learning rate, updated during iteration based on previous steps. 10^{-4} is selected as the initial learning rate after various iterative tests with it from $[10^{-5}, 10^{-2}]$. The loss function is the Huber loss with the δ hyperparameter value of 10^{-3} as mentioned in Sect. 3.5. During training, we calculate MAE and MSE in each epoch. When the epoch progresses, their values decrease.

The features extracted by the convolutional and average pooling layers are passed through the Inception module before reaching the final dense layers. We found that the performance of the model improves with the inclusion of Inception, with notable improvement found in the reduction of the loss function value. We conducted tests with different numbers of Inception modules, ranging from three to five. Our findings suggest that using four Inception modules with varying numbers of filters is optimal for our data. The first inception module receives input from the Average Pooling 2D layer, while each subsequent inception module takes its input from the output of the preceding inception module. The initial inception modules extract low-level features and the final inception layers capture the global patterns in the image. Padding is applied before the convolutional and pooling operations for each Conv2D and AveragePooling2D layer using the ‘same’ padding (He et al. 2015). This ensures that the spatial dimensions (height and width) of the output feature maps are maintained as much as possible relative to the input. When the kernel is applied to the input, padding is added around the edges as necessary to keep the spatial dimensions consistent after each convolution and pooling operation.

As in Li et al. (2022) and Henghes et al. (2022), for our photo-z model, we combined two different types of networks, “shallow” (fully connected) ANN, that we will denote as ordinary neural networks (ONNs) and CNN, hence the name

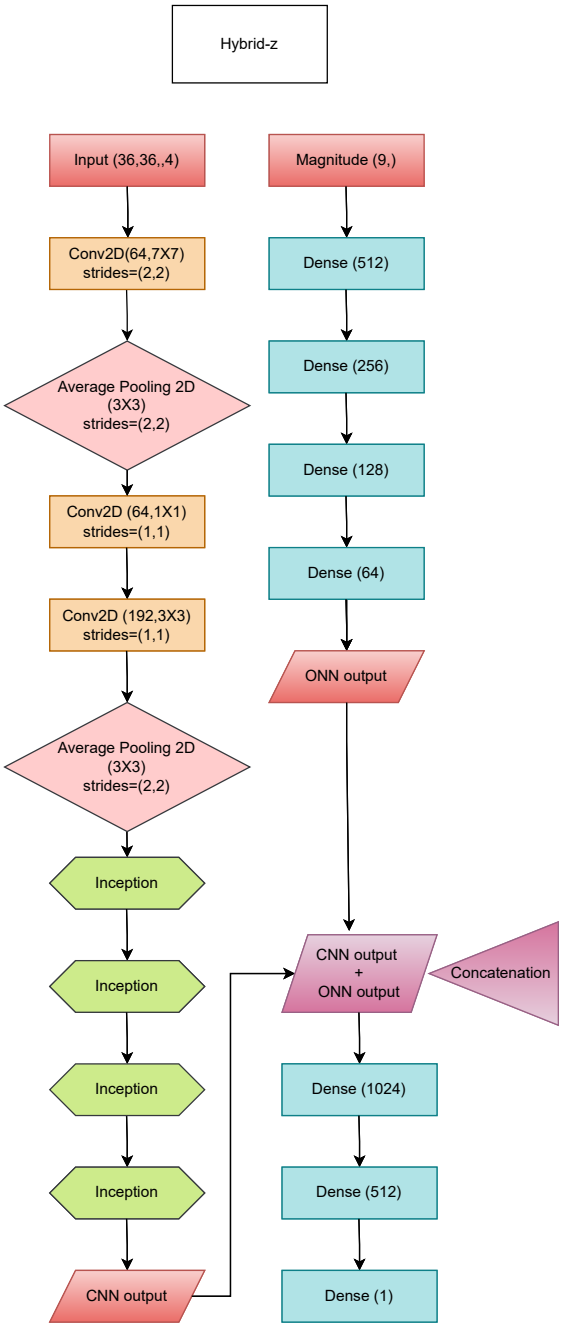


Fig. 2. Architecture of the Hybrid-z model, which employs four-band KiDS galaxy images together with nine-band KiDS+VIKING magnitudes for photo-z derivation. For the CNN part (left-hand side), the symbols used are as in Fig. 1, while inception modules are represented as hexagons. In the fully connected part (right-hand side), dense layers are given as blue rectangles, and the number of neurons is shown in parentheses.

Hybrid-z. There are ∼13.8 million trainable parameters in our model. An important aspect of the Hybrid-z model is the concatenation step. The nine-band magnitudes of galaxies are the additional information for the network. These are processed by dense layers. The flattened feature map and ONN output are combined via depth-wise concatenation. ONN output has 64 features. However, when we include images, the number of features increases to 12 064. Then this concatenated information is passed through final layers for the prediction of photo-zs. This significant increase in features for final dense layers after concatenation highlights the substantial contribution of the image data.

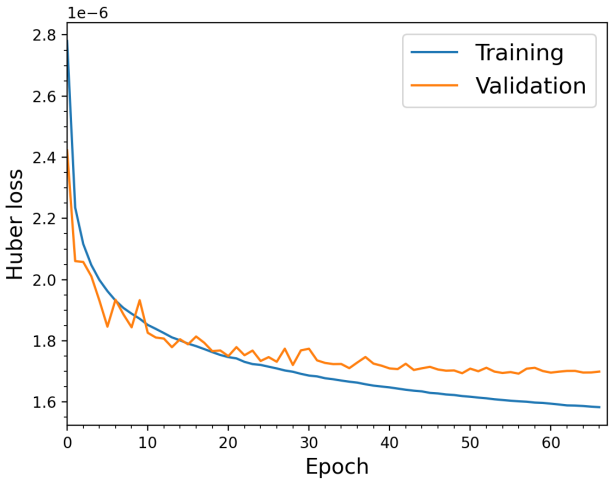


Fig. 3. Performance of the Huber loss function during the training and validation stage of the Hybrid-z.

In Fig. 3, the performance of the Huber loss function for the Hybrid-z is shown. The addition of supplementary information enhances the model's performance with respect to using only 4-band images. We find that for this model, we can achieve R^2 value greater than 0.93. We used early stopping criteria (Prechelt 1996) for determining the number of epochs. The threshold for considering an improvement is zero, meaning that any decrease in the validation loss would count as an improvement. Early stopping will be triggered after ten consecutive epochs where the validation loss does not improve and the training will automatically stop.

5. Results and discussion

Here we present the results of applying the model described in the previous Section to the KiDS-Bright data discussed in Sect. 2. We benchmark our findings against the results obtained previously in B21 using 'shallow' neural-network software ANNz2. We then analyze the Hybrid-z results for the test sample in more detail. Finally, we apply our best model to the entire KiDS-Bright DR4 photometric sample and discuss the properties of the resulting photo-zs. To reiterate, our training and testing samples with redshift labels have been selected from GAMA equatorial data.

5.1. Hybrid-z performance on test data

The comparison of the main statistics of Hybrid-z versus ANNz2 results are provided in Table 1. Our new model employing jointly optical imaging and KiDS+VIKING magnitudes performs clearly better than ANNz2 which used 9-band magnitudes. In the first block of the Table, we provide statistics of a general test sample, which consists of GAMA galaxies cross-matched with KiDS-Bright including both 'clean' (masked=0) and 'contaminated' (masked=1) data, where the 'masked' flag was derived in B21 based on the KiDS-internal 'MASK' bitwise column. The compared methods have comparable mean residuals – that is, they are similarly accurate – but Hybrid-z outperforms ANNz2 in its photo-z precision, quantified as the scatter in Δz. Both in terms of standard deviation and SMAD, our combined CNN+ONN approach performs better, be it for the

Table 1. Statistics of photometric redshift performance obtained for the KiDS-GAMA test sample.

Sample	Size [1]	$\langle z_{\rm spec}\rangle$	Photo-z model	$\langle z_{\rm phot}\rangle$	$\langle \delta z\rangle$	$\langle \Delta z\rangle$	$\sigma_{\Delta z}$	SMAD(Δz)
Test data [2]	26 035	0.230	ANNz2 (9-band)	0.230	−0.0002	0.0005	0.0254	0.0180
			Hybrid-z (this work)	0.230	−0.0003	0.0002	0.0203	0.0145
Clean data [3]	20 965	0.232	ANNz2 (9-band)	0.232	0.00008	0.0006	0.0248	0.0178
			Hybrid-z (this work)	0.232	−0.0002	0.0003	0.0197	0.0142
Red galaxies [4]	11 062	0.240	ANNz2 (9-band)	0.239	−0.0006	0.00005	0.0198	0.0155
			Hybrid-z (this work)	0.239	−0.0006	0.0002	0.0168	0.0129
Blue galaxies [4]	12 096	0.212	ANNz2 (9-band)	0.213	0.0007	0.0012	0.0269	0.0197
			Hybrid-z (this work)	0.212	0.000001	0.0004	0.0212	0.0154

Notes. Angle brackets $\langle\cdot\rangle$ indicate the mean. [1]Number of galaxies in the sample. [2]General test data from the KiDS-DR4 Bright sample crossmatched with GAMA in the equatorial fields, including both clean and contaminated galaxies. [3]Galaxies from the test sample which have the flag masked=0 and are therefore free of any known artifacts B21. [4]Galaxies separated into red and blue based on their position on the absolute r-band magnitude - rest-frame $u-g$ color diagram, see B21 for details.

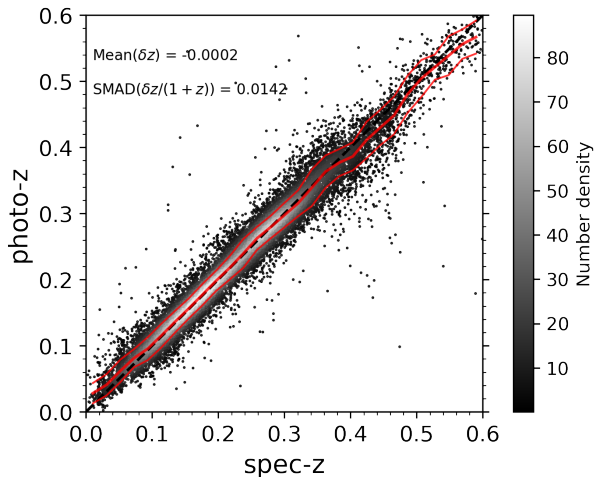


Fig. 4. Comparison of the true spectroscopic redshifts with the photometric ones derived with our framework, Hybrid-z, using four-band KiDS images and nine-band KiDS+VIKING magnitudes. The test data were taken from a subset of GAMA crossmatched with KiDS. The thick red line depicts the median while the thin lines show the scatter (SMAD) around the median.

overall or clean (masked=0) test data. Hybrid-z provides about 20% smaller scatter than the ANNz2-based derivations of B21, where 9-band magnitudes were used. In the case of the fiducial clean data that are recommended for science, we achieve SMAD(Δz) $\simeq$ 0.014, as compared to 0.018 in B21. Our result is at the same level as that by Treyer et al. (2024) where CNNs were used for photo-zs based on SDSS images at similar depths as in our case.

To further evaluate the performance and stability of Hybrid-z, we used the k-fold cross-validation technique. We employ $k = 5$ folds, and the dataset was systematically divided into five distinct subsets. In each iteration of the cross-validation process, one fold was reserved for validation, another for testing, and the remaining three folds were used for training the model. This procedure was repeated five times, each fold taking a turn as the validation and testing set. This ensures that each data point is used for validation, testing, and training, providing a comprehensive evaluation of the model's performance across different subsets of data. The photo-z statistics for each of the folds were very similar, we therefore quote in the tables numbers from only one such runs.

We visualize the performance of the Hybrid-z model in Figs. 4 and 5, where we limit the range of redshift shown to $z < 0.6$ as beyond that value there are practically no galaxies in our samples (for instance, there are only 18 objects with $z > 0.6$ in the test data). Figure 4 directly compares the true redshifts from the test set with our photo-z predictions. The running median (depicted by the thick red line) closely follows the diagonal for most of the $0 < z < 0.6$ range, deviating only at the lowest and highest redshifts. This behavior is typical for ML photo-z derivations, which are unbiased as a function of photometric, but not spectroscopic redshifts. This is further illustrated in panels (a) and (b) of Fig. 5, where photo-z residuals (rescaled by $1 + z$ as typically done) are plotted as a function of respectively predicted photo-z (panel a) and of the true spectroscopic redshift (panel b).

In panels (c) and (d) of Fig. 5 we illustrate the behavior of photo-z errors of our Hybrid-z model as a function of two observables: the r-band apparent magnitude and the observed $u-g$ color. The former is shown for the 'auto' flux measurement, which approximates the total light of a galaxy and was used by B21 to select the flux-limited KiDS-Bright sample. We observe slightly growing scatter (thin red lines indicating SMAD in Fig. 5 (c)) at the faintest end of the sample ($r \lesssim 20$). Otherwise, photo-zs are stable as a function of magnitude, with practically zero bias at most of the range. The $u-g$ color shown in panel (d) can be used to split galaxies into red and blue populations (similarly as $u-r$, Strateva et al. 2001). Indeed, we see the bimodality in the plot, with bluer galaxies to the left and redder to the right. As expected for photo-zs, the former has a wider scatter with respect to the true value than the latter. The difference is, however, not substantial, and again the median biases for the whole range of the $u-g$ color are very close to zero.

We further quantify the performance of Hybrid-z for blue and red galaxies in the two bottom blocks of Table 1. Here, for a direct comparison between this work and B21, the selection of the two galaxy types was based on their intrinsic rather than observed properties, namely via their positioning on the

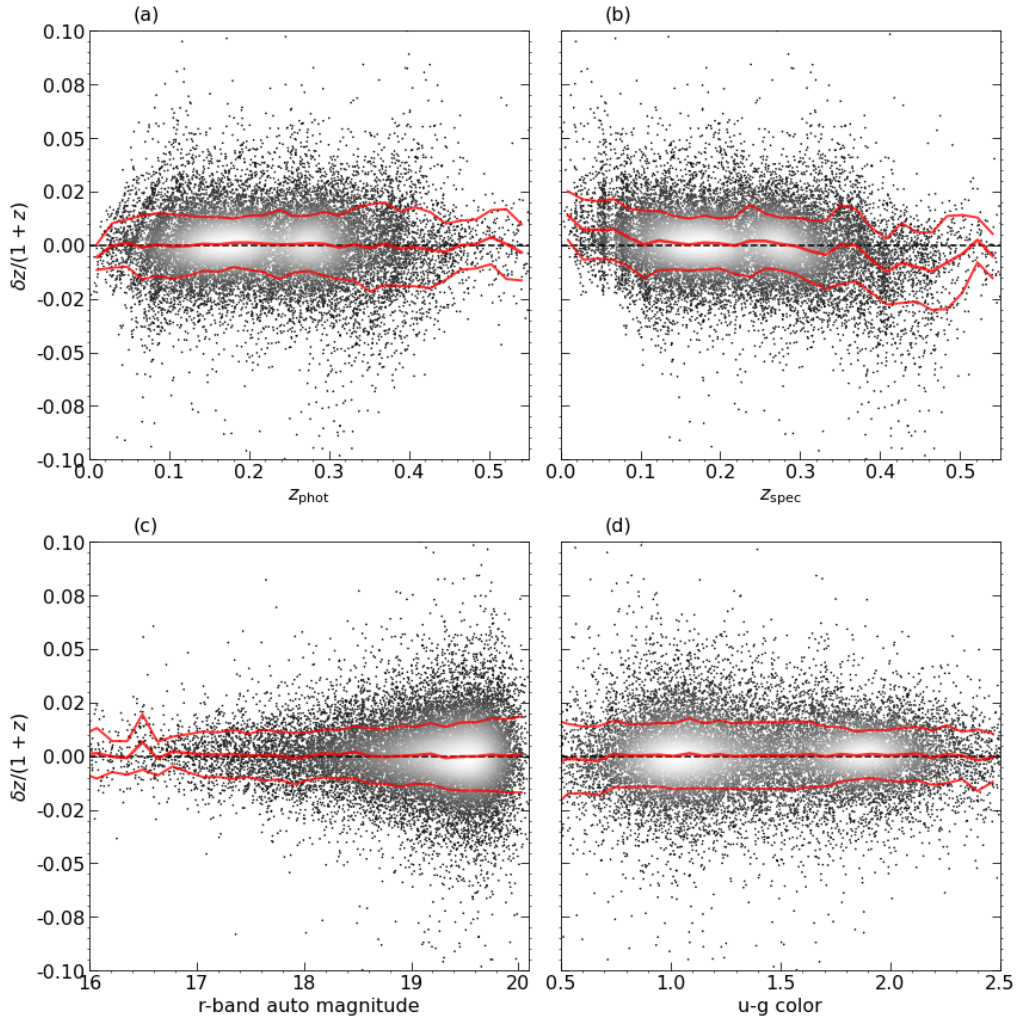


Fig. 5. Photometric redshift errors of Hybrid-z as a function of four quantities, from left to right, (a) photometric redshift as derived in this work; (b) true spectroscopic redshift from the clean sample; (c) apparent r-band magnitude; and (d) observed $u - g$ color. The thick solid red line represents the running median, while the thin red lines illustrate the scatter (SMAD), based on a blind test sample derived from GAMA.

absolute r-band magnitude versus rest-frame $u - g$ color diagram[4] as discussed in B21. Similarly as for the full sample, also for these subsets, Hybrid-z provides considerable improvement in precision over ANNz2. Interestingly, this reduction in SMAD(Δz) is greater for blue galaxies (by ∼22%) than for red ones (∼17%). This is consistent with the fact that, as compared to the approaches relying solely on summary measurements such as magnitudes, the DL-based methods, which directly extract features from images to estimate redshifts, are expected to show larger improvement for galaxies with intricate morphologies such as spirals, that typically have bluer colors (e.g., Schuldt et al. 2021; Treyer et al. 2024). In contrast, for red galaxies – typically ellipticals with fewer features in images – the reduction in scatter for Hybrid-z is smaller than for the overall sample, although we would such as to emphasize that it still does perform better than ANNz2 for the same galaxy selection. In addition, it is worth noting that SMAD(Δz) of 0.0154 for blue galaxies as obtained by Hybrid-z is even smaller than the same statistic of ANNz2 for red ones (cf. blocks #3 and #4 of Table 1).

The final comparison between ANNz2 and Hybrid-z performance is provided in Fig. 6, where we show the distributions of photo-z errors Δz as calibrated on the GAMA test data. As discussed in B21, following earlier work (e.g., Bilicki et al. 2014), such a distribution presents non-Gaussian features in the wings and is therefore better fit with a "generalized Lorentzian" of the form[5]

$$N(\Delta z) \propto \left(1 + \frac{\Delta z^2}{2as^2}\right)^{-a}, \tag{10}$$

where the mean bias is assumed to be negligible (but see Hang et al. 2021 on including non-zero mean). For our Hybrid-z model in the test sample, the best-fitting values are $a = 2.332$ and $s = 0.0116$. If we instead fit a Gaussian, this time allowing the mean to depart from 0, we get best-fit values of $\sigma = 0.0145$ and $\mu = 2.56 \times 10^{-4}$. As visible in Fig. 6, the Gaussian is a worse fit to the residuals than Eq. (10). In the same plot, we also show the histogram for ANNz2 residuals as derived by B21, which is clearly broader. Quantitatively, the above numbers can be compared with best-fit $\{a, s\}_{\rm B21} = \{2.613, 0.0149\}$ and $\sigma_{\rm B21} = 0.0180$. For the Gaussian, the reduction in scatter of Hybrid-z is by 24%. For the modified Lorentzian, the width is

[4] More specifically, we crossmatched our sample with galaxies labeled as red and blue in KiDS-Bright DR4.

[5] It is interesting to note that also spectroscopic redshifts may present Lorentzian rather than Gaussian uncertainties, e.g., Yu et al. (2024).

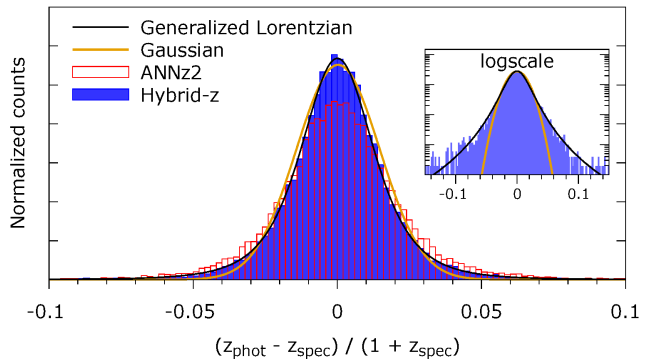


Fig. 6. Comparison of photometric redshift error distributions between Hybrid-z (blue, this work) and ANNz2 (B21, red) shown as normalized counts based on the test sample results. Solid lines illustrate the best-fit Gaussian (with $\sigma = 0.0145$ and $\mu = 2.56 \times 10^{-4}$, yellow) and generalized Lorentzian (Eq. 10, with $a = 2.332$ and $s = 0.0116$, black). The inset compares the histogram of Hybrid-z photo-z residuals with the fit models in logscale of the y-axis.

encoded in the s parameter while a quantifies the extent of the wings; we see a reduction in both for Hybrid-z in comparison to ANNz2.

To summarize this subsection, the Hybrid-z model using CNNs on KiDS optical images together with ONNs on KiDS+VIKING magnitudes provides considerably better results than the previous 9-band ANNz2 derivations of B21. The improvement from ANNz2 to Hybrid-z is at the level of 20% reduction in photo-z scatter (better precision) while maintaining almost zero bias (very good accuracy). Thanks to our new model, we observe smaller photo-z errors for various subsamples of test data, and most remarkably for blue galaxies, where SMAD(Δz) is lowered by about 22% from what was obtained in B21. As such, these are state-of-the-art results for the KiDS-Bright galaxies and in line with independent derivations for SDSS at the same depth (Treyer et al. 2024).

5.2. Validation of Hybrid-z across external spectroscopic samples

In addition to the standard evaluation of Hybrid-z performance from Sect. 5.1 done on test samples statistically consistent with the training data (as both were randomly sampled from input GAMA equatorial catalogs), we have done a further validation of our photo-zs. Still using GAMA-Equatorial for training, we checked Hybrid-z predictions for a number of spectroscopic samples overlapping with the KiDS DR4 area and covering appropriate magnitude ranges and sky areas to provide sufficient statistics. These include GAMA Equatorial and G23 fields (Driver et al. 2022), 2dFGRS (Colless et al. 2001), SDSS DR16 (Ahumada et al. 2020), and 2dFLenS (Blake et al. 2016). Among these, GAMA G23 and 2dFLenS are disjoint with our GAMA-Equatorial training set, as they cover only the southern patch of KiDS (at $\delta < -25°$)[6]. The surveys SDSS and 2dFGRS do have common objects with GAMA in its equatorial patches. However, they extend to the much wider KiDS area, so their overlap with GAMA-eq is also small.

In Table 2 we present the statistics derived from crossmatching the above-mentioned datasets with the full KiDS-Bright 'clean' sample. For each spectroscopic dataset, we compare the performance of Hybrid-z with previous ANNz2 results from B21. For each of these test sets, Hybrid-z performs considerably better in terms of scatter than ANNz2, while generally the former displays larger bias than the latter. This indicates that our new model loses some of the accuracy that ANNz2 had while gaining in precision (typical ML trade-off). We note, however, that in all the inspected cases, for Hybrid-z the mean biases in δz or Δz are much smaller than the scatter, meaning that the model is still highly accurate in its photo-z predictions. In the future, we plan to inspect this further with KiDS DR5 (Wright et al. 2024) 9-band imaging and extending the training with, for example, DESI DR1 (DESI Collaboration 2025), to see if we can minimize both precision and accuracy at the same time.

What is worth noticing is the very small SMAD of Hybrid-z (~0.012) for the SDSS DR16 crossmatch. This dataset, due to the $r < 20$ flux limit of KiDS-Bright, includes galaxies from the SDSS main sample (flux-limited to $r < 17.77$), as well as a subset of BOSS LOWZ and CMASS, dominated by luminous red galaxies. This combination leads to the photo-z statistics for such a SDSS × KiDS-Bright crossmatch being overall better than for the general red galaxies specified in Table 1. However, the same is not the case for ANNz2, which performs only marginally better for such a mixture of SDSS galaxies than it did for the general red galaxy selection.

Among the datasets included in Table 2, GAMA G23 and 2dFLenS are genuinely 'blind' test samples, as they are separated on-sky by many degrees from the area where the GAMA-Equatorial training sets are. Photo-z statistics derived from these catalogs should therefore be robust against possible overfitting thanks to their independence from the training data. While one should remember that neither G23 nor 2dFLenS are as complete and flux-limited samples as KiDS-Bright and GAMA-Equatorial, it is reassuring to find that for both of the former Hybrid-z gives considerable improvement over ANNz2, namely reduction in scatter by respectively 17% and 20%.

To summarize, in addition to evaluating the Hybrid-z model on test samples similar to the training data, we validated its photo-z predictions using several independent spectroscopic datasets within the KiDS DR4 survey area. Results show that Hybrid-z outperforms the previous ANNz2 model in terms of scatter across all test datasets, indicating robust performance also on datasets statistically different from the training.

5.3. Application to the KiDS-DR4 Bright sample

The Hybrid-z model trained on the full GAMA equatorial data, as discussed above, provides satisfactory results when tested on various overlapping spectroscopic datasets. However, applying it directly to the KiDS-Bright sample introduces artifacts in the final photo-z distribution, which we believe are related to the specific properties of the GAMA training set. Namely, GAMA sky coverage of 180 deg^2 is small enough that the survey is affected by cosmic variance, manifesting itself by enhanced effects of the cosmic web, such as voids and filaments (e.g., Eardley et al. 2015). This results in significant 'peaks' and 'dips' in the GAMA redshift distribution, which should average out for larger sky area. However, these kind of features are still imprinted onto our photo-z predictions. Namely, training the Hybrid-z model directly on the full GAMA dataset leads to $dN/dz_{\rm phot}$ of KiDS-Bright which mimics some of the LSS-related properties of GAMA $dN/dz_{\rm spec}$.

As discussed in previous papers (e.g., Bilicki et al. 2016, B21), the redshift distribution of GAMA has strong features, and in particular, there is an under-abundance of galaxies at

[6] A very small fraction (~1%) of 2dFLenS is located in KiDS-N, but the overlap with GAMA-Equatorial fields is negligible (~200 sources).

Table 2. Statistics of photometric redshift performance for KiDS-Bright crossmatched with various spectroscopic samples.

Sample	Size [1]	$\langle z_{\rm spec}\rangle$	Photo-z model	$\langle z_{\rm phot}\rangle$	$\langle\delta z\rangle$	$\langle\Delta z\rangle$	$\sigma_{\Delta z}$	SMAD(Δz)
GAMA Equatorial [2]	145 493	0.229	ANNz2	0.229	0.0005	0.0009	0.0237	0.0178
			Hybrid-z (this work)	0.224	−0.0044	−0.0031	0.0187	0.0144
2dFGRS [3]	53 179	0.119	ANNz2	0.122	0.0022	0.0022	0.0238	0.0158
			Hybrid-z (this work)	0.120	0.0004	0.0006	0.0167	0.0123
SDSS DR16 [4]	43 581	0.221	ANNz2	0.220	−0.0011	−0.0002	0.0221	0.0157
			Hybrid-z (this work)	0.217	−0.0034	−0.0023	0.0158	0.0116
GAMA G23 [5]	34 941	0.215	ANNz2	0.217	0.0021	0.0022	0.0244	0.0166
			Hybrid-z (this work)	0.213	−0.0027	−0.0017	0.0184	0.0137
2dFLenS [6]	22 128	0.251	ANNz2	0.251	−0.0007	0.0005	0.0365	0.0167
			Hybrid-z (this work)	0.247	−0.0042	−0.0025	0.0224	0.0133

Notes. [1]Number of galaxies in the crossmatched sample. [2]Galaxy And Mass Assembly in the equatorial fields (Driver et al. 2022), used for training both models. [3]2-degree Field Galaxy Redshift Survey (Colless et al. 2001). [4]Sloan Digital Sky Survey Data Release 16 (Ahumada et al. 2020). [5]GAMA in the southern G23 field (Driver et al. 2022), disjoint with the training set. [6]2-degree Field Lensing Survey (Blake et al. 2016).

$z_{\rm spec} \sim 0.25$. This redshift is close to the median of the sample, which is where ML models typically work optimally. If we then train on such a specific distribution, our model becomes biased toward this input dN/dz, which results in a 'dip' at $z_{\rm phot} \sim 0.25$ and two peaks below and above this value (see panel (a) of Fig. 7). As photo-zs dilute the structures in the radial direction, for a sample covering $\sim$1000 deg^2 such as ours, this kind of strong features in its $dN/dz_{\rm phot}$ are unlikely to be physical but rather originate from 'redshift focusing' of the model. However, this behavior was not observed in B21, where photo-zs were derived with ANNz2.

Our interpretation is that our new model is more sensitive to such strong features in the training data and this needs to be mitigated. One possibility would be to include training sets covering more of the sky, and hence they would be less affected by cosmic variance. This will be possible thanks to, for example, DESI[7]. Another option is to train several models, with for instance different random seeds and/or architectures, and appropriately combine their outputs into the final prediction. Such an approach is implemented in ANNz2, but it uses much less computationally demanding ONNs. Employing a similar framework with DL would be beyond our scope of research. Below we propose another mitigation strategy consisting of appropriately resampling ("smoothing") the available training set.

The second effect we observe is related to the mismatch between the training and target photometric datasets at the faint end. Namely, GAMA is complete to a brighter magnitude than the KiDS-Bright selection, which will lead to some extrapolation of model predictions at the faint end. As quantified in the recent paper by Jalan et al. (2024), the GAMA-equatorial sample becomes considerably incomplete with respect to KiDS-Bright at $r \gtrsim 19.5$ mag. For a supervised ML model such as ours, this leads to extrapolation resulting from the so-called covariate shift. ML models tend to perform well when the test and training data share a well-matched feature space and distribution of the target quantity. When this distribution shifts or a covariate shift occurs, model performance can be adversely affected (Y et al. 2019). In our case, this affects the predictions at the faintest magnitudes of the sample. In particular, as was already the case for ANNz2, many galaxies with $r > 19.5$ are assigned $z_{\rm phot} > 0.35$ where training data is sparser. For Hybrid-z, training directly on the full GAMA dataset introduces additionally a new peak at $z_{\rm phot} \sim 0.38$, revealing a redshift focusing effect in our model's predictions as evident from panel (a) of Fig. 7).

In order to solve the above-discussed issues, for the final training of the full-sample photo-zs, we decided to subsample the spectroscopic redshifts from the GAMA equatorial dataset in such a way as to smooth out the peaks and dips originally present. The subsampling was done in such a way to not affect the color-redshift relation but instead provide a 'smoother' (more regular) input redshift distribution in GAMA without the strong features discussed above. We create this smoothed subsample of training data by iteratively adjusting the distribution of redshift values to achieve a more uniform histogram. Starting with an initial histogram of true z values, bins with counts that significantly exceed their neighboring bins are identified as spikes, based on a decreasing threshold. In each iteration, if a bin exceeds the count of its neighbors by this threshold, its count is reduced to either the average of its neighbors or its original count, whichever is lower. Randomly selected data points are then drawn from each bin according to the adjusted counts, preserving the dataset's structure but smoothing out extreme values. This iterative process results in a more evenly distributed subsample, which is useful for downstream tasks, by reducing overrepresented regions. Finally, the resulting 'KiDS ID' values and their corresponding z values for this subsample are output, with a calculation of the subsample size as a percentage of the original dataset. This gave us an output of 118k galaxies from the GAMA training set (about 66%of the original one) and their redshift distribution is shown as red bars in panel (b) in Fig. 7. Using these smoothed data we retrained the Hybrid-z model and then applied it to the full $r < 20$ mag dataset of $\sim$1.2 million KiDS DR4 galaxies. Among these, about 996k have the flag 'masked=0' indicating their usefulness for science (see B21 for details).

In panel (b) of Fig. 7 we compare the $dN/dz_{\rm phot}$ of the full sample from our model (Hybrid-z) and ANNz2 from B21. We

[7] DESI Data Release 1 (DESI Collaboration 2025) was released after this work had been completed; hence, we do not include it here.

also show the spec-z distribution of the smoothed training sample. Differences in the photo-z distributions are notable between Hybrid-z and ANNz2, particularly at $z \sim 0.24$ and within the redshift range of $(0.3, 0.4)$. Otherwise, they are very consistent, despite the fact that ANNz2 redshifts were trained on the full GAMA sample as shown in panel (a), while Hybrid-z used the smoothed GAMA subsample for training, as shown in panel (b). What persists in our derivations is the peak at $z_{\rm phot} \sim 0.38$, which is present whether we train on original or smoothed GAMA data. We note that the comparison between Hybrid-z and ANNz2 $dN/dz_{\rm phot}$ serves here just as a cross-check, but the aim is not to make them overlapping. While some consistency between the two approaches is expected, as the models use the same input training data and were applied to the same inference sample, direct comparisons between two photo-z approaches should be done with care. For science applications of such data, further redshift calibration is needed to either reproduce the underlying $dN/dz_{\rm true}$ or to build a photo-z error model.

To summarize, the Hybrid-z model, trained on the full GAMA equatorial dataset, performs well on overlapping spectroscopic surveys, but introduces artifacts when applied directly to the KiDS-Bright sample, due to the specific properties of the GAMA training set. To mitigate these problems, a smoothed subsample of the training data was created to achieve a more uniform redshift distribution, which was then used to retrain the model. As a result, we obtain photo-z predictions for the full KiDS-DR4 Bright galaxy sample, which displays improved performance over previous derivations and also gives a generally artifact-free redshift distribution.

We release the photometric redshifts generated with our Hybrid-z model for the entire KiDS-DR4 Bright Sample as a supplement to the original dataset which was accompanying B21. This is available from the KiDS webpage at `https://kids.strw.leidenuniv.nl/DR4/brightsample.php`.

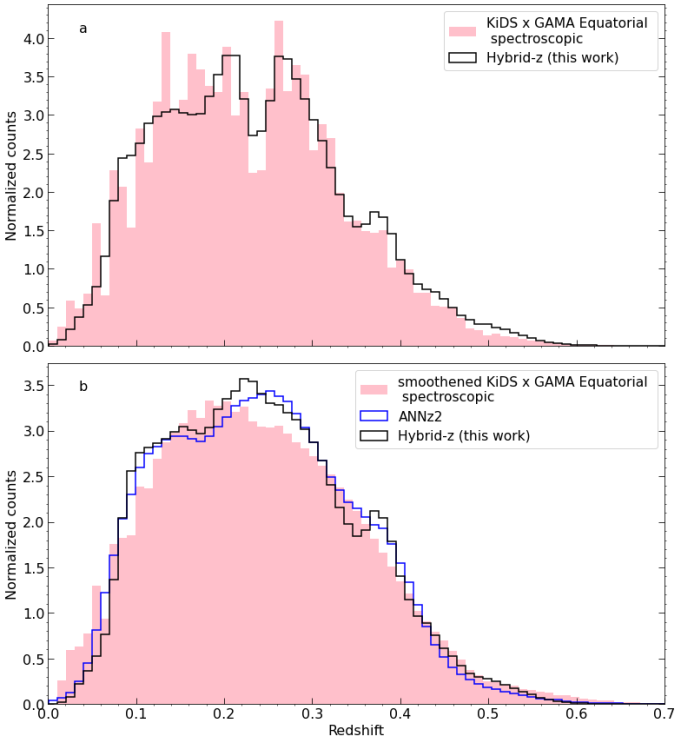


Fig. 7. Photo-z distribution of the KiDS-DR4 Bright sample ($r < 20$ mag) as predicted by ANNz2 (blue, B21) compared with photo-z estimates from Hybrid-z (this work) trained on the spec-z distribution of (a) KiDS x GAMA bright sources (original training sample) (b) smoothed KiDS x GAMA bright sources.

6. Conclusions and future prospects

This work presents the first DL photometric redshift (photo-z) derivations for the flux-limited KiDS-Bright galaxy sample with the selection threshold $r < 20$ mag (Bilicki et al. 2021). Previously, photo-zs for this catalog were estimated using "shallow" learning methods, specifically the ANNz2 neural network package (Sadeh et al. 2016). Our new model, Hybrid-z, is built on recent studies, including Li et al. (2022), where DL was applied for photo-zs in a deeper KiDS galaxy sample, and Treyer et al. (2024), who used a similar training set as us to obtain CNN-based redshifts for an SDSS-selected catalog with the same flux limit.

We built and tested a DL model for photo-z derivation called Hybrid-z that uses four-band KiDS images (*ugri*) processed by CNNs, which are combined with nine-band magnitudes from KiDS+VIKING, processed by an ONN. Rather than simply averaging the outputs of the two networks, we concatenated their outputs before the final three dense layers. This approach was inspired by the previous works of, for example, Li et al. (2022) and Henghes et al. (2022), and it yielded significantly improved photo-z performance compared to the previous ANNz2 derivations of B21, where nine-band KiDS+VIKING magnitudes were used. The Hybrid-z model reduces the scatter (SMAD) of Δz by 20% as compared to ANNz2 for the same test samples. This is true for multiple spectroscopic test datasets, of which some can be considered entirely "blind" in terms of being fully disjointed with the training data. When tested on the fiducial "clean" KiDS-Bright sample, Hybrid-z achieves a SMAD(Δz) of approximately $0.014(1+z)$, representing a clear improvement over the prior ANNz2 results giving $\sim 0.018(1+z)$ while maintaining the same minimal bias in δz of at most a few times 10^{-4}.

The Hybrid-z model shows even greater improvement in photo-z precision when we separate out blue galaxies from the KiDS-Bright sample. For these objects, Hybrid-z reduces the SMAD of Δz by 22% as compared to ANNz2, resulting in a scatter comparable to that which ANNz2 attained for red galaxies. This of course, means that at the same time, the photo-z performance for red galaxies improves less than on average (by 17% in SMAD) after adding DL as compared to the ordinary "shallow" networks. This is consistent with expectations, as CNNs can leverage the detailed varied features in blue (typically spiral) galaxies more effectively than in the smoother elliptical red galaxies. In a flux-limited sample at low redshift, as ours, blue galaxies are more abundant than red galaxies, so such a property of the photo-z model is very useful to improve the overall quality of the derivations. These advancements pave the way for more refined astrophysical and cosmological analyses using KiDS-Bright data, such as of the stellar-to-halo-mass relation (B21) or multi-probe analyses (Dvornik et al. 2023).

Our new photo-z model trained on the GAMA equatorial dataset performs reliably on KiDS overlapping spectroscopic data, but it introduces artifacts in the redshift distribution when directly applied to the KiDS-Bright sample. These artifacts stem from GAMA's limited sky coverage and cosmic variance, which affect the resulting $dN/dz_{\rm phot}$ and introduce specific patterns that mimic GAMA characteristics, such as a notable dip at $z_{\rm phot} \sim 0.25$, and extrapolation issues at faint magnitudes and where $r > 19.5$ mag. To overcome these limitations, we created a smoothed subsample of the GAMA training data by reducing

sharp peaks and dips in its redshift distribution and achieving a more uniform $dN/dz_{\rm spec}$ for final model training. This approach mitigates artifacts in the output $dN/dz_{\rm phot}$, thus enhancing the reliability of our photo-z estimates. These improved photo-zs for the KiDS-DR4 Bright sample (∼1.2M galaxies) are available publicly online[8].

In this work, our DL analysis focused on the four-band KiDS optical images, while the full nine-band KiDS+VIKING photometry was used only in the form of magnitudes. Moving forward, we plan to build an extended DL model incorporating images from all nine bands, from KiDS u to VIKING K_s, once the NIR coadds are available; currently, they are being prepared for the 4MOST WAVES target selection (Driver et al. 2019). We anticipate that this expansion will improve CNN-based photo-z precision and will mirror the gains seen from KiDS DR3 *ugri* (Bilicki et al. 2018) to nine-band inclusion in DR4 (Bilicki et al. 2021). We plan to apply such an enhanced Hybrid-z model to the final KiDS Data Release 5 (Wright et al. 2024) for state-of-the-art bright-end photometric redshifts. Additional improvement in DR5 is expected thanks to the second i-band pass, which deepens effective imaging and should help further refine our photo-z estimates. Last but not least, our approach holds promise for even deeper imaging applications in the forthcoming Legacy Survey of Space and Time (LSST Science Collaboration 2009), paving the way for robust photo-zs in future large-scale sky surveys.

Data availability

The photometric redshift catalog based on the Hybrid-z DL model, containing redshifts for over 1.2 million galaxies in the KiDS-Bright DR4 sample, is available at the CDS via anonymous ftp to cdsarc.cds.unistra.fr (130.79.128.5) or via https://cdsarc.cds.unistra.fr/viz-bin/cat/J/A+A/698/A276

Acknowledgements. We would like to thank Elisa Chisari, Rui Li, Nicola Napolitano & Angus Wright for their valuable comments and suggestions on the manuscript. Based on data products from observations made with ESO Telescopes at the La Silla Paranal Observatory under program IDs 177.A-3016, 177.A-3017 and 177.A-3018, and on data products produced by Target/OmegaCEN, INAF-OACN, INAF-OAPD and the KiDS production team, on behalf of the KiDS consortium. OmegaCEN and the KiDS production team acknowledge support by NOVA and NWO-M grants. Members of INAF-OAPD and INAF-OACN also acknowledge the support from the Department of Physics & Astronomy of the University of Padova, and of the Department of Physics of Univ. Federico II (Naples). GAMA is a joint European-Australasian project based around a spectroscopic campaign using the Anglo-Australian Telescope. The GAMA input catalog is based on data taken from the Sloan Digital Sky Survey and the UKIRT Infrared Deep Sky Survey. Complementary imaging of the GAMA regions is being obtained by a number of independent survey programs including GALEX MIS, VST KiDS, VISTA VIKING, WISE, Herschel-ATLAS, GMRT, and ASKAP providing UV to radio coverage. GAMA is funded by the STFC (UK), the ARC (Australia), the AAO, and the participating institutions. The GAMA website is http://www.gama-survey.org/. This work is supported by the Polish National Science Center through grants no. 2020/38/E/ST9/00395, and 2018/31/G/ST9/03388. We have made use of TOPCAT (Taylor 2005) and STILTS (Taylor 2006) software, as well as of PYTHON (www.python.org), including the packages NUMPY (Harris et al. 2020), SCIPY (Virtanen et al. 2020), and MATPLOTLIB (Hunter 2007).

References

Ahumada, R., Allende Prieto, C., Almeida, A., et al. 2020, ApJS, 249, 3
Ansari, Z., Agnello, A., & Gall, C. 2021, A&A, 650, A90
Baum, W. A. 1957, AJ, 62, 6
Baum, W. A. 1962, in Problems of Extra-Galactic Research, ed. G. C. McVittie, 15, 390
Bellstedt, S., Driver, S. P., Robotham, A. S. G., et al. 2020, MNRAS, 496, 3235
Benítez, N. 2000, ApJ, 536, 571
Bertin, E., & Arnouts, S. 1996, A&AS, 117, 393
Bilicki, M., Jarrett, T. H., Peacock, J. A., Cluver, M. E., & Steward, L. 2014, ApJS, 210, 9
Bilicki, M., Peacock, J. A., Jarrett, T. H., et al. 2016, ApJS, 225, 5
Bilicki, M., Hoekstra, H., Brown, M. J. I., et al. 2018, A&A, 616, A69
Bilicki, M., Dvornik, A., Hoekstra, H., et al. 2021, A&A, 653, A82
Blake, C., Amon, A., Childress, M., et al. 2016, MNRAS, 462, 4240
Bonfield, D. G., Sun, Y., Davey, N., et al. 2010, MNRAS, 405, 987
Capaccioli, M., & Schipani, P. 2011, The Messenger, 146, 2
Carliles, S., Budavári, T., Heinis, S., Priebe, C., & Szalay, A. S. 2010, ApJ, 712, 511
Cavuoti, S., Brescia, M., De Stefano, V., & Longo, G. 2015, Exp. Astron., 39, 45
Chollet, F. 2017, Deep Learning with Python (New York, NY: Manning Publications)
Colless, M., Dalton, G., Maddox, S., et al. 2001, MNRAS, 328, 1039
Collister, A. A., & Lahav, O. 2004, PASP, 116, 345
Connolly, A. J., Csabai, I., Szalay, A. S., et al. 1995, AJ, 110, 2655
Cunha, P. A. C., & Humphrey, A. 2022, A&A, 666, A87
de Jong, J. T. A., Verdoes Kleijn, G. A., Kuijken, K. H., & Valentijn, E. A. 2013, Exp. Astron., 35, 25
de Jong, J. T. A., Verdoes Kleijn, G. A., Boxhoorn, D. R., et al. 2015, A&A, 582, A62
DESI Collaboration (Aghamousa, A., et al.) 2016, arXiv e-prints [arXiv:1611.00036]
DESI Collaboration (Abdul-Karim, M., et al.) 2025, arXiv e-prints [arXiv:2503.14745]
Dey, B., Andrews, B. H., Newman, J. A., et al. 2022, MNRAS, 515, 5285
D'Isanto, A., & Polsterer, K. L. 2018, A&A, 609, A111
Driver, S. P., Hill, D. T., Kelvin, L. S., et al. 2011, MNRAS, 413, 971
Driver, S. P., Liske, J., Davies, L. J. M., et al. 2019, The Messenger, 175, 46
Driver, S. P., Bellstedt, S., Robotham, A. S. G., et al. 2022, MNRAS, 513, 439
Dvornik, A., Heymans, C., Asgari, M., et al. 2023, A&A, 675, A189
Eardley, E., Peacock, J. A., McNaught-Roberts, T., et al. 2015, MNRAS, 448, 3665
Edge, A., Sutherland, W., Kuijken, K., et al. 2013, The Messenger, 154, 32
Gholamalinezhad, H., & Khosravi, H. 2020, arXiv e-prints [arXiv:2009.07485]
Goodfellow, I., Bengio, Y., & Courville, A. 2016, Deep Learning (MIT Press)
Graham, M. L., Connolly, A. J., Željko Ivezic, et al. 2017, AJ, 155, 1
Grespan, M., Thuruthipilly, H., Pollo, A., et al. 2024, A&A, 688, A34
Hang, Q., Alam, S., Peacock, J. A., & Cai, Y.-C. 2021, MNRAS, 501, 1481
Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, Nature, 585, 357
He, K., Zhang, X., Ren, S., & Sun, J. 2015, arXiv e-prints [arXiv:1512.03385]
Henghes, B., Thiyagalingam, J., Pettitt, C., Hey, T., & Lahav, O. 2022, MNRAS, 512, 1696
Hildebrandt, H., Arnouts, S., Capak, P., et al. 2010, A&A, 523, A31
Hildebrandt, H., van den Busch, J. L., Wright, A. H., et al. 2021, A&A, 647, A124
Hoyle, B. 2016, Astron. Comput., 16, 34
Huber, P. J. 1964, Ann. Math. Stat., 35, 73
Hunter, J. D. 2007, Comput. Sci. Eng., 9, 90
Jalan, P., Bilicki, M., Hellwing, W. A., et al. 2024, A&A, 692, A177
Jentzen, A., Kuckuck, B., & von Wurstemberger, P. 2023, arXiv e-prints [arXiv:2310.20360]
Jones, E., Do, T., Li, Y. Q., et al. 2024, ApJ, 974, 159
Kingma, D. P., & Ba, J. 2014, arXiv e-print [arXiv:1412.6980]
Krone-Martins, A., Ishida, E. E. O., & de Souza, R. S. 2014, MNRAS, 443, L34
Kuijken, K. 2008, A&A, 482, 1053
Kuijken, K., Heymans, C., Hildebrandt, H., et al. 2015, MNRAS, 454, 3500
Kuijken, K., Heymans, C., Dvornik, A., et al. 2019, A&A, 625, A2
LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. 1998, Proc. IEEE, 86, 2278
Li, C., Zhang, Y., Cui, C., et al. 2021, MNRAS, 509, 2289
Li, R., Napolitano, N. R., Feng, H., et al. 2022, A&A, 666, A85
Liske, J., Baldry, I. K., Driver, S. P., et al. 2015, MNRAS, 452, 2087
LSST Science Collaboration (Abell, P. A., et al.) 2009, arXiv e-prints [arXiv:0912.0201]
McCulloch, W. S., & P., 1943, Bull. Math. Biophys., 5, 115
McFarland, J. P., Verdoes-Kleijn, G., Sikkema, G., et al. 2013, Exp. Astron., 35, 45
Menou, K. 2019, MNRAS, 489, 4802
Oyaizu, H., Lima, M., Cunha, C. E., et al. 2008, ApJ, 674, 768
Pasquet, J., Bertin, E., Treyer, M., Arnouts, S., & Fouchez, D. 2019, A&A, 628, A10
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, J. Mach. Learn. Res., 12, 2825
Prechelt, L. 1996, Neural Networks, 9, 457
Roster, W., Salvato, M., Krippendorf, S., et al. 2024, A&A, 692, A260

[8] https://kids.strw.leidenuniv.nl/DR4/brightsample.php

Rozo, E., Rykoff, E. S., Abate, A., et al. 2016, MNRAS, 461, 1431
Sadeh, I., Abdalla, F. B., & Lahav, O. 2016, PASP, 128, 104502
Schlafly, E. F., & Finkbeiner, D. P. 2011, ApJ, 737, 103
Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, ApJ, 500, 525
Schuldt, S., Suyu, S. H., Cañameras, R., et al. 2021, A&A, 651, A55
Shorten, C., & Khoshgoftaar, T. M. 2019, J. Big Data, 6, 1
Strateva, I., Ivezić, Ž., Knapp, G. R., et al. 2001, AJ, 122, 1861
Szegedy, C., Liu, W., Jia, Y., et al. 2015, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1
Tagliaferri, R., Longo, G., Andreon, S., et al. 2003, Lect. Notes Comput. Sci., 2859, 226
Taylor, M. B. 2005, in Astronomical Data Analysis Software and Systems XIV, eds. P. Shopbell, M. Britton, & R. Ebert, ASP Conf. Ser., 347, 29
Taylor, M. B. 2006, in Astronomical Data Analysis Software and Systems XV, eds. C. Gabriel, C. Arviset, D. Ponz, & S. Enrique, ASP Conf. Ser., 351, 666
Treyer, M., Ait Ouahmed, R., Pasquet, J., et al. 2024, MNRAS, 527, 651
Vakili, M., Bilicki, M., Hoekstra, H., et al. 2019, MNRAS, 487, 3715
Vakili, M., Hoekstra, H., Bilicki, M., et al. 2023, A&A, 675, A202
Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, Nat. Methods, 17, 261
Wadadekar, Y. 2004, PASP, 117, 79
Way, M. J., & Klose, C. D. 2012, PASP, 124, 274
Way, M. J., & Srivastava, A. N. 2006, ApJ, 647, 102
Wright, A. H., Kuijken, K., Hildebrandt, H., et al. 2024, A&A, 686, A170
Xia, Z. 2024, Appl. Comput. Eng., 37, 212
Y, G. D., Nair, N. G., Satpathy, P., & Christopher, J. 2019, in 2019 Global Conference for Advancement in Technology (GCAT), 1
Yu, J., Ross, A. J., Rocher, A., et al. 2024, arXiv e-prints [arXiv:2405.16657]

Part IV

Angular clustering and bias of photometric quasars in the Kilo-Degree Survey Data Release 4

This work is currently under review by the Astronomy & Astrophysics Journal.

4.1 Introduction

In this chapter, we apply the `Hybrid-z` framework to estimate photo-zs for quasars in the KiDS-DR4. The deep learning model is trained using spec-zs from DESI-DR1 and SDSS-DR17. The results show that `Hybrid-z` achieves improved performance compared to previous approaches that relied solely on photometric features without incorporating imaging data. We constructed a quasar photo-z catalog of size $\sim 157k$.

The clustering analysis is carried out tomographically across four photo-z bins spanning $0.1 \leq z_{\rm phot} \leq 2.7$. The quasar bias ($b$) is calculated by comparing the angular 2PCF of KiDS-DR4 quasars with the theoretical angular 2PCF of matter, as described in Sections 1.4 and 1.5. This study finds that the quasar bias increases with redshift, from $b \approx 1.6$ at $z \approx 0.6$ to $b \approx 4.0$ at $z \approx 2.2$. These results suggest that KiDS quasars reside in dark matter halos with typical masses in the range $\log_{10}(M_{\rm eff}/h^{-1}, M_{\odot}) \approx 12.7$ to 12.9. The work also emphasizes the critical importance of precise redshift calibration, as uncertainties in assumed redshift distribution significantly impact bias estimation, whereas stellar contamination has a negligible effect.

Astronomy & Astrophysics manuscript no. main
November 24, 2025

Angular clustering and bias of photometric quasars in the Kilo-Degree Survey Data Release 4

Anjitha John William, Maciej Bilicki, Wojciech A. Hellwing, Szymon J. Nakoneczny, and Priyanka Jalan

Center for Theoretical Physics, Polish Academy of Sciences, al. Lotników 32/46, 02-668 Warsaw, Poland

ABSTRACT

We investigate the angular clustering and effective bias of photometrically selected quasars in the Kilo-Degree Survey Data Release 4 (KiDS DR4). We update the previous photometric redshifts (photo-zs) of the KiDS quasars using Hybrid-z, a deep learning framework combining four-band KiDS images and nine-band KiDS+VIKING magnitudes. Hybrid-z is trained on the latest Dark Energy Spectroscopic Instrument (DESI) DR1 and Sloan Digital Sky Survey (SDSS) DR17 quasars matching with KiDS, and achieves average bias $\langle \delta z \rangle < 0.01$ and scatter $\sim 0.04(1+z)$ on a test sample. The updated catalog of $\sim 157k$ quasars over 777 deg^2 is divided into four tomographic bins spanning $0.1 \leq z_{\rm phot} \leq 2.7$. In each bin, we measure the angular two-point correlation function and compare it with theoretical predictions for dark matter clustering. We estimate the best-fit scale-independent quasar bias, which increases from $b \approx 1.6$ at $z \approx 0.6$ to $b \approx 4.0$ at $z \approx 2.2$, and is well matched by a quadratic relation in redshift. Our clustering analysis indicates that KiDS quasars reside in dark matter halos of mass $\log_{10}(M_{\rm eff}/h^{-1}M_\odot)$ in the range $\sim$ 12.7–12.9 and effective peak heights $\nu_{\rm eff}$ rising from $\sim$ 1.5 to 2.9 over our redshift span. We study two systematics that could affect the bias derivation: stellar contamination and the redshift distribution assumed in the theoretical modeling. The former has a negligible effect, whereas the latter significantly impacts the derived $b(z)$, emphasizing the importance of redshift calibration. Our work is the first cosmological application of quasars selected from KiDS and paves the way for future extensions in the final KiDS DR5, the Legacy Survey of Space and Time, or the 4-metre Multi-Object Spectroscopic Telescope.

1. Introduction

Quasars, also known as quasi-stellar objects (QSO, Schmidt 1963), are the brightest among the active galactic nuclei (AGN, Rees 1984). Their high luminosity ($\sim 10^{10} - 10^{14} L_\odot$) is due to the matter accretion on a supermassive black hole (SMBH) situated at the center of their host galaxy (Salpeter 1964; Zel'dovich & Novikov 1964). This allows us to observe quasars up to very high redshifts ($z \gtrsim 7$), which makes them valuable astrophysical and cosmological probes of both late and early-time Universe (Fan et al. 2023). Quasar emission originates from regions well within the gravitational influence of the supermassive black hole (SMBH), providing information on the physics of black hole accretion and AGN activity (Marziani et al. 2006; Spilker et al. 2025). At galactic scales (~1–10 kpc), quasars help in understanding galaxy formation, growth, and quenching processes (Prochaska & Hennawi 2009). On cosmological scales ($\gtrsim$ 1 Mpc), they trace the large-scale structure (LSS) of the Universe (Song et al. 2016). For instance, measuring quasar clustering allows us to constrain the redshift evolution of their bias with respect to the underlying dark matter (DM), reveal their host halo masses, and provide access to cosmological information such as detecting baryon acoustic oscillations (e.g. Myers et al. 2007; Ata et al. 2018; Adame et al. 2025). Their spatial distribution can be cross-correlated with galaxy and cosmic microwave background weak gravitational lensing maps to probe the growth of structure and constrain cosmological parameters (e.g. Ménard & Bartelmann 2002; DiPompeo et al. 2016; Luo et al. 2024). Additionally, the large-scale clustering of quasars is sensitive to primordial non-Gaussianity through scale-dependent bias effects (Slosar et al. 2008).

Despite quasars being among the brightest extragalactic objects, selecting them in surveys remains challenging. Their compact emission region outshines the host galaxy, and at high redshifts, their small angular size, combined with the limited resolution of telescopes, causes them to appear point-like, making them morphologically indistinguishable from stars in imaging data. As a result, quasar selection must rely on other properties, such as colors, variability, or spectroscopy, to separate them from the vastly more numerous stellar population. However, the spectroscopic quasar identification method (Richards et al. 2002; Lyke et al. 2020; Storey-Fisher et al. 2024) is time-consuming and observationally expensive, restricting surveys to relatively bright sources and limiting sky coverage and depth of quasar samples. To overcome these limitations, we can apply photometric selection techniques to produce larger and more complete QSO catalogs (e.g. Carrasco et al. 2015; Nakoneczny et al. 2019; Chaussidon et al. 2023; Feng et al. 2025). Photometrically classified quasars offer higher completeness but at the cost of increased contamination compared to their spectroscopic counterparts. However, they provide substantial cosmological information due to their large numbers and wide sky coverage.

Photometric target selection technique for quasar identification is applied within the Kilo-Degree Survey (KiDS, Kuijken et al. 2019) by Nakoneczny et al. 2021, hereafter N21. KiDS covers about 1000 deg^2 of the southern sky in its fourth data release. This paper presents the first study of angular clustering and effective bias for KiDS DR4 quasars. Our results demonstrate that the quasar bias evolution is consistent with a quadratic dependence on redshift, in agreement with prior findings from other surveys (Myers et al. 2007; Shen et al. 2009; Laurent et al. 2017; Eltvedt et al. 2024). We investigate potential systematic effects, finding that uncertainties in the redshift distribution significantly impact the quasar bias estimation, highlighting the necessity of precise redshift calibration. In contrast, stellar contamination within the sample has a negligible effect on the clustering measurements and bias inference. Together, these findings establish a framework for quasar clustering studies with

KiDS and contribute insight into quasar bias evolution across cosmic time.

A prerequisite to study the evolution of quasar properties – such as their bias – is the availability of redshift measurements. Similar to non-active galaxies, the redshifts of quasars can be determined through both spectroscopic and photometric methods. Measuring the accurate and precise spectroscopic redshifts (hereafter spec-zs) is, however, both expensive and time-consuming. The less precise photometric redshift (photo-z) estimation is based on broadband photometry (Baum 1957; Koo 1985; Salvato et al. 2019; Newman & Gruen 2022). This alternative approach can be relatively quick, as compared to spectroscopy, and may give a redshift estimate for all sources within an imaging survey. In this work, we use the empirical approach of machine learning (ML) to derive photo-zs for KiDS DR4 quasars. Such algorithms, typically belonging to the supervised ML category, find the relation between multi-band photometric quantities and the redshift by being trained on datasets containing both photometry and corresponding spec-zs. Examples of their applications to quasars include Brescia et al. (2013); D'Isanto & Polsterer (2018); Nakoneczny et al. (2021); Nakazono et al. (2024).

As compared to other galaxies, quasars exhibit prominent broad emission lines, such as Lyα, C IV, Mg II (Peterson 1997), and they move across different photometric bands at different redshifts. Accurate photometric redshift estimation for quasars therefore benefits from broad wavelength coverage, extending from the blue to the infrared (IR), to help break degeneracies caused by different emission lines entering similar filter combinations at different redshifts, as well as ambiguities from their smooth, power-law continua (Salvato et al. 2019). Still, these effects often produce characteristic peaks and dips in the quasar photometric redshift distribution. Furthermore, QSO spectra are dominated by non-thermal emission, giving them colors distinct from most other extragalactic sources and necessitating training sets specifically tailored for quasar photo-z measurements.

Within the broader ML framework, deep learning (DL), often using images among the inputs, has been finding increasingly numerous uses for photo-z estimation, in particular for quasars (e.g. Pasquet-Itam & Pasquet 2018; Yao et al. 2023; Roster et al. 2024). The term "deep" indicates artificial neural networks (ANN) with multiple layers (often tens or hundreds of them concatenated) and numerous neurons in each layer. In this work, we integrated two types of neural networks, dense (McCulloch 1943; Rosenblatt 1958) and convolutional (CNN, Lecun & Bengio 1995). Our model is called Hybrid-z[1] since it uses both magnitudes and multi-channel images, and was originally developed for the KiDS bright galaxy sample in John William et al. 2025, hereafter A25. In its previous application, Hybrid-z improved the earlier ANN-based photo-zs for the KiDS DR4 bright sample (Bilicki et al. 2021), reducing the scatter of photo-z residuals by 20%. In this work, we apply a similar DL model to update the photo-zs of the photometrically classified KiDS DR4 QSO sample, previously derived with ANNs employing quasar magnitudes and colors by N21.

In addition to developing a new ML model for photo-zs, a key improvement of this work over N21 is the use of the recent Dark Energy Spectroscopic Instrument Data Release 1 (DESI DR1, DESI Collaboration et al. 2025) and Sloan Digital Sky Survey Data Release 17 (SDSS DR17, Abdurro'uf et al. 2022) to build the spectroscopic training set for our photo-z model. As compared to the SDSS DR14 (Abolfathi et al. 2018) spec-z data used previously by N21, adding DESI DR1 considerably improves the quality of our training set and the resulting photo-z estimation. DESI not only includes many more quasars than SDSS in the KiDS fields; it also extends the QSO spectroscopic coverage across the color-redshift space.

We use the KiDS DR4 quasar sample with the updated photo-zs to study the connection between QSO distribution and the underlying DM field via angular clustering, which is a common observational statistic to study the properties of the LSS projected on the sky (Peebles 1973, 1980). Following previous studies (e.g. Myers et al. 2007), we consider quasar auto-correlations in photo-z bins. We use these measurements to constrain the evolution with redshift of the effective quasar bias and of their effective host halo mass.

In the standard cosmological framework, the evolution of density perturbations yields predictions for the power spectrum of the total matter density contrast field, including contributions from both dark and baryonic matter. Observationally, we do not directly probe the clustering of the total matter density field. Instead, we measure the spatial distribution of luminous tracers - in our case, quasars. On sufficiently large scales (megaparsecs), the DM and quasar overdensities ($\delta_{\rm q}$) are related via the bias parameter, such that $\delta_{\rm q} = b\delta_{\rm m}$. This bias is expected to evolve with redshift (Shen et al. 2009; Laurent et al. 2017) and its evolution can be traced from angular clustering measured 'tomographically', i.e., in redshift bins. If additional information on quasar redshift distribution per bin is available – i.e., via photo-zs – the angular two-point correlation function (2PCF) can be related to the theoretical three-dimensional 2PCF or its Fourier counterpart, power spectrum. The factor linking the observed quasar correlations and those theoretically expected for DM is the bias b. In summary, incorporating the redshift distribution enables the projection of theoretical three-dimensional dark matter clustering into angular space, allowing the quasar bias to be inferred from the observed 2PCF.

This paper is structured as follows: In Sec.2, we describe the photometric and spectroscopic datasets used in our analysis. Sec.3 presents our Hybrid-z model and the photo-z estimation for the quasar sample. The clustering analysis of this sample is detailed in Sec.4. Finally, we summarize our findings and discuss future prospects in Sec.5.

2. Data

In this section, we describe the data we use to select quasars and estimate their photo-zs. The quasar sample is derived from KiDS Data Release 4 (DR4, Kuijken et al. 2019), and the spectroscopic data for the training sample are obtained from DESI DR1 and SDSS DR17.

KiDS is a multi-band imaging survey designed for weak lensing studies and it was carried out by the European Southern Observatory (ESO) at the VLT Survey Telescope (VST, Capaccioli & Schipani 2011). The images were taken in four optical broad bands (ugri) and supplemented with five near-infrared bands (ZYJHK$_{\rm s}$) from the VISTA Kilo-degree INfrared Galaxy survey (VIKING, Edge et al. 2013) overlapping with KiDS on the sky. Here we employ KiDS DR4, covering about 1000 deg^2 before masking. The nine-band images were processed by the KiDS team, using in particular the Gaussian aperture and point spread function method (GAaP, Kuijken 2008) to obtain uniformly measured u to K_s magnitudes, which we employ throughout. In addition to the magnitudes, similarly as in A25, for photo-z derivations we also use four-band ugri images, leaving the extension to VIKING images for future work. The optical images

[1] https://github.com/Anjithajm/Hybrid-z.git

in KiDS DR4 are organized into 4×1006 survey tiles[2] with a uniform pixel scale of 0.2 arcsec (de Jong et al. 2015). For the Hybrid-z model described below, we made cutouts centered at quasar positions with a size of $5'' \times 5''$ (25×25 pixels), as most of our quasars are smaller than this.

Our quasar sample is selected from KiDS DR4 following the previous work by N21[3]. That selection was obtained by applying an ANN-based classification model using 9-band detections, trained and tested on SDSS DR14 spectroscopy. In the KiDS DR4 QSO sample, the objects are further categorized into "safe" and "extrapolation" regimes based on the feature space coverage with respect to the training set. In this work, for the clustering analysis presented in Sec. 4, we only use the safe subset, limited to $r < 22$ mag, whose feature space lies well within the domain of the N21 training data. For this subset, the N21 classification model achieves test-data purity of 97% and completeness of 94%. As we do not update the quasar selection over that previous study, limiting our analysis to only the safe QSO set mitigates potential contamination from stars and non-active galaxies, which could affect photo-z quality and subsequent clustering analyses. There are $\sim 157k$ KiDS DR4 objects in this safe QSO category. Unless otherwise specified, the term KiDS DR4 quasar sample hereafter refers to this subset of safe quasars. Note, however, that the general training set does not have the safe selection applied (see below). We leave for future study a possible analysis of quasars derived from KiDS DR5 (Wright et al. 2024; Feng et al. 2025).

While we adopt the same KiDS DR4 QSO sample as selected by N21, we update and improve the photo-zs derived there. This is achieved by both changing the methodology (ML photo-z model) as well as by adding considerably more training data with spectroscopic labels. The former is provided thanks to using the Hybrid-z approach, employing both 9-band magnitudes and 4-band optical images, previously developed and validated on the KiDS-bright galaxy sample (Bilicki et al. 2021) by A25. We discuss the particular Hybrid-z implementation and its performance for KiDS DR4 quasars in the next Sec. 3.

As for the new training data, we replace the SDSS DR14 QSOs previously used by N21 by a joint sample of SDSS DR17 and DESI DR1 quasars. These two datasets cover the northern (equatorial) KiDS area and provide spectroscopically confirmed quasars with spectroscopic redshifts (that may be considered exact) to train and validate our photo-z model. The joint SDSS+DESI sample, with only unique objects kept[4], is dominated by DESI and has about 115k counterparts in the full KiDS DR4 catalog before we apply any further cuts. Of these, about 104k have all nine ugriZYJHK$_s$ bands measured in KiDS DR4 – a selection we apply as our neural network model requires a numerical value to be provided for each of the magnitudes. This general cross-match of DESI+SDSS with KiDS DR4 will be our training and test set for the Hybrid-z model. Note that this training sample goes deeper in magnitude and redshift than the final safe-QSO sample we use for the clustering analysis, which is a conscious choice to provide photo-z estimates also for quasars fainter than eventually employed. The median r-band magnitude of the quasar sample with spectroscopic redshift labels is $r \sim 21.4$ mag, while the depth expressed as the 99th percentile is $r \sim 23.7$. The median spectroscopic redshift of this training set is $z = 1.66$ and 99% of the quasars are contained within the redshift range $0.22 \leq z \leq 3.32$.

[2] https://kids.strw.leidenuniv.nl/DR4

[3] https://kids.strw.leidenuniv.nl/DR4/quasarcatalog.php

[4] Note that majority of SDSS quasars are also present in the DESI catalog. In the cross-match with KiDS DR4, only about 1900 out of 17,000 SDSS QSOs are not in DESI.

3. Photometric redshift derivation

The architecture and configuration of our photo-z model, Hybrid-z, are detailed in A25. In brief, Hybrid-z comprises two parallel branches: a CNN branch that processes ugri KiDS galaxy cutouts through initial convolutional layers and four Inception modules to extract multi-scale features, and an ordinary (fully-connected) neural network (ONN) branch that processes standardized nine-band KiDS+VIKING magnitudes through a sequence of dense layers. The CNN and ONN outputs are concatenated and passed through fully connected layers, with the final layer providing a photo-z estimate. The primary modification between the current model configuration and that detailed in A25 is the exclusion of the sigmoid activation function in the output layer. This adjustment was made since the redshift distribution of quasars within our dataset extends well above $z > 1$, while previously we were limiting the predictions to the $0 < z_{\rm phot} < 1$ range, appropriate for the KiDS-bright sample. Therefore, here a linear activation function is applied in the final layer. The rest of the Hybrid-z architecture remains identical to that of A25. The hyperparameters differ from those in A25 and were selected through empirical testing across various configurations and evaluating the corresponding model's performance on the quasar dataset used in this study. We employed a batch size of 32 and an initial learning rate of 0.001.

As far as data preprocessing is concerned, we adapt the same techniques as in A25; normalization to the image pixel values and standardization of the 9-band magnitude features. Furthermore, we employed data augmentation, including 90° rotation, flipping, and translations (height and width shifts), to expand our training set (Yang et al. 2022). We used $\zeta = \ln(1 + z_{\rm spec})$ as labels (Baldry 2018) to achieve a more balanced distribution in redshift. This transformation of true values helps to maintain stable gradients in the descent-based optimization of the model by ensuring numerical stability. Although the model is trained in this log-transformed space, all evaluations and error metrics are reported after inverting the transformation.

The dataset is divided into training, validation, and testing disjoint sets using the SCIKIT-LEARN library function `train_test_split`, with approximately two-thirds of the data assigned to training and the remainder evenly split between validation and testing subsets. In our case, this corresponds to ∼104k quasars in total, with $\sim 73k$ allocated for training and $\sim 18k$ each for validation and testing.

The performance of the Hybrid-z model is primarily evaluated using the Huber loss function (Huber 1992). The loss curves for both training and validation sets show a steady decline over successive epochs, indicating efficient learning and convergence. The close agreement between the two curves suggests that the model generalizes well to unseen data without signs of overfitting. We used early stopping criteria (Prechelt 2012) to determine the optimal number of epochs. Training ceased automatically after 10 consecutive epochs without improvement, with any reduction in validation loss considered as an indication of improvement. The model was trained over 70 epochs in our model setup. Finally, the estimated QSO photo-zs for the test sample are evaluated by standard statistical metrics. To quantify the statistical accuracy, we calculated the bias, $\delta z = z_{\rm phot} - z_{\rm spec}$, and normalized bias, $\Delta z = \frac{\delta z}{1+z_{\rm spec}}$. Standard deviation of normalized bias $\sigma_{\Delta z}$ and scaled median absolute deviation (SMAD)

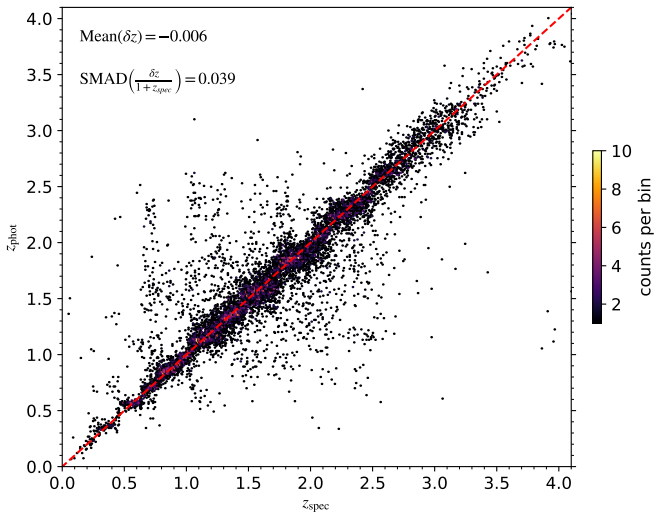


Fig. 1. Density plot comparing spectroscopic and photometric redshifts of KiDS DR4 quasars derived with our Hybrid-z method. The color bar indicates counts per hex-bin for a test sample not seen in the training. The red dashed line corresponds to the identity relation $z_{spec} = z_{phot}$.

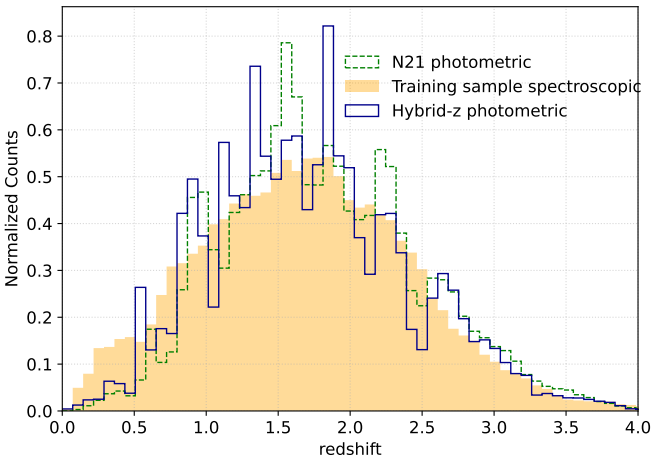


Fig. 2. Comparison of redshift distributions for the KiDS DR4 safe quasar sample. The orange-filled histogram shows the spectroscopic training sample used in our Hybrid-z model. The green dotted line represents the photo-zs of KiDS DR4 safe quasars from N21, while the blue line corresponds to the photo-zs derived in this work using the Hybrid-z framework.

of Δz are the measures of scatter, i.e. statistical precision. Here, SMAD(Δz) = 1.4826 × median($|\Delta z$ − median(Δz)|).

3.1. Results for test data

In this section, we analyze the KiDS DR4 QSO photo-zs derived using the Hybrid-z model for the test sample, with true redshift known, but not seen by the model in training nor validation. Starting with the statistics averaged over the entire test set of about 18,200 quasars drawn from a general cross-match of KiDS with DESI+SDSS spectroscopy, the mean bias is below 0.01 in absolute terms (see first row of Table 1), while the scatter of $\Delta z \equiv (z_{phot} - z_{spec})/(1 + z_{spec})$ is $\sigma_{\Delta z} = 0.133$ (standard deviation) or SMAD(Δz) = 0.039 (scaled median absolute deviation). Therefore, our photo-zs are on average unbiased i.e., the overall scatter is considerably larger than the mean bias. On the other hand, the fact that SMAD is much smaller than SD indicates their non-Gaussian nature with extended tails (outliers), which is also visible in the direct $z_{spec} - z_{phot}$ comparison, presented in Fig. 1, as estimates far apart from the diagonal. This is typical behavior for quasar photo-zs (e.g. Curran 2022; Yao et al. 2023).

In order to directly compare the Hybrid-z performance for KiDS DR4 QSOs with the previous derivations by N21, we cross-matched the above-mentioned test sample with the safe set from that earlier work. This reduces the number of common objects to about 10k mostly because of the $r < 22$ mag limit in that sample. The results are presented in rows 2-3 of Table. 1 and we notice that the Hybrid-z model exhibits a notable improvement over N21 in all the statistics. As we use the N21 safe quasar sample for clustering measurements below, the statistics provided in row 3 of Table. 1 us general quantification of the photo-z performance for our dataset.

We illustrate a direct comparison of our photo-z predictions for the test sample with the true ones in Fig. 1. While the data generally follow the diagonal (identity line $z_{spec} = z_{phot}$), as expected for photo-zs unbiased on average, characteristic 'step-like' behavior is observed in detail. This specific redshift focusing, where a range of spec-zs have similar photo-zs predicted, or vice versa, is commonly observed in quasar photo-z derivations (e.g. Nakoneczny et al. 2021; Kunsági-Máté et al. 2022; Yao et al. 2023; Moss et al. 2025). We interpret these steps as artifacts driven by emission line transitions across filters. For instance, At lower redshifts (e.g., $z \sim 0.4$), the Hα emission line (restframe $\lambda = 6562.8$Å) lies within the Z-band (central $\lambda = 8800$Å), contributing significantly to redshift estimation. As redshift increases, Hα shifts out of that filter coverage, and although Hβ remains visible, this transitional zone introduces increased uncertainty. Around $z \sim 1.1$, Hβ (4861Å) also exits the optical range, and the model increasingly relies on redshifted MgII (2800Å) and CIII] (1908.73Å) emission lines, which can lead to degeneracies due to similar observed colors across different redshifts. These transitions result in redshift confusion, where the model tends to assign many objects the same redshift, producing the observed step-like artifacts in the distribution. Each "step" reflects a regime where the model's redshift estimation is dominated by a different combination of emission lines.

The KiDS quasar clustering analysis presented below in Sec. 4 is done tomographically in four photo-z bins, it is therefore relevant to analyze the redshift performance for each of these bins. This is presented in rows 4-7 of Table.1, where the numbers refer to a cross-match of our test set with the N21 safe quasar sample, binned in Hybrid-z redshift estimates. This time we observe more considerable biases in our photo-zs, with mean $\delta z = z_{phot} - z_{spec}$ surpassing 0.01 in absolute terms, with some improvement if the $1 + z$ scaling is applied (i.e. for Δz). On the other hand, the scatter is better behaved, staying at a similar level for the particular bins as for the full test sample, both in terms of the standard deviation and SMAD of the residuals. This indicates that, once binned in photo-z, our redshift estimates retain their statistical precision, but lose on accuracy.

3.2. Updated KiDS DR4 quasar photometric redshifts

Having checked the Hybrid-z performance for quasars with known redshifts from spectroscopic data, we trained the model

Table 1. Photometric redshift statistics of the Hybrid-z model for quasars in a test sample derived from a cross-match of KiDS DR4 with DESI+SDSS spectroscopy. Row 1 applies to the entire test sample (not seen by the model in training). Rows 2-3 are for a cross-match with the earlier Nakoneczny et al. (2021) derivations: original results in row 2, and our updates in row 3. Finally, rows 4-7 present a break-down into photometric redshift bins which are employed for tomographic angular clustering in this paper

Selection	Size	$\langle z_{\rm spec}\rangle$	$\langle z_{\rm phot}\rangle$	$\langle \delta z\rangle$	$\langle \Delta z\rangle$	$\sigma_{\Delta z}$	SMAD(Δz)
full test set	18,274	1.662	1.656	-0.006	0.008	0.133	0.039
test set cross-matched with the N21 sample:							
previous N21 derivations	10,226	1.672	1.738	0.067	0.038	0.131	0.054
this work	10,226	1.672	1.664	-0.007	0.005	0.116	0.033
$0.1 \le z_{\rm phot} \le 0.8$	718	0.656	0.596	-0.061	-0.019	0.104	0.029
$0.8 < z_{\rm phot} \le 1.2$	1996	1.079	1.018	-0.061	-0.017	0.095	0.030
$1.2 < z_{\rm phot} \le 1.8$	3459	1.510	1.489	-0.021	0.002	0.110	0.038
$1.8 < z_{\rm phot} \le 2.7$	3290	2.110	2.158	0.047	0.027	0.139	0.031

Number of quasars in the subsample.
Entire spectroscopic test set without cross-matching to N21 selection nor any cuts.
Spectroscopic test set cross-matched with the N21 safe quasar sample ($r < 22$ mag).
N21 derived photo-z statistics in the spectroscopic test set cross-matched with the N21 safe quasar sample ($r < 22$ mag).

on the KiDS DR4 × DESI+SDSS cross-match and applied it to the entire safe quasar sample as selected by N21. As already clarified, our feature space are 4-band ugri images and 9-band KiDS+VIKING magnitudes, with a requirement of having detections in each. This dataset, flux-limited to $r < 22$ mag, includes 157,419 quasar candidates on $\sim$ 777 deg^2(effective sky coverage of KiDS DR4 after masking). In Fig. 2 we compare the redshift distribution of the spectroscopic training quasars present also in our photometric set (filled histogram) with two photometric redshift distributions: previous from N21 (green dashed) and the one derived here with Hybrid-z (blue solid). Each of the histograms was normalized to unity under the curve. The comparison indicates that the original relatively smooth $dN/dz_{\rm spec}$ is mapped to $dN/dz_{\rm phot}$ with visible peaks and dips, both for N21 and Hybrid-z derivations. This is another manifestation of the effects already discussed previously in the context of Fig. 1 that quasar photo-zs tend to be focusing around particular redshift values due to redshifted line transitions. This is a caveat one should bear in mind when analyzing the overall photo-z performance for quasars. Indeed, as shown above in Table 1, once broken down into particular redshift bins, the statistics may deteriorate. On the other hand, for the subsequent clustering analysis we chose sufficiently broad photo-z ranges to include neighboring 'peaks' and 'dips' in the redshift distribution and hence partly mitigate the related issues.

4. Angular clustering and bias evolution of KiDS quasars

In this section, we study the clustering of KiDS quasars along with their effective bias, peak height and host halo mass evolution. Our observable is the angular 2PCF measured in photo-z bins and compared to theoretical predictions for the underlying DM field. We also examine two possible systematics – stellar contamination and redshift distribution modeling.

4.1. Measurements: angular correlation function

We use the angular two-point correlation function (2PCF), $\omega(\theta)$, as our observable. The 2PCF quantifies the excess probability of finding a pair of quasars separated by a given distance, here expressed as an angle θ on the sky (Peebles 1973), as we are dealing with photometric data without exact redshifts. To compute the 2PCF, we use the standard Landy & Szalay (1993) estimator:

$$\omega_{obs}(\theta) = \frac{DD(\theta) - 2DR(\theta) + RR(\theta)}{RR(\theta)} \,. \quad (1)$$

Here, $DD(\theta)$, $DR(\theta)$, and $RR(\theta)$ represent the pair counts normalized by number density within an angular bin centered at θ, for the data-data, data-random, and random-random pairs, respectively. The random catalog covers the same footprint as the real data and includes about 100x more points than the quasar sample, for each of the redshift bins used. It is generated by first uniformly distributing points within the range of the minimum and maximum values of right ascension and declination of the data and then applying the KiDS DR4 binary HEALPix mask with $N_{\rm side} = 4096$ to restrict the random dataset to the survey coverage. In this work we do not employ any corrections to the randoms that would compensate for survey systematics. While appropriate methodology has been developed for KiDS (Johnston et al. 2021; Yan et al. 2025), our quasar sample is too sparse for such 'organized randoms' approach to be applicable. Position-dependent variations of survey depth could introduce systematic effects in clustering measurements, which usually lead to the amplitudes being overestimated (e.g. Johnston et al. 2021; Vakili et al. 2023; Yan et al. 2025). However, our sample is limited to $r < 22$ mag, which is well below the KiDS limiting magnitude of $r \sim 25$ mag. In addition, we work with point sources, which should be less sensitive to effects such as varying PSF. Therefore, we do not expect the effective depth to be a dominant source of uncertainty for our analysis.

Our measurements are performed in nine logarithmically-spaced angular bins in the range $0.05° < \theta < 1°$. The lower value is driven by shot noise effects in our relatively sparse dataset,

while the upper one corresponds to very large physical scales once deprojected from angles to distances, and additionally is chosen to mitigate the main KiDS systematic, which is number density variations between tiles of 1 degree at a side. We have checked that indeed the measured $\omega(\theta)$ displays 'jumps' when passing to larger scales, hence we discard them in the analysis. Restricting measurements to scales below 1°, while necessary, will of course not remove the systematics arising from tile-to-tile variations, as correlations between quasars near tile edges will still be present. However, similarly as with the general effect of variable depth, we assess this to be of less importance here as our sample covers the bright end of KiDS. As for small scales, while we do measure the correlations down to arcminutes, the first angular bins are not used to derive the best-fit quasar bias when comparing with theoretical predictions. This is because the corresponding physical scales fall within the non-linear regime where our assumed simple relation between matter and quasar clustering is no longer adequate.

In our analysis we employ four photo-z bins with edges $z_{\rm phot} = \{0.1, 0.8, 1.2, 1.8, 2.7\}$ (Table.1). The minimum and maximum values were chosen to cover the bulk of the redshift distribution (Fig. 2), while the bin widths and divisions are selected to account for photo-z scatter, quasar number density and the artifacts in the redshift distribution, as discussed in Sec. 3.2. The angular 2PCF is computed separately in each photo-z bin using the `TreeCorr` package (Jarvis et al. 2004), and the measurements are shown as points in Fig. 3. The number of sources in each redshift bin is indicated in the legend of each panel in Fig. 3. The error bars attached to the measurements in Fig. 3 are the square root of the diagonal elements of the covariance matrix (Sec. 4.3). Generally, the larger error bars in the first redshift bin than those in the other three bins are primarily driven by the smaller number of sources and the increased relative impact of shot noise in that bin.

In this work, we use only auto-correlations of quasars in the redshift bins. However, due to photo-z errors, there is also clustering signal in bin cross-correlation (Blake et al. 2006; Agarwal et al. 2014; Ho et al. 2015), notwithstanding the magnification lensing effect correlating the bins even for exact redshift tomography (Breton et al. 2022). What is more, photo-z bin cross-correlations could be used as a further diagnostic for systematics, for instance, in case of redshift outliers (Balaguera-Antolínez et al. 2018). However, in this work we chose to focus on auto-correlations only for two main reasons. First, proper modeling of cross-correlations would require deeper knowledge of the underlying true redshift distributions for the bins, which we do not have. Second, such modeling would need to properly account for lensing magnification, which is usually quantified by assuming magnitude-limited sampling (Maartens et al. 2021). While our quasar selection does have a flux cut applied, it is far from magnitude-limited due to complexities of ML-based catalog construction. We are therefore not able to properly model the cross-correlation measurements to either use them as extra information or to evaluate residual systematics.

4.2. Theoretical modeling

We compare the measured angular 2PCF with the theoretical prediction derived for matter clustering to estimate quasar bias. The theoretical angular 2PCF is obtained from the angular power spectrum C_ℓ via the Legendre transformation:

$$\omega_{\rm m}(\theta) = \frac{1}{4\pi} \sum_{\ell=\ell_{min}}^{\ell_{\rm max}} (2\ell+1) C_\ell P_\ell(\cos\theta) \,, \tag{2}$$

where $P_\ell(\cos\theta)$ is the Legendre polynomial of degree ℓ, and the multipole moment ℓ relates to the angular scale θ via $\ell = 180°/\theta$. The sum in the above equation should formally run from $\ell = 1, ..., +\infty$, but we restrict the summation to $\ell_{\rm min} = 30$ and $\ell_{\rm max} = 10^4$ since our angular binning ranges from 0.05° to 1°.

The angular power spectrum (PS) is derived by projecting the three-dimensional PS, $P(k, z)$, along the line of sight, typically using the Limber approximation:

$$C_\ell = \int \frac{{\rm d}z}{H(z)\chi^2(z)} \left(\frac{{\rm d}N}{{\rm d}z}\right)^2 P_{\rm m}\left(\frac{\ell+0.5}{\chi(z)}, z\right), \tag{3}$$

where $H(z)$ is the Hubble parameter and $\chi(z)$ is the comoving radial distance to redshift z. The function ${\rm d}N/{\rm d}z$ represents the normalized redshift distribution of the quasar population for a given redshift bin; ideally this should be the *true* redshift distribution, which we do not know exactly not having spec-zs for the full sample. As discussed below, we will compare two approaches to approximate dN/dz, one by using photo-zs directly and the other by cross-matching with spec-z samples.

The term $P_{\rm m}(k, z)$ is the three-dimensional non-linear matter PS evaluated at the wavenumber $k = \frac{\ell+0.5}{\chi(z)}$ and redshift z. To obtain it, for the matter-only case, we use the Core Cosmology Library (`CCL`; Chisari et al. (2019)) which uses the CLASS algorithm (Blas et al. 2011) for the linear PS and non-linear corrections to the matter PS are applied using the Halofit prescription (Smith et al. 2003; Takahashi et al. 2012). This requires cosmological parameters and here we adopt the flat ΛCDM cosmology with $\Omega_c = 0.2642$ (cold dark matter fraction), $\Omega_b = 0.0493$ (baryonic matter fraction), $n_s = 0.965$ (spectral index), $H_0 = 67.4 {\rm kms}^{-1}{\rm Mpc}^{-1}$ (Hubble constant), $\sigma_8 = 0.811$ (amplitude of matter density fluctuations on a scale of $8\,h^{-1}$ Mpc) and a cosmological constant (Planck Collaboration et al. 2020). In principle, the cosmological model could be varied in the analysis, however our measurements are not sufficiently sensitive to constrain both quasar bias and cosmological parameters, we therefore fix the latter to best-fit Planck values. The resulting $\omega_{\rm m}(\theta)$ curves, computed for each redshift bin using $dN/dz_{\rm phot}$ as the assumed redshift distribution, are shown as black dashed lines in Fig. 3. In Sec. 4.4 we detail how we use these theoretical predictions to derive quasar bias by comparing with the measured 2PCF. First we however discuss in Sec. 4.3 the details of the covariance matrix needed for such calculations.

4.3. Covariance estimation

In order to derive quasar bias by comparing the observed 2PCF with its theoretical counterpart, we need the covariance matrix (CM) quantifying the correlations of $\omega(\theta)$ at different scales, which in our case are particular bins in θ. We estimate the CM directly from the data via the the Bootstrap method (Mohammad & Percival 2022). It is a non-parametric internal error estimation method for a measured quantity, where multiple realizations of the dataset are generated by randomly sampling with replacement. In our analysis, we divided the KiDS DR4 survey footprint into $N_{\rm patch} = 200$ spatial patches, and each bootstrap sample is constructed by randomly selecting $N_{\rm patch}$ patches with replacement, allowing some patches to be repeated while others may be omitted.

For each bootstrap realization, $\omega(\theta)$ is recalculated using the selected patches. This procedure is repeated $N_{\rm boot} = 15,000$ times, resulting in $N_{\rm boot}$ bootstrap realizations of $\omega(\theta)$. $N_{\rm boot}$ is determined by checking that the eigenvalues of the covariance matrix have converged. The bootstrap estimate of the covariance matrix is then computed as:

$$C_{ij} = \frac{1}{N_{\rm boot} - 1} \sum_{k=1}^{N_{\rm boot}} [\omega_k(\theta_i) - \bar{\omega}(\theta_i)] \left[\omega_k(\theta_j) - \bar{\omega}(\theta_j)\right], \quad (4)$$

where $\omega_k(\theta_i)$ and $\omega_k(\theta_j)$ are the angular 2PCF measured in the k-th bootstrap sample at angular bins θ_i and θ_j, respectively, and $\bar{\omega}(\theta_i)$ is the mean value across all bootstrap samples:

$$\bar{\omega}(\theta_i) = \frac{1}{N_{\rm boot}} \sum_{k=1}^{N_{\rm boot}} \omega_k(\theta_i). \quad (5)$$

The inverse of the CM, $C_{ij}^{-1,boot}$, estimated using Bootstrap method is biased because of the finite number of subsamples; that is, $C_{ij}^{-1,boot}$ does not exactly correspond to the true inverse CM. We applied the Anderson-Hartlap-Kaufman (AHK) debiasing factor (Hartlap et al. 2007; Kaufman et al. 2008; Vakili et al. 2023) to correct for this bias in the inverse CM calculation. AHK debiasing factor is defined as:

$$f_{\rm AHK} = \frac{N_{\rm patch} - N_d - 2}{N_{\rm patch} - 1}, \quad (6)$$

where $N_d = 10$ is the number of angular bins used. The corrected inverse CM thus becomes:

$$C_{\rm ij}^{-1,AHK} = f_{\rm AHK} \times C_{ij}^{-1,boot}. \quad (7)$$

This corrected inverse covariance is then used for uncertainty-weighted model fitting to derive the quasar bias. We use the `TreeCorr` package to compute the Bootstrap estimate of the CM. `TreeCorr` employs k-means clustering using the angular positions (right ascension, declination) of quasars to generate the patches for covariance estimation.

4.4. Quasar bias and host halo mass estimation

Here we detail how we derive the quasar effective bias of KiDS quasars, $b(z)$, from the angular correlation measurements. Under the assumption of scale-independent biasing, the observed angular 2PCF of quasars is modeled as a scaled version of the matter $\omega(\theta)$, with the scaling governed by the square of the effective bias parameter:

$$\omega_q(\theta) = b^2(z)\omega_{\rm m}(\theta)\ . \quad (8)$$

In each redshift bin, square of quasar bias $b^2(z)$ is inferred through a chi-squared minimization:

$$\chi^2 = \sum_{i,j} \left[\omega_q(\theta_i) - b^2(z)\omega_{\rm m}(\theta_i)\right]^{\rm T} C_{ij}^{-1,AHK} \left[\omega_q(\theta_j) - b^2(z)\omega_{\rm m}(\theta_j)\right], \quad (9)$$

In this analysis, we fit the observed quasar clustering over the angular range $0.05 \leq \theta({\rm deg}) \leq 1.0$. We employed a Monte Carlo approach to robustly estimate the goodness-of-fit (Fumagalli et al. 2022). We utilized the χ^2 calibration method to account for covariance matrix estimation effects and correlated data points. This is done through moment-matching to a scaled χ^2 distribution. The result is a calibrated reduced χ^2, denoted as $\chi^2_{\rm red}$ in Fig. 3. We determine b^2 in each redshift bin by minimizing the χ^2 statistic, and derive b and its uncertainties from the resulting $\chi^2(b)$ curve. Owing to the asymmetry of $\chi^2(b)$, the inferred bias values have asymmetric 1σ uncertainties. These are presented in Figs. 3 and 5, as well as in Table. A.1.

Given a normalized redshift distribution dN/dz, the effective redshift in each bin is computed as:

$$z_{\rm eff} = \frac{\int_{z_{\rm min}}^{z_{\rm max}} z \left(\frac{{\rm d}N}{{\rm d}z}\right) {\rm d}z}{\int_{z_{\rm min}}^{z_{\rm max}} \left(\frac{{\rm d}N}{{\rm d}z}\right) {\rm d}z} \quad (10)$$

In Fig. 3, the $z_{\rm eff}$ was computed using photometric dN/dz, which we take as our fiducial approach.

The fitted quasar bias increases steadily from $\sim$ 1.6 in the first redshift bin to $\sim$ 4.0 in the last z-bin (Fig. 3 and Table A.1). This trend agrees with the expectations from halo bias models and previous studies (Myers et al. 2007; Ross et al. 2009; Eftekharzadeh et al. 2019). We calculated the effective peak height $\nu_{\rm eff}$ directly from our bias measurements using the Tinker et al. (2010) and Comparat et al. (2017) (hereafter T10 and C17 respectively) bias–$\nu_{\rm eff}$ relations, and derived the effective host halo masses $M_{\rm eff}$ for each redshift bin by inverting their halo mass–bias relations. Table.2 presents the $\nu_{\rm eff}$ and $\log_{10} M_{\rm eff}(h^{-1}M_\odot)$ and their lower and upper bounds at each effective redshift. For both models, $\nu_{\rm eff}$ increases with redshift, and the corresponding halo masses lie in the range $\log_{10} M_{\rm eff}(h^{-1}M_\odot) \simeq 12.7$–$12.9$. These trends are consistent with the widely accepted picture of quasar host halo mass (Shen et al. 2009; Eftekharzadeh et al. 2019; Petter et al. 2023; Pizzati et al. 2024). We note that the selection effects of our quasar sample, particularly at high redshifts, are not accounted for in this analysis.

4.5. Stellar contamination correction

Stellar contamination of the quasar sample is a possibly significant factor affecting the clustering measurements and quasar bias derivation. Stars are expected to have a nearly flat angular 2PCF, in contrast to steep power-law for extragalactic sources such as quasars. Stellar contamination would therefore affect the slope of the measured quasar 2PCF by making it less steep.

In order to inspect and mitigate the stellar contamination effect, we take a similar approach as in Myers et al. (2006). Namely, we model the observed angular 2PCF of our sample as being composed of a true quasar clustering component and that from stars, appropriately weighted by the purity of the quasar sample.[5] As discussed in Myers et al. (2006), it can be shown that this leads to the following expression for true quasar clustering (neglecting cross-terms from star-quasar correlations):

$$\omega_{\rm q}(\theta) = \frac{\omega_{\rm obs}(\theta) - (1 - a^2)\omega_{\rm s}(\theta)}{a^2}\ , \quad (11)$$

where $\omega_{\rm s}(\theta)$ is angular 2PCF of stars, while a is quasar sample purity, for which we take $a = 0.98$ from the KiDS DR4 QSO classification model of N21.

[5] In what follows we assume that only stars can contaminate the quasar sample, i.e. we neglect possible contamination from non-QSO galaxies.

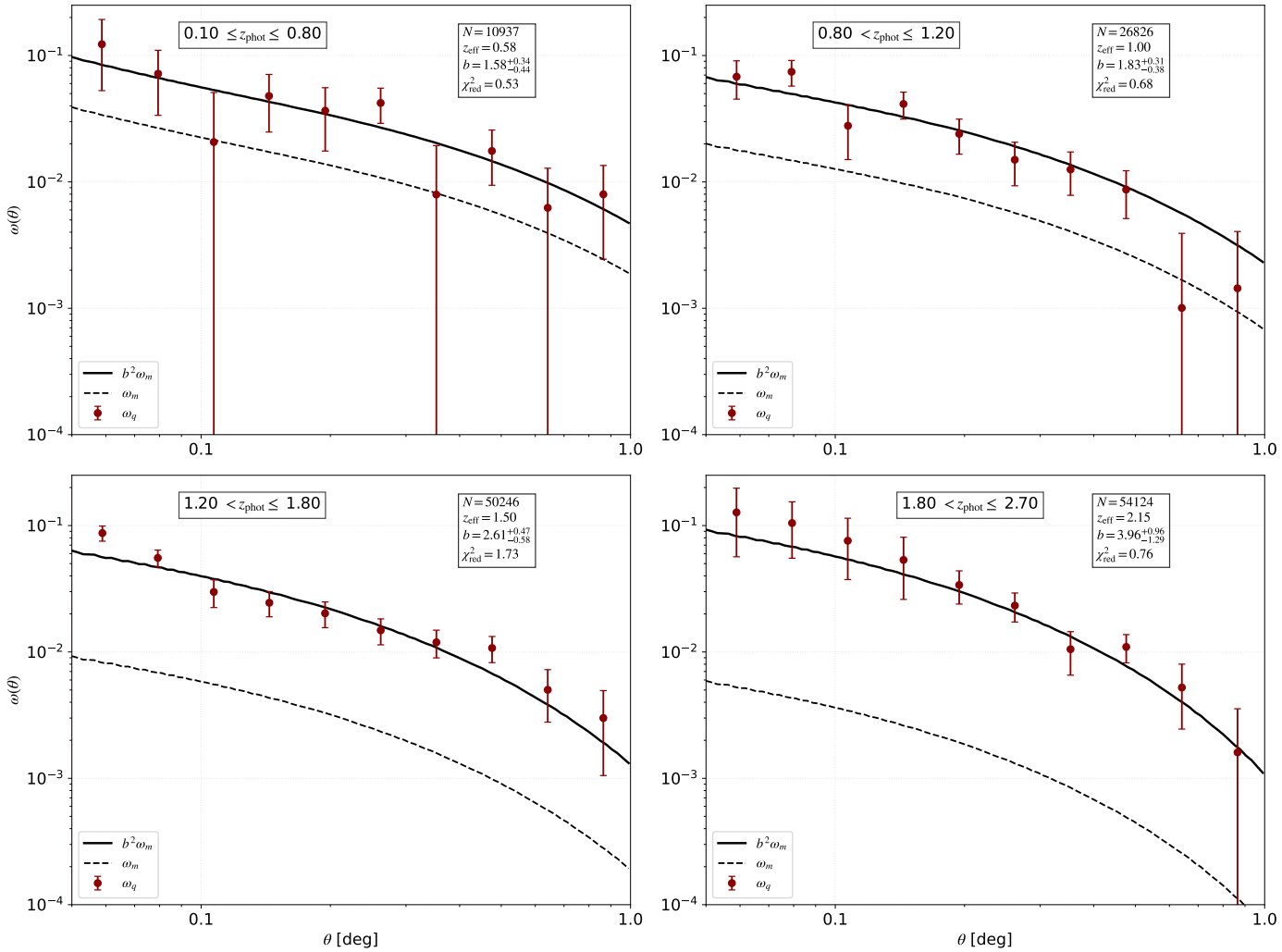


Fig. 3. Angular auto-correlation function of KiDS DR4 quasars measured in four photometric redshift bins. Red data points with error bars are the quasar 2PCF measurements, while the dashed lines show the predicted matter 2PCF $\omega_m(\theta)$, obtained for Planck ΛCDM cosmology and quasar redshift distribution taken as $dN/dz_{\rm phot}$. The solid black line shows $b^2\omega_m$, with the best-fit bias value indicated in the legend. N and $z_{\rm eff}$ are the number of quasar objects and the effective redshift, respectively. The measurements shown here were corrected for stellar contamination using $p_{\rm star} = 0.99$ as discussed in Sec. 4.5.

We measure the star 2PCF ω_s from samples of star candidates provided by the N21 classification model [6]. For that, we select objects based on different stellar classification probability thresholds, $p_{\rm star}$. A threshold of $p_{\rm star} \geq 0.99$ yields a star catalog of higher purity, containing fewer quasars misclassified as stars compared to catalogs with $p_{\rm star} \geq 0.90$ and $p_{\rm star} \geq 0.75$. The autocorrelation function of these stellar samples, $\omega_{\rm s}(\theta)$, was measured over the angular range $0.05° \leq \theta \leq 1.0°$ following the procedure described in Sec. 4.1, and the results are presented in Fig. B.1. Notably, the clustering amplitude of the $p_{\rm star} \geq 0.99$ sample remains approximately flat across the considered angular scales, indicating a low residual clustering signal from extragalactic sources, while the angular 2PCF for the $p_{\rm star} \geq 0.90$ and $p_{\rm star} \geq 0.75$ samples exhibit a clear decline with increasing angular separation.

We used the above-discussed measurements of $\omega_s(\theta)$ for the various $p_{\rm star}$ thresholds in Eq. (11) to correct the quasar clustering measurements for stellar contamination. These corrected measurements were further employed to estimate quasar bias as discussed in Sec. 4.4. We found no significant variation in the bias values across the different $p_{\rm star}$ thresholds. This suggests that our bias estimates are robust to moderate levels of residual stellar contamination. For the final analysis and interpretation, we adopted the stellar catalog with $p_{\rm star} \geq 0.99$ for the stellar contamination correction in the quasar clustering measurements (Fig. 3).

4.6. Impact of the assumed redshift distribution

Bias inference from the angular 2PCF depends on the assumed redshift distribution dN/dz, as this quantifies the projection of the 3D quasar distribution onto the sky, and enters the model quadratically (see Eq. 3). To investigate the impact of the assumed redshift distribution on quasar bias measurements, we compared two approaches. The fiducial one, for which the results are presented in Fig. 3, uses the photometric estimates, $dN/dz_{\rm phot}$, directly. The second method follows Myers et al. (2007): KiDS DR4 quasars in each of the photo-z bins are cross-matched with the DESI+SDSS spec-z quasar sample, which gives $dN/dz_{\rm spec}$ for a given bin. The derived dN/dz are com-

[6] https://kids.strw.leidenuniv.nl/DR4/quasarcatalog.php

Table 2. Effective peak height $\nu_{\rm eff}$ and halo mass, $M_{\rm eff}$ of KiDS quasars obtained by using T10 and C17 models for each redshift $z_{\rm eff}$.

Model	$z_{\rm eff}$	$\nu_{\rm eff}$	$\log_{10} M_{\rm eff}(h^{-1}M_\odot)$
T10	0.58	$1.54^{+0.25}_{-0.41}$	$12.86^{+0.18}_{-0.22}$
	1.00	$1.73^{+0.21}_{-0.29}$	$12.65^{+0.11}_{-0.13}$
	1.50	$2.21^{+0.24}_{-0.34}$	$12.70^{+0.07}_{-0.07}$
	2.15	$2.85^{+0.38}_{-0.61}$	$12.73^{+0.13}_{-0.15}$
C17	0.58	$1.54^{+0.25}_{-0.41}$	$12.93^{+0.17}_{-0.21}$
	1.00	$1.73^{+0.21}_{-0.29}$	$12.68^{+0.12}_{-0.13}$
	1.50	$2.21^{+0.24}_{-0.34}$	$12.72^{+0.07}_{-0.07}$
	2.15	$2.85^{+0.38}_{-0.61}$	$12.76^{+0.13}_{-0.15}$

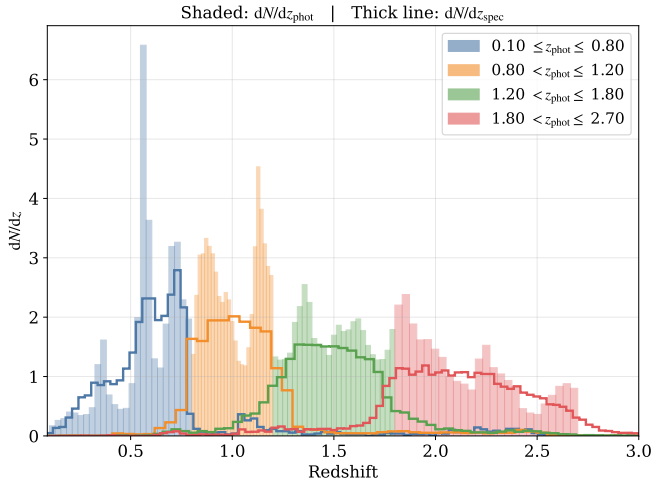


Fig. 4. Quasar redshift distributions for the four tomographic bins used in this study. Shaded histograms are the photo-z distributions of the KiDS DR4 QSO sample, while the thick lines correspond to the spec-z distributions obtained by direct cross-match of the quasars from a given photo-z bin with overlapping DESI and SDSS spectroscopy. Each distribution is normalized to unity under the curve.

pared in Fig. 4,where shaded bars present the photo-z distributions, while the thick lines illustrate $dN/dz_{\rm spec}$. In that plot, each of the tomographic redshift distributions is normalized to unity under the curve to appreciate better the differences in the redshift support for each of the bins.

By design, the $dN/dz_{\rm phot}$ are truncated at bin edges and do not reflect the photo-z uncertainties scattering the objects between the bins. The true redshift distributions will therefore always be broader than the photo-zs suggest. What is more, at higher redshifts, photo-z errors generally increase, leading to a larger fraction of outliers. Consequently, $dN/dz_{\rm phot}$ will deviate more substantially from the true underlying redshift distribution as the redshifts increase.

On the other hand, the spec-z distributions derived via the cross-match are broader than the photometric ones and indicate some outliers at values far apart from bin centers. We should however emphasize that the spec-z distribution derived from the cross-matched sample is not necessarily representative of the true redshift distribution of our entire quasar sample, as the N21 selection may contain quasar candidates not represented by DESI or SDSS in the color space. Therefore, bias values inferred using $dN/dz_{\rm spec}$ may not be inherently more accurate than those obtained from the photo-z distribution of the entire sample.

Both the shape and width of the redshift distribution significantly influence the theoretical predictions of the angular 2PCF. If dN/dz is narrower, the integrand in Eq. (3) results in a higher angular clustering amplitude compared to that of a broader distribution. This in turn will lead to *lower* best-fit quasar bias for the narrower dN/dz for a given redshift bin. This is indeed what we find, as illustrated in Fig.5 and quantified in Table.A.1. Both compare the best-fit effective bias derivations for the KiDS DR4 quasars when using $dN/dz_{\rm phot}$ (blue points in Fig.5) or $dN/dz_{\rm spec}$ (green points) in the modeling. In Fig.5 we additionally show simple phenomenological fits for the the redshift dependence of quasar bias: $b(z) = b_0 + b_1 z + b_2 z^2$ for $dN/dz_{\rm phot}$ and $dN/dz_{\rm spec}$. The best-fit values for b_0, b_1, b_2 are provided in the plot while their errors are listed in Table C.1 and illustrated as shaded bands in the Figure. We note that such quadratic relations fit the datapoints very well. On the other hand, as we do not account for redshift-bin correlations (which would require building a relevant covariance matrix), the fit uncertainties will be somewhat underestimated. We can however firmly conclude that using spectroscopic redshifts to model dN/dz gives bias values larger than the photometric case for the considered redshift range.

These results show that the choice of the underlying redshift distribution has a non-negligible impact on the inferred quasar bias estimation for photometric samples. This emphasizes the importance of reliable dN/dz calibration for clustering-based cosmological analyses with such quasar datasets. Here, we adopt the quasar bias derived from the $dN/dz_{\rm phot}$ of the full sample as our fiducial result, because this redshift distribution corresponds directly to the dataset used in the observed angular 2PCF measurements, unlike the spec-z distribution of the cross-match, which may not be fully representative of the entire sample.

4.7. Comparison with previous derivations of quasar bias

The effective bias of quasars has been measured in a number of past studies and here we compare our results with the literature. For this comparison, we chose those where the large-scale bias was estimated in tomographic bins, both using photometric and spectroscopic redshifts. These are in particular: Croom et al. (2005), who employed the spectroscopic sample from the 2dF QSO Redshift Survey (2QZ, Croom et al. 2004) covering the range $0.3 < z < 2.2$; already discussed Myers et al. (2007), who used photometrically classified quasars drawn from SDSS DR4 spanning $0.4 < z < 2.8$; Ross et al. (2009), who presented bias measurements for spectroscopic quasars from SDSS DR5 within the redshift range $0.3 \leq z \leq 2.2$; and Laurent et al. (2017), who studied eBOSS spectroscopic quasar sample for redshifts $0.9 < z < 2.2$. We note that these past works used different cosmological parameters than us, care should be therefore taken when comparing the bias derivations. The main factor influencing the bias estimate will be different adopted σ_8, which rescales the amplitude of the theoretical power spectrum, and hence directly affects the best-fit effective bias. Moreover, the previous analyses reported $b(z)$ values at the mean redshifts of the adopted bins, which depends on the assumed redshift distributions per bin.

We illustrate the comparison of our results from this paper with the past quasar bias measurements in Fig. 6. We observe very good consistency of our estimates with the 2QZ and SDSS derivations respectively by Croom et al. (2005) and Myers et al.

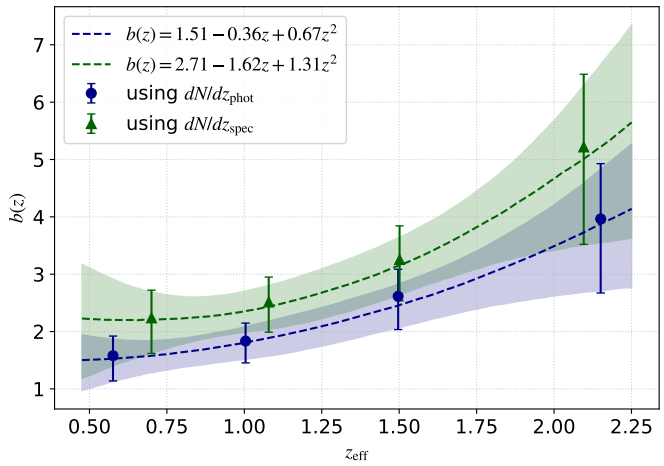


Fig. 5. Effective bias of KiDS DR4 quasars a function of redshift, estimated from angular clustering analysis. We present results from two modeling choices: blue circles were obtained by assuming $dN/dz_{\rm phot}$ directly as the underlying redshift distribution in the photo-z bins, while and for green triangles we used per-bin cross-matches of KiDS photometric quasars with DESI and SDSS spectroscopy. Other model assumptions are the same in both cases. The dashed lines indicate the best-fit models of the form $b(z) = b_0 + b_1 z + b_2 z^2$. The corresponding best-fit parameters are shown in the legend, while their uncertainties are reported in Table C.1. The shaded regions represent the $\pm 1\sigma$ uncertainty bands of the fitted models.

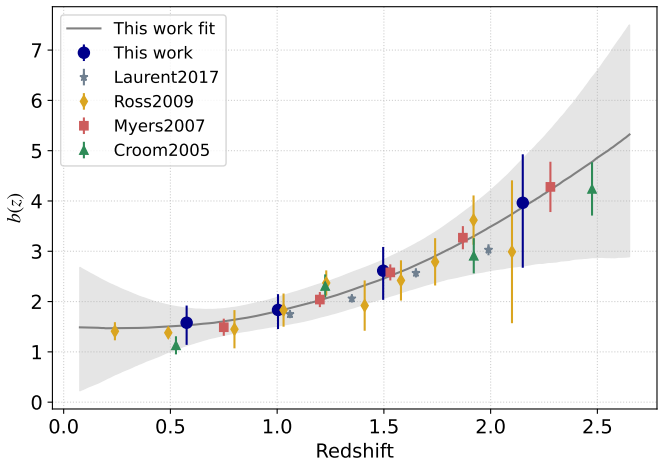


Fig. 6. Comparison of the the effective quasar bias as a function of redshift between this work and previous analyses. The dark blue circles represent the KiDS DR4 measurements from this paper, while the others are respectively: 2QZ from Croom et al. (2005) (upward green triangles); SDSS DR4 from Myers et al. (2007) (red squares); SDSS DR5 from Ross et al. (2009) (orange diamonds); and eBOSS from Laurent et al. (2017) (downward grey triangles). The background grey line with errorband is our best-fit quadratic $b(z)$.

(2007); Ross et al. (2009), while the more recent eBOSS constraints by Laurent et al. (2017) lie systematically below ours, especially for the highest-z bins. This difference is enhanced especially due to considerably smaller errorbars on $b(z)$ in this latter paper as compared to other works, including ours. However, the discrepancy hardly exceeds 2σ even for the most diverging datapoint at $z \sim 2$, which is appreciated especially when compared to our best-fit quadratic model of the form $b(z) = 1.51 - 0.36z + 0.67z^2$ (Sec. 4.4 & Fig. 5). This overall consistency between our quasar bias fits and those from the previous analyses is especially remarkable taking into account various surveys and sample selections, as well as cosmologies used. We conclude that within the explored redshift ranges and scales the bias measurement for quasars is relatively insensitive to such factors.

5. Conclusions

This study provides the first comprehensive measurement of angular clustering and scale-independent effective bias for photometrically classifiedquasars from the Kilo-Degree Survey Data Release 4 (Kuijken et al. 2019; Nakoneczny et al. 2021). Using the original quasar selection from this latter paper (N21), we restricted our analysis to the safe sample, where object features lie within the range of the training data of the QSO classification model. Although we used the previous photometrically classified quasar catalog, we did not adopt the photometric redshifts derived originally by N21. Instead, we applied our new deep learning model, Hybrid-z, developed earlier by John William et al. (2025) Hybrid-z combines KiDS ugri imaging with KiDS+VIKING nine-band photometry (ugriZYJHK$_s$) and in this study it is trained on DESI DR1 and SDSS DR17 spectroscopic quasars. This way we update the approach of N21, who relied solely on SDSS DR14 and used only colors and magnitudes without incorporating imaging information.

Our new model significantly improves the previous KiDS QSO photo-z derivation in practically all the statistics, including the mean bias and scatter of residuals. On the other hand, despite better statistical accuracy and precision of our photo-zs, we still observe the characteristic artefacts in quasar redshift estimates as in past studies: redshift focusing is due to the emission line shifts across photometric filters, resulting in narrow peaks in photometric redshift distributions.

Having tested the Hybrid-z model on spectroscopic data with known redshifts, we trained it on the overlap between KiDS DR4 and DESI+SDSS quasars, and applied it to the entire photometric sample of safe quasars from N21. The final product, used for the clustering analysis, is an updated KiDS DR4 QSO catalog containing about 157k objects on an effective area of ~ 777 deg^2.

We divided our quasar sample into four photometric redshift bins in the range $0.1 \leq z \leq 2.7$, where most of the objects are located. For each of the bins, we measured the quasar angular two-point correlation function (2PCF) over scales from 0.05° to 1°. These measurements were then compared with theoretical predictions for dark-matter 2PCF assuming Planck Collaboration et al. (2020) ΛCDM cosmological parameters. Using a simple bias relation of the form $\omega_Q(\theta) = b^2(z)\omega_m(\theta)$, we obtained best-fit bias values per bin. Our analysis of quasar clustering exhibits a clear increase in bias with redshift, from $b \sim 1.6$ at $z \sim 0.6$ to $b \sim 4.0$ at $z \sim 2.2$. We calculated the effective peak height $\nu_{\rm eff}$ and host halo mass $M_{\rm eff}$ of our quasars.$\nu_{\rm eff}$ increases from ~ 1.5 at $z \sim 0.6$ to ~ 2.85 at $z \sim 2.15$. This trend is consistent with halo bias models and previous observational studies, confirming that quasars observed at higher redshifts reside in progressively more massive dark matter halos.

We have assessed the impact of stellar contamination on our quasar clustering measurements by applying the correction formula from Myers et al. (2006), using angular scale-dependent stellar clustering based on stellar samples with varying classification probabilities. Our analysis shows that the inferred quasar

bias remains largely unchanged across different stellar purity thresholds, indicating that the results are robust to moderate levels of residual stellar contamination. For the final analysis, we adopt the highest-purity stellar sample ($p_{\rm star} \geq 0.99$) to correct the quasar angular correlation function, ensuring reliable bias estimates.

We also investigated how our quasar bias derivations depend on the per-bin redshift distributions employed in the theoretical model. We used two approaches: taking $dN/dz_{\rm phot}$ for each bin directly, or employing as redshift distributions the cross-matches of the bins with spectroscopic quasars from DESI+SDSS. These latter dN/dz are broader than the photometric ones due to photo-z errors, and display tails far away from bin centers related to catastrophic outliers. Consequently, using $dN/dz_{\rm spec}$ yields systematically higher bias values compared to $dN/dz_{\rm phot}$ as the amplitudes of the theoretical predictions for dark matter 2PCF are lower in the former case than in the latter. Nevertheless, the spectroscopic distribution is not fully representative of the entire sample, especially that the original quasar selection from N21 that we employ was trained on SDSS DR14. We therefore adopt the bias derived from the photo-z distribution of the full KiDS DR4 quasar sample as our fiducial result. For the future robust clustering-based cosmological inference from photometric quasar samples, such as from the final KiDS DR5 or forthcoming LSST, accurate characterization of the underlying redshift distribution will be crucial.

Our effective quasar bias measurements for the four effective redshifts of the bins can be modeled as $b(z) = b_0 + b_1 z + b_2 z^2$ with best-fit parameters $(b_0, b_1, b_2) = (1.51, -0.36, 0.67)$. A comparison with the previous derivations of quasar $b_Q(z)$, using samples such as 2QZ (Croom et al. 2005), SDSS: photometric (Myers et al. 2007) and spectroscopic (Ross et al. 2009), as well as eBOSS (Laurent et al. 2017), shows overall consistency with our results, except for the latter work where the highest-z bias estimates depart from our best-fit by more than 2σ. A caveat regarding these comparison is that these past works used different cosmologies than us, which is especially important when σ_8 are not matching, as these directly rescale the matter power spectrum.

In this work, we estimated updated photo-zs for quasars selected from the previous KiDS DR4, and performed the first measurements of KiDS quasar clustering and of their scale-independent effective bias. In the near future, we plan to extend such studies to the final KiDS DR5 (Wright et al. 2024). A quasar selection from that release was already presented in Feng et al. (2025), where, however SDSS data were used to train the classifier, and updating that to DESI should allow us to further improve the previous classification, especially at the faint end. Other avenues to explore it to use all 9 KiDS+VIKING passbands in the convolutional part of the Hybrid-z model, i.e., incorporate both imaging and magnitudes from the whole u to K_s range.

Angular 2PCF measurements give high signal-to-noise down to 0.1° (at least) which could be further explored with more sophisticated theoretical frameworks, such as the halo occupation distribution (Mitra et al. 2018; Eftekharzadeh et al. 2019; Petter et al. 2023; Chowdhary & Chatterjee 2025). Furthermore, both the quasar bias and other cosmological parameters could be also derived by cross-correlating the quasar distribution with CMB lensing (Hirata et al. 2008; Sherwin et al. 2012; Eltvedt et al. 2024; de Belsunce et al. 2025) A crucial ingredient for the robustness of the cosmological constraints when using photometric quasars will be to properly calibrate their redshift distributions, and techniques such as self-organizing maps could be employed for that purpose (Jalan et al. 2024; Wright et al. 2025).

On a longer term, forthcoming big data from, e.g., the Vera Rubin Observatory Legacy Survey of Space and Time (LSST, Ivezić et al. 2019), as well as from the 4-metre Multi-Object Spectroscopic Telescope (4MOST, de Jong et al. 2019) will allow to build and explore new quasar samples over most of the sky. This gives great perspectives not only to understand better their large-scale clustering properties, but also to connect these to smaller-scale environments of the cosmic web.

Code availability

The Deep learning framework for the photometric redshift estimation, named Hybrid-z, will be publicly available at `https://github.com/Anjithajm/Hybrid-z.git`.

Acknowledgements. We thank Angus Wright for his valuable comments and suggestions on the manuscript, and Andrej Dvornik and Ziang Yan for their helpful feedback and discussions during the early stages of this project.
This work is supported by the Polish National Science Center through grants no. 2020/38/E/ST9/00395, and 2020/39/B/ST9/03494.
Based on data products from observations made with ESO Telescopes at the La Silla Paranal Observatory under program IDs 177.A-3016, 177.A-3017 and 177.A-3018, and on data products produced by Target/OmegaCEN, INAF-OACN, INAF-OAPD and the KiDS production team, on behalf of the KiDS consortium. OmegaCEN and the KiDS production team acknowledge support by NOVA and NWO-M grants. Members of INAF-OAPD and INAF-OACN also acknowledge the support from the Department of Physics & Astronomy of the University of Padova, and of the Department of Physics of Univ. Federico II (Naples).
We have made use of TOPCAT (Taylor 2005) software, as well as of PYTHON (`www.python.org`), including the packages NUMPY (Harris et al. 2020), SCIPY (Virtanen et al. 2020), COLOSSUS (Diemer 2018), and MATPLOTLIB (Hunter 2007).

References

Abdurro'uf et al., 2022, ApJS, 259, 35
Abolfathi B., et al., 2018, ApJS, 235, 42
Adame A. G., et al., 2025, J. Cosmology Astropart. Phys., 2025, 012
Agarwal N., et al., 2014, J. Cosmology Astropart. Phys., 2014, 007
Ata M., et al., 2018, MNRAS, 473, 4773
Balaguera-Antolínez A., Bilicki M., Branchini E., Postiglione A., 2018, MNRAS, 476, 1050
Baldry I. K., 2018, arXiv e-prints, p. arXiv:1812.05135
Baum W. A., 1957, AJ, 62, 6
Bilicki M., et al., 2021, A&A, 653, A82
Blake C., Pope A., Scott D., Mobasher B., 2006, MNRAS, 368, 732
Blas D., Lesgourgues J., Tram T., 2011, CLASS: Cosmic Linear Anisotropy Solving System, Astrophysics Source Code Library, record ascl:1106.020
Brescia M., Cavuoti S., D'Abrusco R., Longo G., Mercurio A., 2013, ApJ, 772, 140
Breton M.-A., de la Torre S., Piat J., 2022, A&A, 661, A154
Capaccioli M., Schipani P., 2011, The Messenger, 146, 2
Carrasco D., et al., 2015, A&A, 584, A44
Chaussidon E., et al., 2023, ApJ, 944, 107
Chisari N. E., et al., 2019, ApJS, 242, 2
Chowdhary A., Chatterjee S., 2025, ApJ, 992, 21
Comparat J., Prada F., Yepes G., Klypin A., 2017, MNRAS, 469, 4157
Croom S. M., Smith R. J., Boyle B. J., Shanks T., Miller L., Outram P. J., Loaring N. S., 2004, MNRAS, 349, 1397
Croom S. M., et al., 2005, MNRAS, 356, 415
Curran S. J., 2022, MNRAS, 512, 2099
DESI Collaboration et al., 2025, arXiv e-prints, p. arXiv:2503.14745
D'Isanto A., Polsterer K. L., 2018, A&A, 609, A111
DiPompeo M. A., Hickox R. C., Myers A. D., 2016, MNRAS, 456, 924
Diemer B., 2018, ApJS, 239, 35
Edge A., Sutherland W., Kuijken K., Driver S., McMahon R., Eales S., Emerson J. P., 2013, The Messenger, 154, 32
Eftekharzadeh S., Myers A. D., Kourkchi E., 2019, MNRAS, 486, 274
Eltvedt A. M., Shanks T., Metcalfe N., Ansarinejad B., Barrientos L. F., Murphy D. N. A., Alexander D. M., 2024, MNRAS, 535, 2105
Fan X., Bañados E., Simcoe R. A., 2023, ARA&A, 61, 373
Feng H.-C., et al., 2025, ApJS, 279, 26

Fumagalli A., Biagetti M., Saro A., Sefusatti E., Slosar A., Monaco P., Veropalumbo A., 2022, J. Cosmology Astropart. Phys., 2022, 022
Harris C. R., et al., 2020, Nature, 585, 357
Hartlap J., Simon P., Schneider P., 2007, A&A, 464, 399
Hirata C. M., Ho S., Padmanabhan N., Seljak U., Bahcall N. A., 2008, Phys. Rev. D, 78, 043520
Ho S., et al., 2015, J. Cosmology Astropart. Phys., 2015, 040
Huber P. J., 1992, Robust Estimation of a Location Parameter. Springer New York, New York, NY, pp 492–518, doi:10.1007/978-1-4612-4380-9_35, https://doi.org/10.1007/978-1-4612-4380-9_35
Hunter J. D., 2007, Computing in Science and Engineering, 9, 90
Ivezić Ž., et al., 2019, ApJ, 873, 111
Jalan P., et al., 2024, A&A, 692, A177
Jarvis M., Bernstein G., Jain B., 2004, MNRAS, 352, 338
John William A., Jalan P., Bilicki M., Hellwing W. A., Thuruthipilly H., Nakoneczny S. J., 2025, A&A, 698, A276
Johnston H., et al., 2021, A&A, 648, A98
Kaufman C., Schervish M., Nychka D., 2008, Journal of the American Statistical Association, 103, 1545
Koo D. C., 1985, AJ, 90, 418
Kuijken K., 2008, A&A, 482, 1053
Kuijken K., et al., 2019, A&A, 625, A2
Kunsági-Máté S., Beck R., Szapudi I., Csabai I., 2022, MNRAS, 516, 2662
Landy S. D., Szalay A. S., 1993, ApJ, 412, 64
Laurent P., et al., 2017, J. Cosmology Astropart. Phys., 2017, 017
Lecun Y., Bengio Y., 1995, Convolutional networks for images, speech, and time-series. MIT Press
Luo W., et al., 2024, ApJ, 977, 59
Lyke B. W., et al., 2020, ApJS, 250, 8
Maartens R., Fonseca J., Camera S., Jolicoeur S., Viljoen J.-A., Clarkson C., 2021, J. Cosmology Astropart. Phys., 2021, 009
Marziani P., Dultzin-Hacyan D., Sulentic J. W., 2006, in Kreitler P. V., ed., , New Developments in Black Hole Research. p. 123, doi:10.48550/arXiv.astro-ph/0606678
McCulloch W. S., 1943, Bulletin of mathematical biophysics, 5, 115
Ménard B., Bartelmann M., 2002, A&A, 386, 784
Mitra K., Chatterjee S., DiPompeo M. A., Myers A. D., Zheng Z., 2018, MNRAS, 477, 45
Mohammad F. G., Percival W. J., 2022, MNRAS, 514, 1289
Moss J. P., Curran S. J., Perrott Y. C., 2025, arXiv e-prints, p. arXiv:2507.03260
Myers A. D., et al., 2006, ApJ, 638, 622
Myers A. D., Brunner R. J., Nichol R. C., Richards G. T., Schneider D. P., Bahcall N. A., 2007, ApJ, 658, 85
Nakazono L., et al., 2024, MNRAS, 531, 327
Nakoneczny S., Bilicki M., Solarz A., Pollo A., Maddox N., Spiniello C., Brescia M., Napolitano N. R., 2019, A&A, 624, A13
Nakoneczny S. J., et al., 2021, A&A, 649, A81
Newman J. A., Gruen D., 2022, ARA&A, 60, 363
Pasquet-Itam J., Pasquet J., 2018, A&A, 611, A97
Peebles P. J. E., 1973, ApJ, 185, 413
Peebles P. J. E., 1980, The large-scale structure of the universe
Peterson B. M., 1997, An Introduction to Active Galactic Nuclei
Petter G. C., Hickox R. C., Alexander D. M., Myers A. D., Geach J. E., Whalen K. E., Andonie C. P., 2023, ApJ, 946, 27
Pizzati E., et al., 2024, MNRAS, 534, 3155
Planck Collaboration et al., 2020, A&A, 641, A6
Prechelt L., 2012, Early Stopping — But When?. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 53–67, doi:10.1007/978-3-642-35289-8_5, https://doi.org/10.1007/978-3-642-35289-8_5
Prochaska J. X., Hennawi J. F., 2009, ApJ, 690, 1558
Rees M. J., 1984, ARA&A, 22, 471
Richards G. T., et al., 2002, AJ, 123, 2945
Rosenblatt F., 1958, Psychological Review, 65, 386
Ross N. P., et al., 2009, ApJ, 697, 1634
Roster W., et al., 2024, A&A, 692, A260
Salpeter E. E., 1964, ApJ, 140, 796
Salvato M., Ilbert O., Hoyle B., 2019, Nature Astronomy, 3, 212
Schmidt M., 1963, Nature, 197, 1040
Shen Y., et al., 2009, ApJ, 697, 1656
Sherwin B. D., et al., 2012, Phys. Rev. D, 86, 083006
Slosar A., Hirata C., Seljak U., Ho S., Padmanabhan N., 2008, J. Cosmology Astropart. Phys., 2008, 031
Smith R. E., et al., 2003, MNRAS, 341, 1311
Song H., Park C., Lietzen H., Einasto M., 2016, ApJ, 827, 104
Spilker J. S., Champagne J. B., Fan X., Fujimoto S., van der Werf P. P., Yang J., Yue M., 2025, ApJ, 982, 72
Storey-Fisher K., Hogg D. W., Rix H.-W., Eilers A.-C., Fabbian G., Blanton M. R., Alonso D., 2024, ApJ, 964, 69
Takahashi R., Sato M., Nishimichi T., Taruya A., Oguri M., 2012, ApJ, 761, 152
Taylor M. B., 2005, in Shopbell P., Britton M., Ebert R., eds, Astronomical Society of the Pacific Conference Series Vol. 347, Astronomical Data Analysis Software and Systems XIV. p. 29
Tinker J. L., Robertson B. E., Kravtsov A. V., Klypin A., Warren M. S., Yepes G., Gottlöber S., 2010, ApJ, 724, 878
Vakili M., et al., 2023, A&A, 675, A202
Virtanen P., et al., 2020, Nature Methods, 17, 261
Wright A. H., et al., 2024, A&A, 686, A170
Wright A. H., et al., 2025, arXiv e-prints, p. arXiv:2503.19440
Yan Z., et al., 2025, A&A, 694, A259
Yang S., Xiao W., Zhang M., Guo S., Zhao J., Shen F., 2022, arXiv preprint arXiv:2204.08610
Yao L., et al., 2023, MNRAS, 523, 5799
Zel'dovich Y. B., Novikov I. D., 1964, Soviet Physics Doklady, 9, 246
de Belsunce R., et al., 2025, arXiv e-prints, p. arXiv:2506.22416
de Jong J. T. A., et al., 2015, A&A, 582, A62
de Jong R. S., et al., 2019, The Messenger, 175, 3

Appendix A: Measured bias values

Table A.1. Comparison of effective redshifts ($z_{\rm eff}$) and bias values obtained using photo-z distributions of KiDS DR4 quasar sample ($dN/dz_{\rm phot}$) and spec-z distributions of KiDS DR4 × (DESI DR1 + SDSS DR17) quasar data ($dN/dz_{\rm spec}$) for the theoretical prediction of angular clustering.

photo-z bin	using $dN/dz_{\rm phot}$		using $dN/dz_{\rm spec}$	
	$z_{\rm eff}$	bias	$z_{\rm eff}$	bias
$0.1 \leq z_{\rm phot} \leq 0.8$	0.58	$1.58^{+0.34}_{-0.44}$	0.70	$2.24^{+0.48}_{-0.62}$
$0.8 < z_{\rm phot} \leq 1.2$	1.00	$1.83^{+0.31}_{-0.38}$	1.08	$2.52^{+0.43}_{-0.53}$
$1.2 < z_{\rm phot} \leq 1.8$	1.50	$2.61^{+0.47}_{-0.58}$	1.50	$3.26^{+0.59}_{-0.72}$
$1.8 < z_{\rm phot} \leq 2.7$	2.15	$3.96^{+0.96}_{-1.29}$	2.10	$5.22^{+1.27}_{-1.69}$

Appendix B: Angular clustering of stars

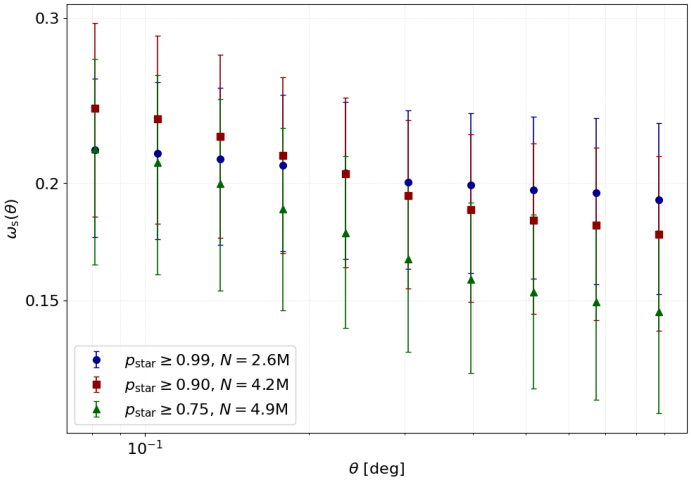


Fig. B.1. Angular two-point correlation function for stellar sources, $\omega_s(\theta)$, measured from 0.05 to 1.0 deg selected by different stellar classification probability ($p_{\rm star}$) threshold. The corresponding number of sources N is indicated in the legend.

Appendix C: Best-fit parameters

Best-fit parameters and corresponding covariance matrix for the quadratic fit,$b_0 + b_1 z + b_2 z^2$, describing the evolution of quasar bias with redshift as follows. Covariance matrix of best-fit parameters when using $dN/dz_{\rm phot}$ is:

$$C_{ij} = \begin{bmatrix} 2.06 & -3.65 & 1.41 \\ -3.65 & 6.84 & -2.75 \\ 1.41 & -2.75 & 1.16 \end{bmatrix}$$

and using $dN/dz_{\rm spec}$ is:

$$C_{ij} = \begin{bmatrix} 7.38 & -12.16 & 4.55 \\ -12.16 & 20.9 & -8.1 \\ 4.55 & -8.1 & 3.26 \end{bmatrix}$$

Table C.1. The parameter uncertainties, corresponding to 1σ deviation, are derived from the square root of the diagonal elements of their covariance matrix.

Parameters	Using $dN/dz_{\rm phot}$	Using $dN/dz_{\rm spec}$
b_0	1.51 ± 1.44	2.71 ± 2.72
b_1	-0.36 ± 2.62	-1.62 ± 4.57
b_2	0.67 ± 1.08	1.31 ± 1.81

Part V

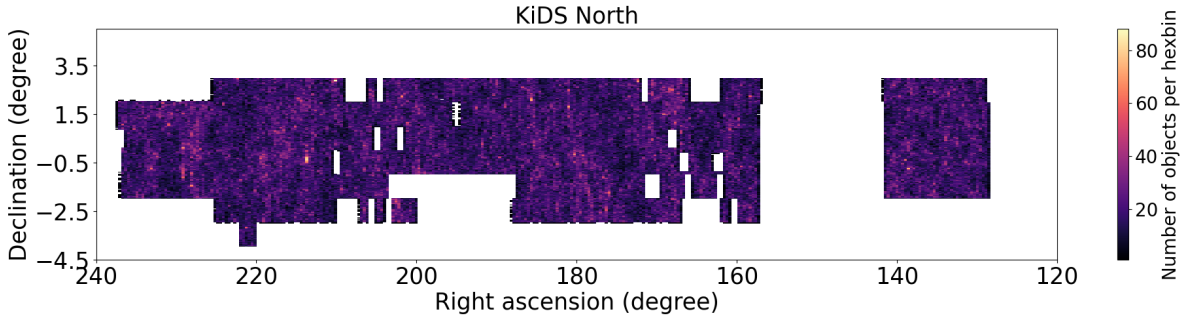
Higher order clustering in the Kilo-Degree Survey bright galaxy sample

In the previous chapter, we analysed the angular clustering of quasars at a given angular scale using the two-point correlation function. This provides insight into effective bias and halo masses. As discussed earlier, this statistic captures only the Gaussian properties of the matter distribution in our Universe. In this chapter, we extend the analysis to probe the non-Gaussian features of the matter distribution.

To this end, we present Count-in-Cells (CiC) measurements of the angular distribution of galaxies. We investigate the clustering strength within an angular cell of radius θ using the angular (or area)- average correlation function $\overline{\omega}_J(\theta)$, and characterise the shape of the galaxy overdensity probability distribution (PDF) on the sky through the reduced cumulants $S_J(\theta)$. The concepts of CiC, $\overline{\omega}_J(\theta)$, and $S_J(\theta)$ were introduced in Section 1.5.5. These statistical tools provide a more complete description of galaxy clustering, capturing not only the amplitude of clustering but also the deviations from Gaussianity in the underlying matter distribution.

In the framework of gravitational instability, the observed galaxy distribution represents a biased realisation of the underlying dark matter field. Since the KiDS survey covers only a fraction of the sky and samples a finite number of galaxies, we adopt the *fair sample hypothesis* [97], assuming that this data set is representative of the large-scale galaxy distribution in the Universe. For this analysis, we focus on the KiDS-DR4 bright galaxy sample selected from the KiDS-DR4 survey footprint [13]. The use of this sample is primarily motivated by the availability of high-quality, deep learning-based photo-zs, as estimated in Chapter 3. This work represents the first CiC-based analysis of higher-order clustering statistics performed using the KiDS-DR4 bright galaxy catalogue. In KiDS, higher-order statistics have also been explored using different techniques, such as the density-split method [20] and peak-count statistics [51].

The large survey area, high-quality photometry, availability of galaxy physical properties, and deep learning-based photo-zs of KiDS make it particularly well suited for investigating the dependence of higher-order clustering on redshift, galaxy colour, and stellar



(a) Northern sky patch

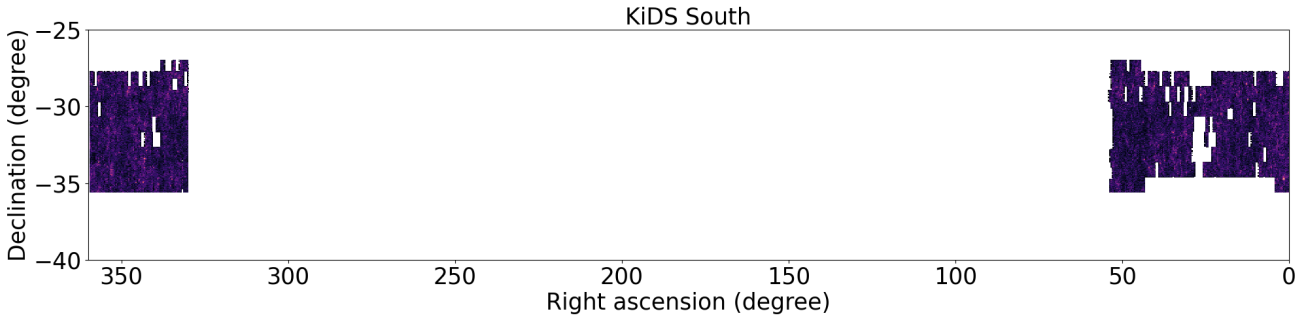


Figure 5.1: Angular footprint of the KiDS-DR4. The data are binned into hexagonal bins (hexbin) in right ascension and declination, with colour indicating the number of objects per bin. The shaded regions indicate the observed fields in right ascension and declination, both given in degrees in the J2000 reference frame.

mass. The significance of this project can be summarised as follows. CiC measurements performed on galaxy samples divided into tomographic redshift bins allow us to trace the evolution of the galaxy overdensity PDF with time. Moreover, the different galaxy population traces the underlying matter distributions in a different manner. To explore this, we divide the galaxy sample into subsamples based on colour and stellar mass. This enables a detailed investigation of how higher-order clustering statistics depend on intrinsic galaxy properties. Such an analysis provides valuable insight into the connection between galaxies and the underlying dark matter distribution.

The chapter is organised as follows. Section 5.1 describes the KiDS-DR4 bright sample dataset and the construction of the galaxy samples used in this analysis. Section 5.2 outlines the methodology adopted to estimate the CiC statistics. Section 5.3 presents the results and analysis. For clarity of presentation, individual data points with error bars are not shown in the main figures of this chapter. The associated uncertainties are discussed in detail in Section 5.3.5.

5.1 Data

The galaxy sample used in this analysis is drawn from the KiDS-DR4 [70] bright galaxy catalogue [13]. KiDS survey details were explained in Chapter 3 & 4. KiDS observations target a square region of the sky of approximately $1\,\mathrm{deg}^2$, referred to as *tiles*. During the observation of a tile, the telescope takes multiple exposures of the same region. These

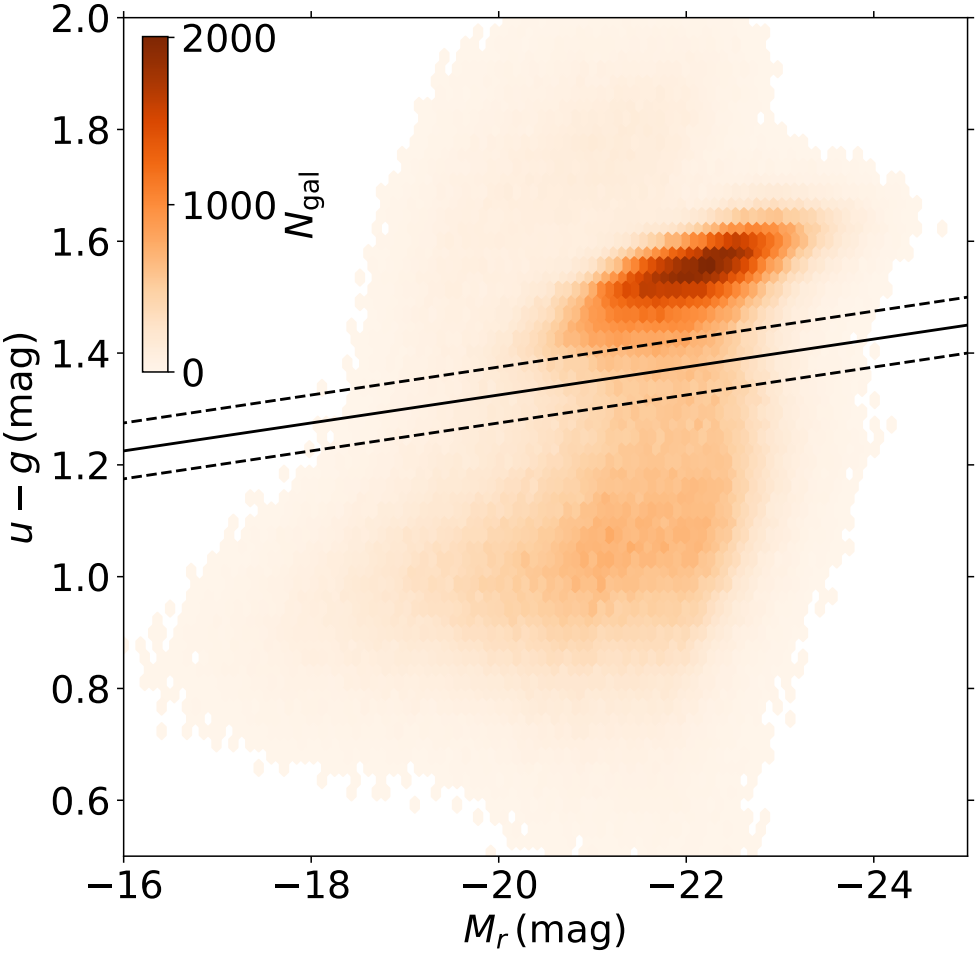


Figure 5.2: colour–magnitude diagram of the KiDS-DR4 bright galaxy sample showing the distribution of galaxies in rest-frame $u-g$ colour versus absolute r-band magnitude M_r. The colour scale indicates the number density of galaxies in each bin. The solid black line represents the empirical relation $u-g = 0.825 - 0.025\,M_r$ used to trace the green valley separating the blue cloud from the red sequence. The dashed lines indicate the adopted boundaries used to define the red, green, and blue galaxy populations. (*Image credit: Andrej Dvornik*)

exposures are taken with small offsets between them using a dithering strategy. After the observations are completed, the individual exposures of the same tile are stacked to produce a single deep image that covers the full $1\,\mathrm{deg}^2$ region[1]. The KiDS data release used in this work consists of 1006 such tiles and provides a sky footprint with an effective area of $\sim 1000\,\mathrm{deg}^2$. The KiDS-DR4 footprint is plotted in Figure 5.1. For more details, refer [70].

The bright sample is defined through a flux limit of $r_{\mathrm{auto}} < 20$, chosen to closely match the depth of the highly complete GAMA spectroscopic survey (see [13]). The flux-limited selection produces about 1.2 million galaxies. The catalogue includes a `masked` flag to exclude artifacts. We selected bright sample sources where `masked=0`, yielding roughly 1 million objects. Photometric redshifts for all objects are those derived in Chapter 3 using the supervised deep-learning framework `Hybrid-z` [64]. The resulting redshift distribution peaks at $z \sim 0.2$ with a median redshift $z_{\mathrm{med}} \approx 0.23$ and photo-z scatter of $0.015(1+z)$. Stellar masses for this catalogue were derived in [13] using the LePhare [4, 61] template fitting code. It is the total stellar mass of a galaxy in units of solar mass. The median of stellar mass is $\log_{10}(M_\star/M_\odot) \approx 10.5$.

The separation between red and blue galaxies is performed using the rest-frame $u-g$ colour and the absolute r-band magnitude (M_r). The colour–magnitude diagram of the sample exhibits the well-known bimodality of the galaxy population, consisting of blue cloud and red sequence galaxies, with a lower-density region between them known as the "green valley". This is shown in Figure 5.2. To determine the boundary between the two populations, an empirical relation was derived from the distribution of galaxies in the $(u-g)$ versus M_r plane [13]. The ridge of the blue cloud was used to determine the slope of the relation, while the location of the minimum galaxy density near $M_r \simeq -19$ was used to anchor the dividing line. This procedure yields the relation

$$u - g = 0.825 - 0.025\, M_r,$$

which traces the green valley in the colour–magnitude diagram. Galaxies are then classified as red or blue according to their position relative to this line, with an additional colour offset applied to avoid contamination from transition objects located within the green valley [13]. Galaxies in this intermediate region are commonly referred to as green galaxies.

5.2 Count-in-Cells estimator

The count-in-cells measurements presented in this chapter were produced using the publicly available `Avcorr` package[2], developed by Paweł Drozda [37, 38]. In this section,

[1] https://kids.strw.leidenuniv.nl/techspecs.php
[2] https://github.com/Pawel-96/Avcorr

we describe the practical implementation adopted for the CiC measurements within the KiDS-DR4 bright sample.

The CiC analysis is performed by randomly placing circular cells of radius θ on the survey footprint and counting the number of galaxies within each cell. This galaxy count in each cell, N, is the fundamental quantity of CiC estimation derived directly from the data. By placing the cells across the survey area, we obtain a distribution of galaxy counts. Angular average correlation function, $\overline{\omega}_J(\theta)$, and reduced cumulants, $S_J(\theta)$, are derived from this count distribution by the mathematical expressions explained in Section 1.5.5.

The angular scales θ are logarithmically spaced between 0.05° and 3.5°. Due to the logarithmic binning used by the `Avcorr` code, the largest sampled angular scale corresponds to $\theta \simeq 2.45°$. The chosen lower angular scale is much larger than the KiDS instrumental resolution to avoid the pixel scale effects. Higher-order moments approach zero for $\theta > 2.45°$, where the measurements become dominated by finite boundary effects rather than intrinsic clustering fluctuations [26, 47, 84].

The KiDS survey footprint consists of three disconnected patches with complex masking due to bright stars and image artifacts (Figure 5.1). This area incompleteness can affect the galaxy count measurements within a cell [40, 28, 105]. In `Avcorr`, the effects of complex survey geometry on CiC measurements are handled using a random catalogue. The random catalogue is constructed to follow the same footprint and mask as the observed data. Instead of generating cell centers uniformly within the survey limits, the centers of the circular cells are drawn from the positions of the random catalogue. The random points follow the same footprint and masking as the data.

The statistical uncertainties are estimated by dividing the full set of cell counts into $n_{\rm reals}$ sub-samples. Each sub-sample contains a subset of the total counts selected without replacement. The statistical uncertainty is determined from the standard deviation across the $n_{\rm reals} = 100$ sub-samples. We tested several values of $n_{\rm reals}$ (50, 100, 150, and 200) to assess the stability of the resampling procedure used for error estimation. The resulting measurements showed no significant variation when increasing $n_{\rm reals}$ beyond 100.

5.3 Results and discussion

In this section, we present the higher-order measurements for the different galaxy subsamples. For clarity of presentation, the results are shown as line plots, which highlight the overall angular scale dependence of the hierarchical moments and allow an easier comparison between the different galaxy subsamples. The measurements, including the individual data points with their corresponding statistical uncertainties are also discussed.

5.3.1 Redshift bins

We divided the KiDS-DR4 bright sample into three photo-z bins with equal width and the bin edges are: $\{0.0, 0.2, 0.4, 0.6\}$. These ranges are chosen to include the majority of

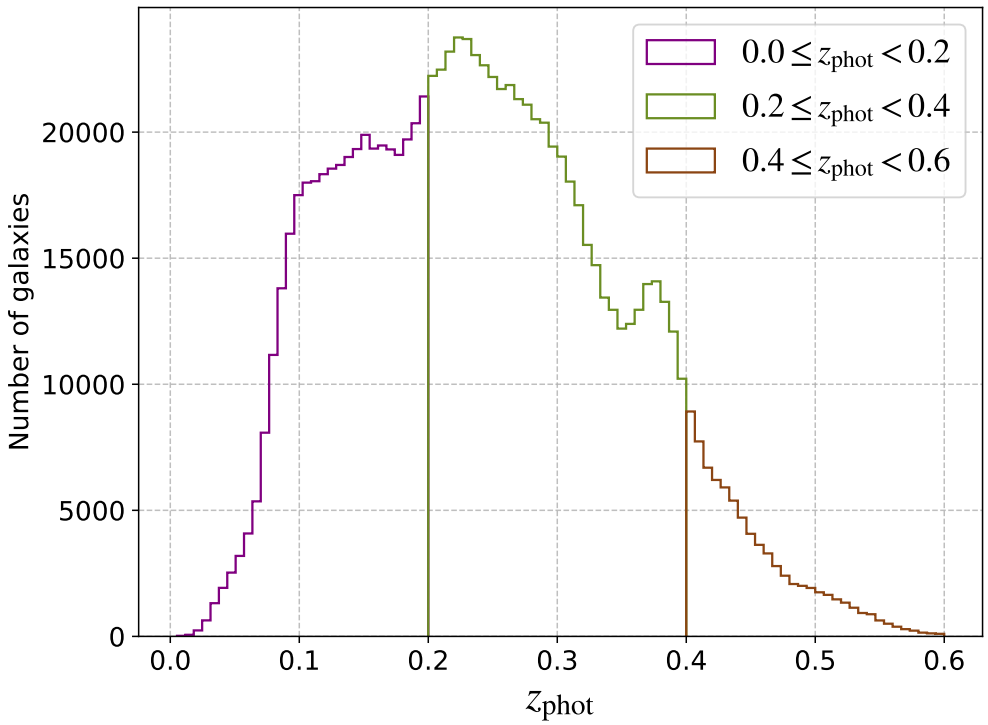


Figure 5.3: Photometric redshift distribution of KiDS-DR4 bright sample in three tomographic bins used in this analysis. Photo-zs are obtained from the deep learning framework `Hybrid-z` [64].

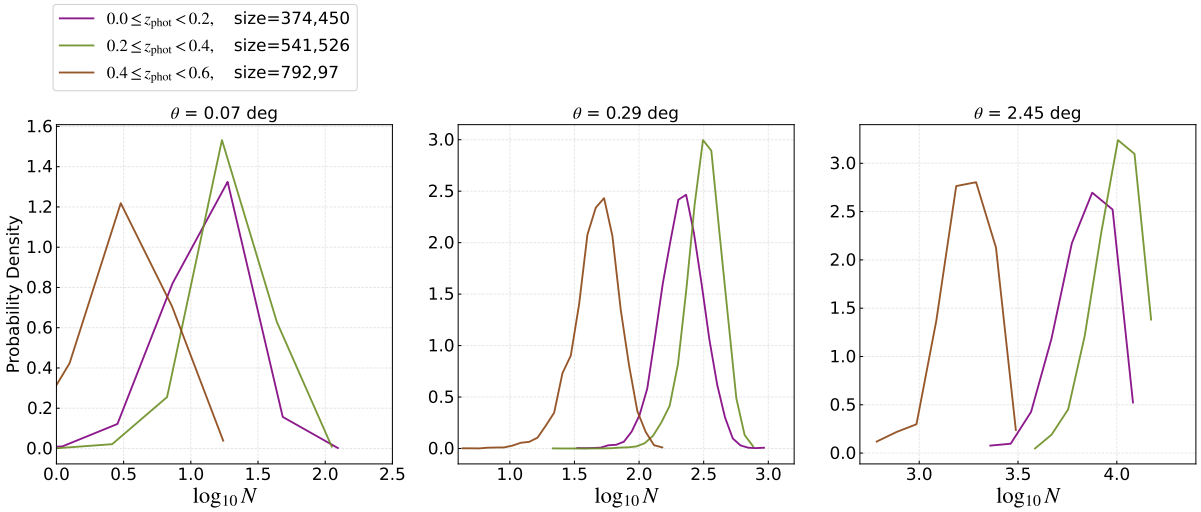


Figure 5.4: Probability density distributions of the logarithm of galaxy counts per cell, $\log_{10} N$, for three photo-z bins: $0.0 \leq z_{\rm phot} < 0.2$, $0.2 \leq z_{\rm phot} < 0.4$, and $0.4 \leq z_{\rm phot} < 0.6$. The distributions are shown for three representative angular scales, $\theta = 0.07°$, $0.29°$, and $2.45°$, corresponding to small, intermediate, and large angular scales, respectively. Size represents the total number of galaxies in each redshift bin.

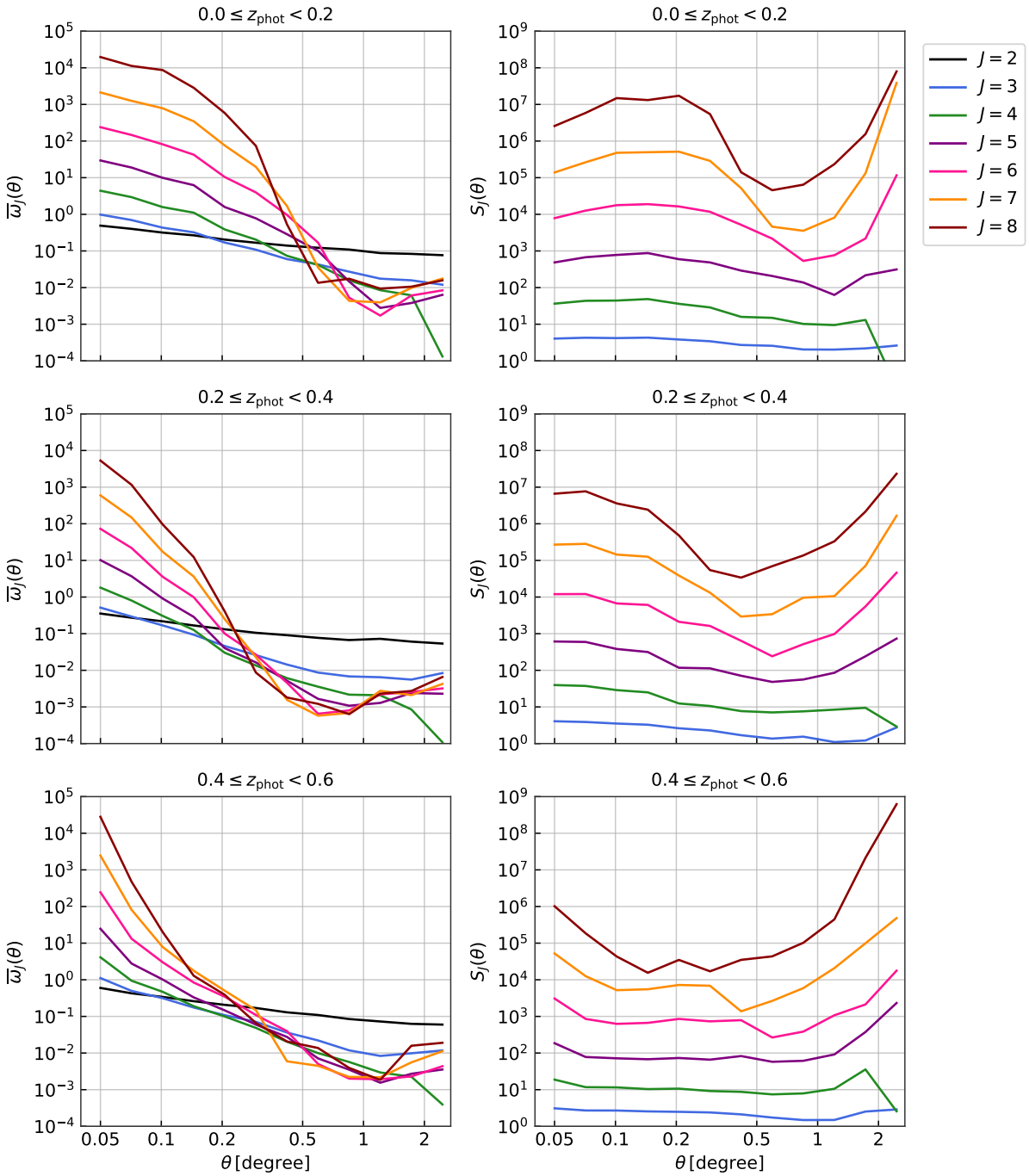


Figure 5.5: Angularly averaged connected moments $\overline{\omega}_J(\theta)$ (left panels) and hierarchical amplitudes $S_J(\theta)$ (right panels) measured for three photometric redshift bins: $0.0 \leq z_{\rm phot} < 0.2$, $0.2 \leq z_{\rm phot} < 0.4$, and $0.4 \leq z_{\rm phot} < 0.6$. Different colours correspond to moment orders $J = 2$–8.

galaxies used in our analysis while ensuring sufficient statistics within each bin. The bin-wise photo-z distribution obtained from is shown in Figure 5.3. The number of galaxies in each three tomographic bin are around $374k$, $541k$, and $79k$, respectively.

`Avcorr` estimated the higher-order moments from the count distribution in each cell of radius θ. Figure 5.4 shows the probability density distributions of the logarithm of galaxy counts per cell, $\log_{10} N$, for three representative angular scales and three photo-z bins. As the angular radius of the cells increases, the distributions shift toward larger $\log_{10} N$ values, reflecting the larger number of galaxies enclosed within bigger apertures. The distributions also exhibit extended high-count tails, especially at smaller and intermediate angular scales. These tails indicate the presence of strong non-Gaussian features in the galaxy density field and contribute significantly to the higher-order moments measured in the count-in-cells analysis. Differences between the redshift bins primarily reflect variations in galaxy number density across the tomographic samples, with lower-redshift bins showing broader distributions due to their higher galaxy density. Overall, the count distribution appears Gaussian in logarithmic space, supporting the log-normal nature of the matter distribution [23].

The left panels of Figure 5.5 display the angularly averaged correlation functions $\overline{\omega}_J(\theta)$ for $J = 2, \ldots, 8$, while the right panels present the corresponding reduced cumulants $S_J(\theta)$ for $J = 3, \ldots, 8$. The angular dependence of $\overline{\omega}_J(\theta)$ is pronounced across all redshift bins, exhibiting a clear decrease with increasing θ. Additionally, the shape of the average correlation function varies between redshift bins. At lower redshifts, the decline is more gradual, in contrast to the steeper drop observed in the higher redshift bins.

A notable feature is that the first tomographic bin exhibits higher values of $\overline{\omega}_J(\theta)$, followed by a slight decrease in the second bin and a subsequent increase in the highest redshift bin. This trend arises from the interplay between geometric volume effects and luminosity-dependent galaxy bias. In the lowest redshift bin ($0 \leq z_{\rm phot} < 0.2$), the high correlation amplitude is primarily driven by physical proximity. At these redshifts, a fixed angular scale θ probes small physical scales within the non-linear regime, where the matter distribution is highly clustered. Moreover, the high level of sample completeness allows the inclusion of relatively faint galaxies, with a median absolute r-band magnitude of $M_r \approx -19.82$. In the intermediate redshift bin ($0.2 \leq z_{\rm phot} < 0.4$), the same angular scale corresponds to a larger physical volume, sampling more linear scales. This effectively smooths density fluctuations, leading to the observed decrease in $\overline{\omega}_J(\theta)$. In the highest redshift bin, the galaxy sample is dominated by intrinsically luminous galaxies with a median magnitude of $M_r \approx -22.19$. These galaxies are strongly biased tracers of the underlying dark matter distribution, resulting in the observed increase in clustering amplitude at higher redshift. The observed non-monotonic redshift trend reflects the competition between projection effects (reducing clustering) and luminosity-dependent bias (enhancing clustering at high redshift).

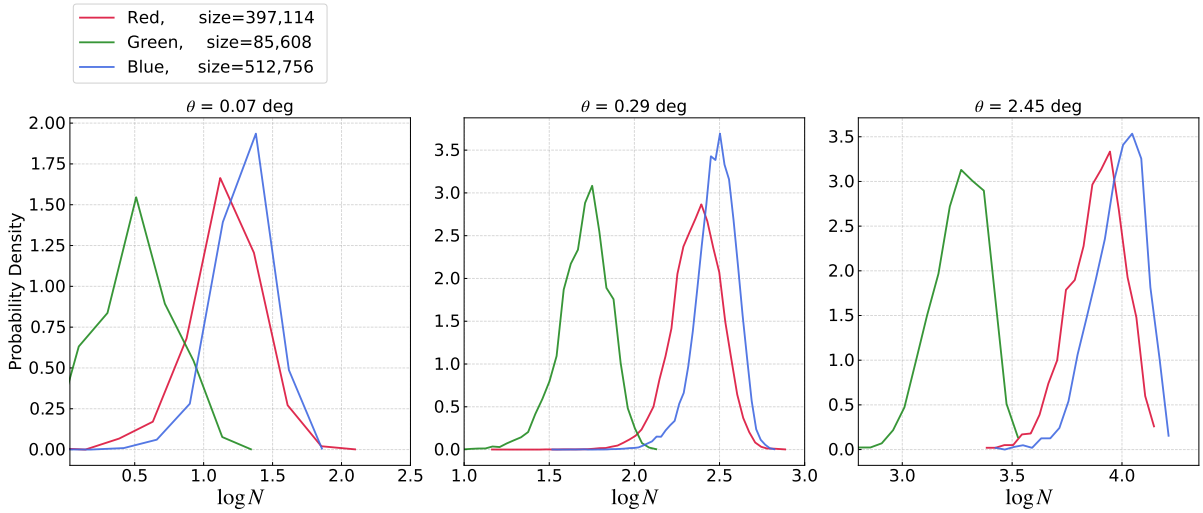


Figure 5.6: Probability density distributions of the logarithm of galaxy counts per cell, $\log_{10} N$, for galaxies separated by colour into red, green, and blue populations. The distributions are shown for three representative angular scales, $\theta = 0.07°$, $0.29°$, and $2.45°$, corresponding to small, intermediate, and large angular scales, respectively.

The redshift evolution of $S_J(\theta)$ is not uniform across angular scales because a fixed angular scale θ corresponds to different physical scales at different redshifts. At low redshift, θ probes small, non-linear scales where the density field is highly non-Gaussian, resulting in larger values of S_J. At higher redshifts, the same θ corresponds to progressively larger, more linear scales, where the density field becomes more Gaussian and S_J decreases. However, this trend is further modulated by galaxy bias and sample selection effects, particularly at high redshift, where luminous galaxies enhance higher-order clustering. As a result, the observed redshift dependence of $S_J(\theta)$ varies with angular scale, reflecting the combined effects of scale-dependent clustering, projection, and galaxy bias [48].

5.3.2 Dependence on colour and stellar mass

We further investigate the CiC statistics for subsamples defined by galaxy colour and stellar mass. The corresponding distributions of galaxy counts per cell are shown in Figure 5.6 & 5.9. As the angular scale increases, the distributions shift toward larger values of $\log_{10} N$, reflecting the increasing number of galaxies enclosed within larger cells. At the same time, the extended high-count tails become less pronounced at larger θ.

The CiC statistics were measured separately for the red, green, and blue galaxy populations. Red galaxies exhibit the strongest clustering signal, followed by green galaxies, with blue galaxies showing the weakest clustering [124, 29]. A similar trend is observed in the non-Gaussianity of the distribution: red galaxies display higher values of the reduced cumulants compared to green and blue galaxies. Red and blue galaxies populate dark matter halos differently. Red galaxies reside in significantly more massive dark matter halos than green and blue galaxies [125, 13]. Red galaxies exhibit both stronger clustering

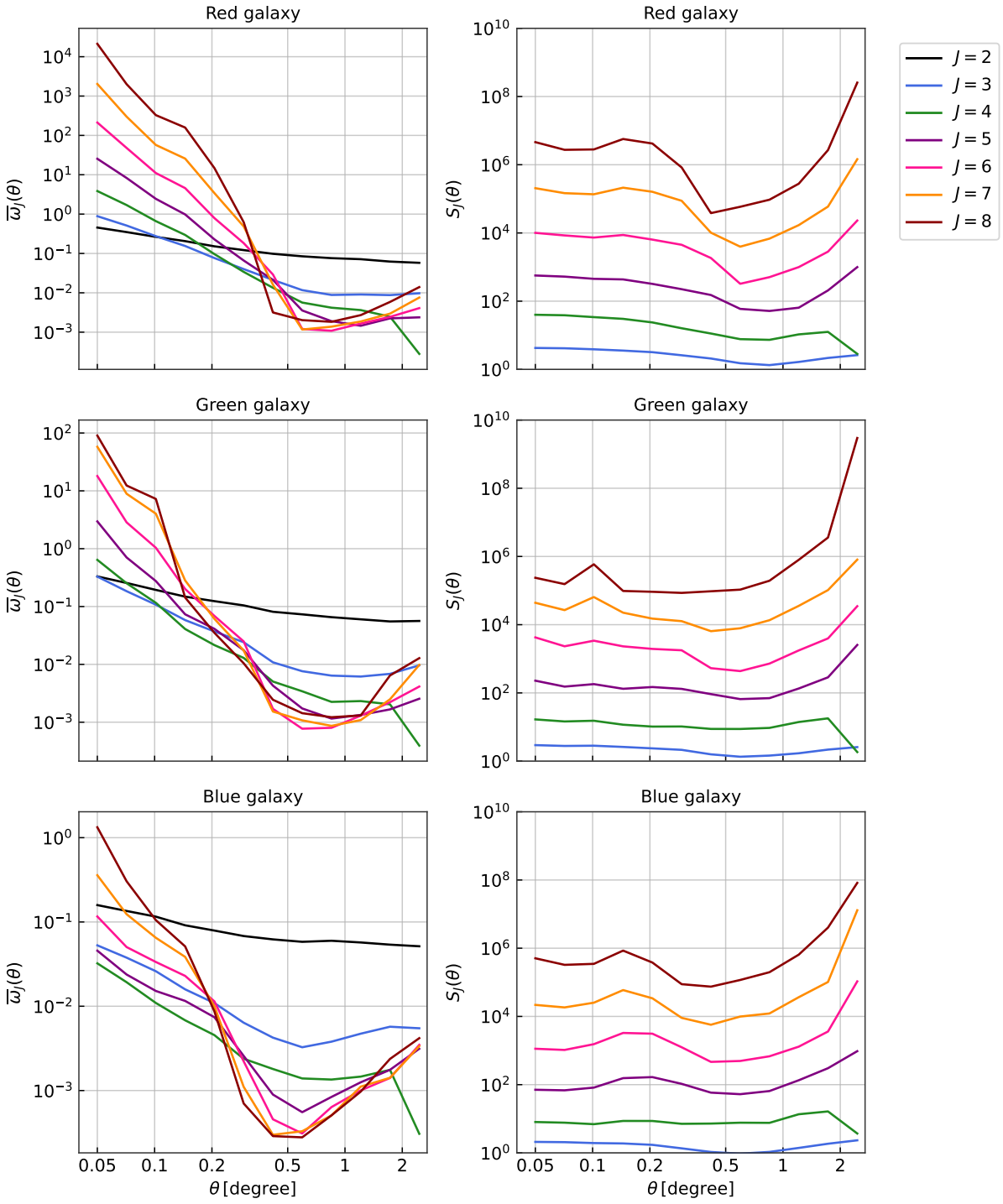


Figure 5.7: $\overline{\omega}_J(\theta)$ (left panels) and $S_J(\theta)$ (right panels) for galaxies separated by colour into red, green, and blue populations. Different colours correspond to moment orders $J = 2$ to 8.

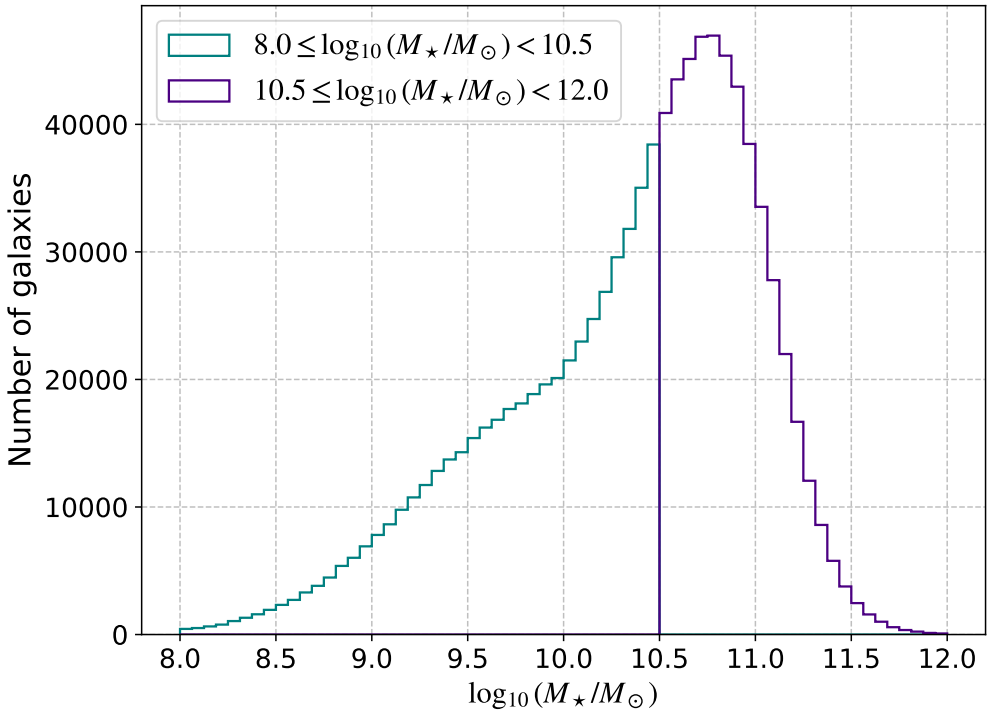


Figure 5.8: Stellar mass distribution of KiDS-DR4 bright sample in two stellar mass bins used in this study. x-axis is the stellar mass in solar mass ($M_\odot$)units. y-axis is the number of objects.

and enhanced non-Gaussianity across all orders relative to blue galaxies. These findings based on angular CiC are in agreement with previous 3D CiC analyses within the 2dFGRS survey [29].

We further divided the KiDS-DR4 bright sample into two stellar-mass bins. The stellar mass distribution is shown in Figure 5.8. The binning strategy is designed to balance statistical robustness and physical interpretability. The median stellar mass of the KiDS-Bright sample is $\log_{10}(M_\star/M_\odot) \approx 10.5$, which naturally separates the galaxy population into lower- and higher-mass systems. We therefore adopt $\log_{10}(M_\star/M_\odot) = 10.5$ as the threshold between the two bins, resulting in subsamples of comparable size ($N = 506k$ and $N = 486k$ galaxies, respectively). The distributions of galaxy counts per cell for the two stellar-mass bins are shown in Figure 5.9. As in the redshift and colour-selected samples, the extended high-count tails become less pronounced with increasing θ, reflecting the averaging of density fluctuations over larger apertures.

Galaxies form and evolve within dark matter halos, where more massive halos generally accrete a larger amount of baryonic matter and host more massive galaxies. However, the relationship between the baryonic properties of galaxies and their host dark matter halos is complex, as it is shaped by a range of astrophysical processes. The Stellar-to-Halo Mass Relation provides a statistical connection between galaxy stellar mass and halo mass [119], and has been extensively studied within KiDS [128, 39, 13]. In this framework,

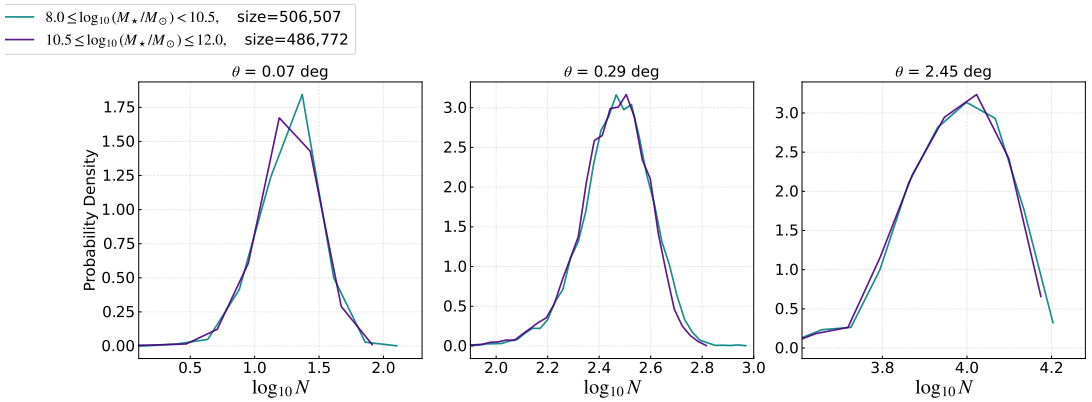


Figure 5.9: Probability density distributions of the logarithm of galaxy counts per cell, $\log_{10} N$, for galaxies divided into two stellar-mass bins: $8.0 \leq \log_{10}(M_\star/M_\odot) < 10.5$ and $10.5 \leq \log_{10}(M_\star/M_\odot) \leq 12.0$. The distributions are shown for three representative angular scales, $\theta = 0.07°$, $0.29°$, and $2.45°$. The number of galaxies in each stellar bin is indicated by size.

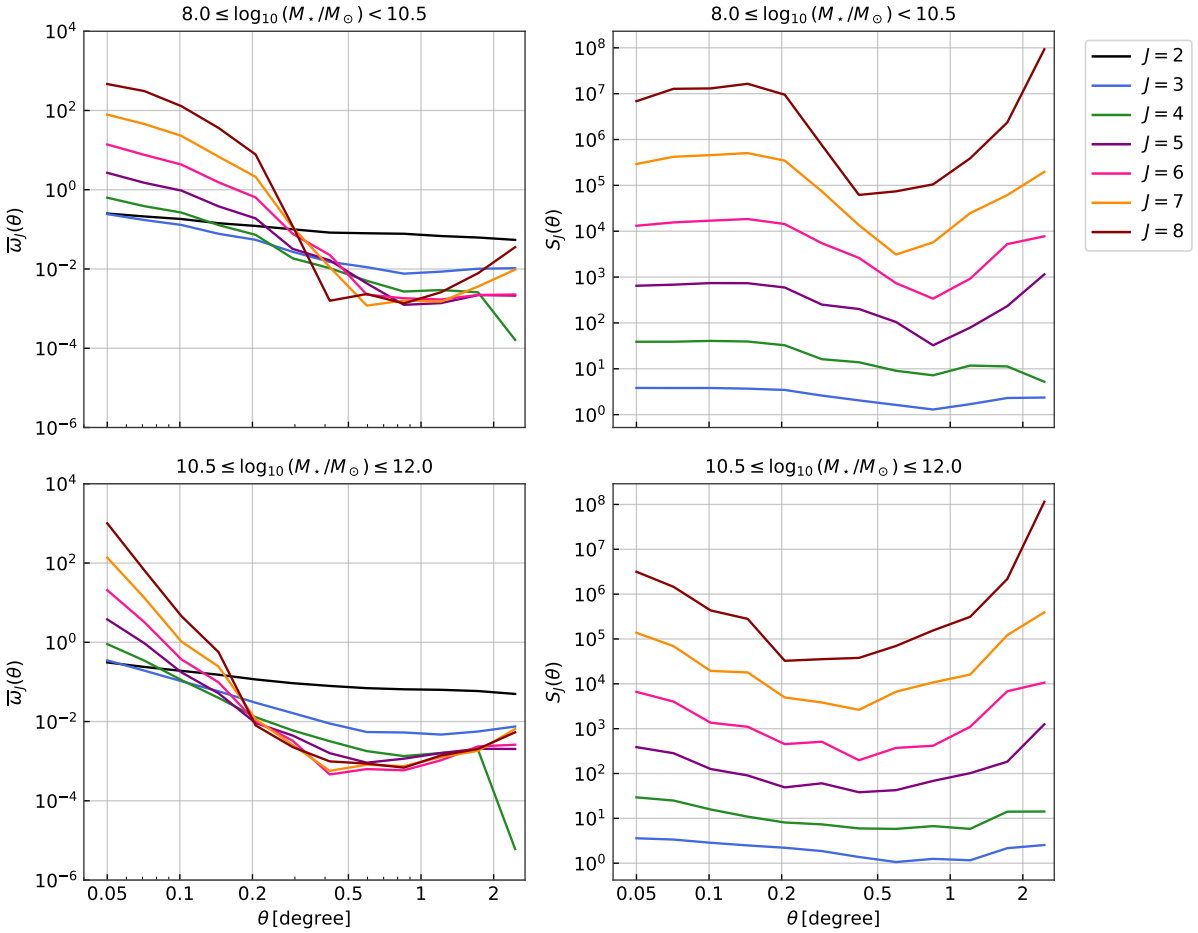


Figure 5.10: Angularly averaged connected moments $\overline{\omega}_J(\theta)$ (left panels) and hierarchical amplitudes $S_J(\theta)$ (right panels) for galaxies divided into two stellar-mass bins: $8.0 \leq \log_{10}(M_\star/M_\odot) < 10.5$ (top panels) and $10.5 \leq \log_{10}(M_\star/M_\odot) \leq 12.0$ (bottom panels). Different colours correspond to moment orders $J = 2$–8.

lower stellar mass galaxies are typically hosted by less massive halos, while higher stellar mass galaxies reside in more massive halos. This relation has direct implications for galaxy clustering, as more massive halos are more strongly biased tracers of the underlying dark matter distribution. Consistent with this expectation, our CiC measurements for stellar mass–binned samples show that galaxies with higher stellar mass exhibit stronger clustering compared to their lower stellar mass counterparts.

While both the stellar mass binned samples and the red galaxy population trace massive dark matter halos, resulting in similarly high clustering amplitudes. However, their clustering patterns differ in shape due to the underlying physical processes. The red galaxy population is characterised by a higher satellite fraction [100]. As a result, their clustering is more strongly influenced by intra-halo processes, with a substantial contribution from satellite galaxies residing within the same dark matter halo. This enhances the small-scale clustering signal and gives rise to a steeper, more non-linear clustering behaviour. In contrast, the high stellar mass sample is primarily dominated by central galaxies [11] that trace the distribution of distinct dark matter halos. As a result, their clustering is governed more by inter-halo correlations, producing a smoother and flatter clustering profile that reflects the large-scale structure of the cosmic web. This can be seen in Figure 5.7 & 5.10.

5.3.3 Transition to Gaussian behaviour

The dominance of the second-order moment, $\overline{\omega}_2$, shows a clear dependence on redshift (Figure 5.5). In higher redshift bins, $\overline{\omega}_2$ begins to overtake the higher-order moments at relatively smaller angular scales compared to the low-redshift bin, indicating a more rapid suppression of higher-order contributions. This reflects the fact that, at higher redshifts, a fixed angular scale probes larger physical scales, where the density field is smoother and closer to Gaussian.

A similar behaviour is observed across different galaxy populations. Blue and green galaxies exhibit an earlier dominance of $\overline{\omega}_2$ compared to red galaxies, where higher-order moments persist down to smaller angular scales (see Figure 5.7). Likewise, in stellar mass–selected samples, $\overline{\omega}_2$ starts to dominate at larger angular scales for high stellar mass galaxies compared to lower stellar mass galaxies (Figure 5.10). Together, these trends highlight the combined influence of scale, redshift, and galaxy properties on the suppression of higher-order clustering and the transition toward Gaussian behaviour.

5.3.4 Upturn behaviour

At the largest angular scales, the reduced cumulants show an artificial rise in all subsamples. In particular, we observe a significant rise in $S_J(\theta)$ for $J > 4$ at $\theta \gtrsim 1^\circ$ since higher-orders are more sensitive to the non-linear tails [10]. This behaviour can arise not

only from finite survey effects in our galaxy sample but also from the presence of superstructures such as clusters. The similar upturn in $S_J(\theta)$ is also observed by past studies [48, 114, 5, 28]. The detailed study of disentangling these effects is an ongoing work. The KiDS survey is conducted with OmegaCAM and the tile size is approximately 1°. At these scales, circular apertures begin to cross tile boundaries, making the measurements sensitive to small calibration residuals and masking patterns at the edges of the dithered exposures. The appearance of this deviation at the same hardware-defined scale in all redshift bins strongly suggests that the signal is dominated by the survey tiling strategy rather than by cosmological fluctuations. Although the CiC method averages over many cells, the higher-order moments remain highly sensitive to localised discontinuities. When apertures cross the $\sim 1°$ boundaries of the OmegaCAM tiles, small systematic variations in galaxy counts are introduced due to localised survey inhomogeneities. This effect is further amplified in the hierarchical amplitudes. Fluctuations in the cell counts are raised to high powers in the numerator (up to $J = 8$), while the denominator $\overline{\omega}_2(\theta)$ simultaneously decreases on large scales. This combination significantly amplifies the influence of outliers, producing the pronounced non-physical rise observed in all redshift bins at $\theta \gtrsim 1°$. This demonstrates that higher-order statistics are particularly sensitive to survey geometry and observational systematics. Our analysis also strengthened the past study conclusion that the clustering measurements depend on sample construction [93].

5.3.5 Uncertainties

Higher-order clustering and its uncertainties are shown in Figure 5.12, 5.13 & 5.11. The uncertainties in the CiC measurements show a strong dependence on galaxy sample, redshift, stellar mass, and moment order. Overall, the error bars increase systematically with increasing moment order J, reflecting the intrinsically higher variance of higher-order statistics. While the second-order moment ($J = 2$) is relatively well constrained across all samples and angular scales, higher-order moments ($J \geq 5$) exhibit substantially larger uncertainties, particularly at small and large angular scales.

A clear dependence on galaxy type is observed. The red galaxy sample, despite having a smaller number of objects compared to the blue sample, shows relatively moderate uncertainties, especially at intermediate angular scales. In contrast, the green galaxy sample exhibits significantly larger error bars, which can be attributed to its smaller sample size. The blue galaxy sample, being the largest, generally provides tighter constraints, particularly for lower-order moments, although uncertainties still grow rapidly for higher-order moments.

The uncertainties also vary with stellar mass bins. The lower stellar mass sample shows relatively smaller error bars due to its larger number of galaxies, whereas the higher stellar mass sample exhibits increased uncertainties. This is expected, as massive galaxies are rarer, leading to larger shot noise and reduced statistical precision. This effect

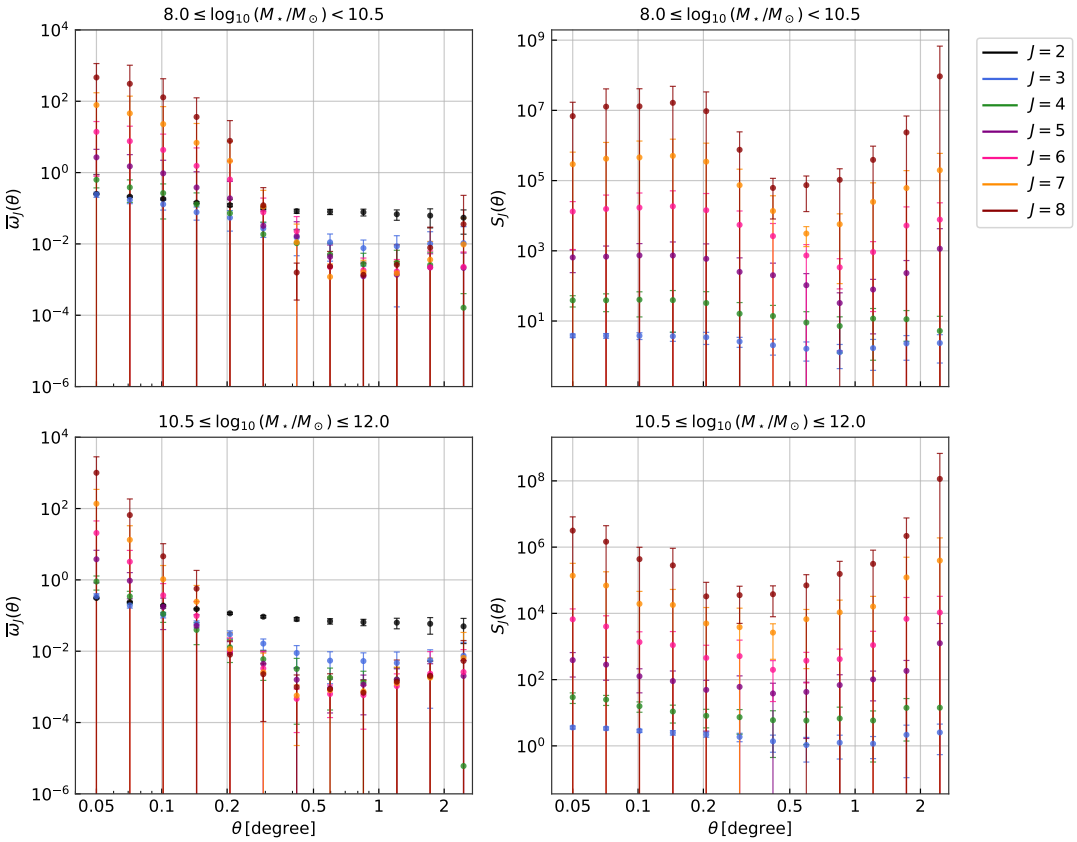


Figure 5.11: Angularly averaged connected moments $\overline{\omega}_J(\theta)$ (left panels) and hierarchical amplitudes $S_J(\theta)$ (right panels) for galaxies divided into two stellar-mass bins: $8.0 \leq \log_{10}(M_\star/M_\odot) < 10.5$ (top panels) and $10.5 \leq \log_{10}(M_\star/M_\odot) \leq 12.0$ (bottom panels).

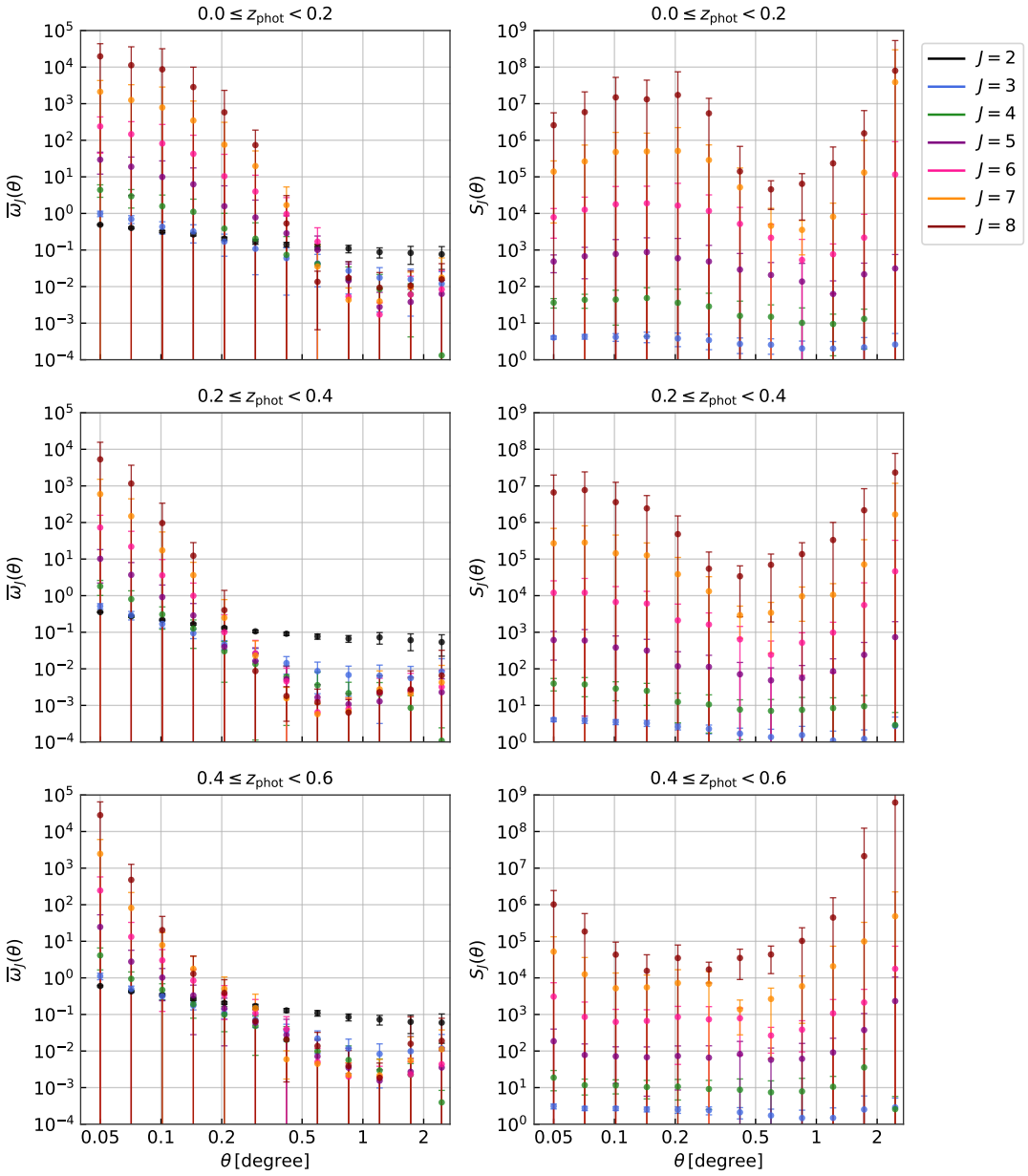


Figure 5.12: Angularly averaged connected moments $\overline{\omega}_J(\theta)$ (left panels) and hierarchical amplitudes $S_J(\theta)$ (right panels) measured using the count-in-cells method for three photometric redshift bins: $0.0 \leq z_{\rm phot} < 0.2$, $0.2 \leq z_{\rm phot} < 0.4$, and $0.4 \leq z_{\rm phot} < 0.6$. Different colours correspond to moment orders $J = 2$–8.

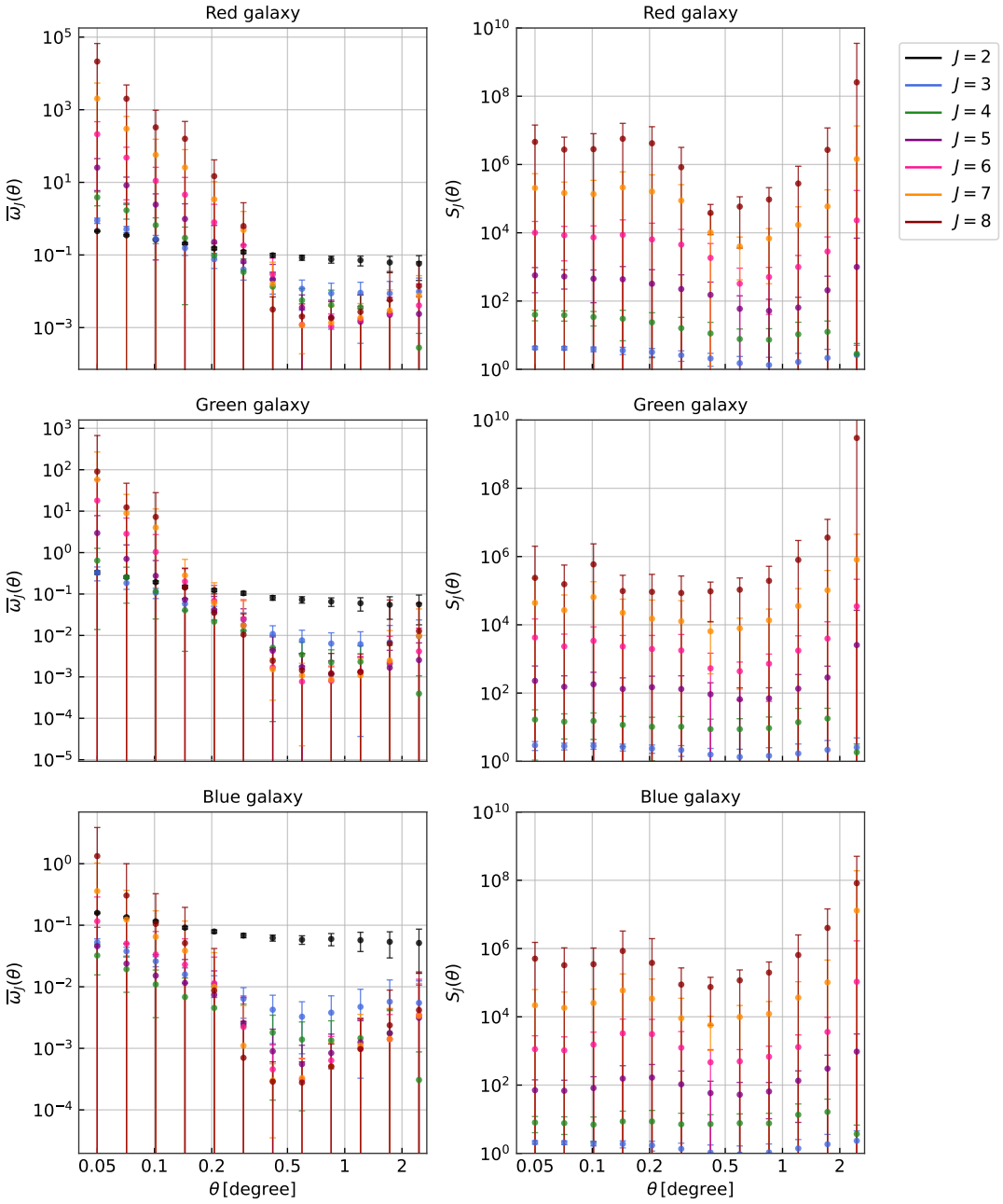


Figure 5.13: Angularly averaged connected moments $\overline{\omega}_J(\theta)$ (left panels) and hierarchical amplitudes $S_J(\theta)$ (right panels) measured for galaxies separated by colour into red, green, and blue populations.

is particularly pronounced for higher-order moments, where the measurements become increasingly sensitive to rare high-density fluctuations.

A similar trend is observed across redshift bins. The lowest redshift bin, which contains a relatively large number of galaxies, shows comparatively smaller uncertainties. As redshift increases, the number density decreases, leading to larger error bars. The highest redshift bin, with the smallest sample size, exhibits the largest uncertainties, especially for higher-order moments and at larger angular scales.

In addition to sample size effects, the uncertainties increase toward both small and large angular scales. At small angular scales, the increase is driven by shot noise and the discrete nature of galaxy counts, while at large angular scales, the uncertainties are dominated by cosmic variance and survey geometry effects. The combination of these factors results in the characteristic U-shaped error behaviour as a function of angular scale.

Overall, the uncertainties are consistent with expectations from Poisson noise, sample variance, and the hierarchical nature of higher-order statistics. While the second- and third-order moments are robustly measured across most samples, caution is required when interpreting higher-order moments, particularly in low-number-density subsamples such as high stellar mass and high-redshift bins.

5.4 Summary

In this chapter, we investigated higher-order clustering of galaxies in the KiDS-DR4 bright sample using the Count-in-Cells formalism, enabling a characterisation of the non-Gaussian features of the galaxy distribution beyond the two-point statistics.

We first examined the dependence of clustering on galaxy subsamples defined by redshift, colour, and stellar mass. The angular averaged correlation functions $\overline{\omega}_J(\theta)$ show a strong scale dependence across all subsamples, decreasing with increasing angular scale. A non-monotonic evolution with redshift is observed, reflecting the interplay between projection effects and luminosity-dependent galaxy bias. At low redshift, strong clustering arises from small-scale non-linear structures, while at intermediate redshift the signal is suppressed due to larger physical scales being probed. At higher redshift, the dominance of intrinsically luminous, highly biased galaxies enhances the clustering amplitude.

Clear trends are also seen with galaxy properties. Red galaxies exhibit the strongest clustering amplitudes and highest values of reduced cumulants $S_J(\theta)$, followed by green and blue galaxies. Similarly, galaxies with higher stellar mass show stronger clustering than their lower-mass counterparts. These trends are consistent with expectations from the stellar-to-halo mass relation, where red and massive galaxies reside in more massive dark matter halos. Furthermore, differences in the shape of the statistics suggest that red galaxies are more influenced by intra-halo (satellite-driven) clustering, leading to

enhanced small-scale non-Gaussianity, whereas high stellar mass galaxies more strongly trace large-scale (inter-halo) structure.

We then investigated the transition toward Gaussianity through the hierarchy of moments. The second-order moment, $\overline{\omega}_2$, progressively dominates over higher-order moments at larger angular scales, with this transition occurring earlier (i.e., at larger θ) for blue and green galaxies compared to red galaxies, and for high stellar mass samples compared to lower-mass ones. A similar trend is observed with redshift, where the dominance of $\overline{\omega}_2$ shifts to smaller angular scales at higher redshift. This behaviour reflects the combined effects of scale-dependent clustering and projection: larger angular cells probe more linear regimes and average over small-scale fluctuations, leading to a suppression of higher-order, non-Gaussian contributions.

Finally, we observe a noticeable upturn in both $\overline{\omega}_J(\theta)$ and $S_J(\theta)$ at large angular scales ($\theta \gtrsim 1^\circ$). This feature is unlikely to be of purely physical origin and may instead be driven by systematic effects such as survey geometry, edge effects, or the presence of large structures such as clusters. The consistency of this upturn across different subsamples suggests a common origin, highlighting the importance of carefully accounting for observational systematics in higher-order clustering analyses.

Overall, this chapter demonstrates that CiC statistics provide a powerful and complementary approach to probing the non-Gaussian nature of large-scale structure, while also emphasising the sensitivity of higher-order moments to both astrophysical effects and observational systematics.

5.5 Future directions

The analysis presented in this chapter provides an initial exploration of higher-order clustering statistics in the KiDS-DR4 bright galaxy sample. Several avenues can further extend this work and fully exploit the cosmological information contained in the Count-in-Cells framework.

First, a more comprehensive treatment of survey systematics is essential to improve the robustness of measurements, particularly on large angular scales. The artificial upturn observed at $\theta \gtrsim 1^\circ$ indicates that higher-order statistics are highly sensitive to observational effects and survey geometry. Future analyses will incorporate improved masking strategies, tile-weight corrections, and alternative estimators that mitigate boundary effects. In addition, mock galaxy catalogues derived from cosmological simulations will be crucial for quantifying the impact of survey geometry and for validating the CiC measurement pipeline.

Furthermore, the clustering and non-Gaussianity of red, green, and blue galaxies can be explored in greater detail by subdividing each population into stellar mass bins. Previous studies [13] have shown that red galaxies reside in more massive dark matter halos than blue galaxies at fixed stellar mass. Testing this result within the CiC framework

would provide an independent validation of these findings. In addition, identifying central and satellite galaxies would help disentangle the relative contributions of halo mass, environment, and galaxy type to the observed clustering signal. All these analyses must be accompanied by careful control of systematics to avoid confusion between physical and observational effects.

The CiC measurements presented here offer a promising route to constraining non-linear galaxy bias. Higher-order clustering statistics provide unique sensitivity to the relationship between the galaxy and matter density fields [47, 10, 105]. In particular, the reduced cumulants S_J encode information about higher-order bias parameters that cannot be accessed through two-point statistics alone. By comparing the observed CiC measurements with predictions from dark matter fields in cosmological N-body simulations, it will be possible to constrain scale-dependent, non-linear galaxy bias.

Higher-order statistics are powerful probes of deviations from General Relativity. Modified gravity models can alter the non-linear growth of structure, leading to measurable changes in higher-order moments of the density field [55, 2, 38]. Since the reduced cumulants S_J are sensitive to non-Gaussian features, they provide a promising observable for testing such theories. Future work will therefore focus on comparing the KiDS CiC measurements with predictions from numerical simulations that incorporate a modified gravity model.

Finally, the methodology developed in this chapter can be extended to upcoming large-scale surveys such as Euclid and LSST. These surveys will provide significantly larger galaxy samples, improved photo-z estimates, and wider sky coverage, enabling much more precise measurements of higher-order clustering. With better control over observational systematics and the support of realistic simulations, CiC statistics have the potential to become a powerful complementary probe of non-linear structure formation, galaxy bias, and fundamental physics.

Part VI

Summary and future prospects

This thesis has addressed key challenges in precision cosmology in the era of large photometric surveys, where accurate three-dimensional information must be inferred from datasets with limited spectroscopic coverage. The work combines theoretical foundations in cosmology with modern machine learning techniques to improve photometric redshift (photo-z) estimation and to study the large-scale structure of the Universe.

The introductory chapters establish the theoretical and methodological framework that underpins the entire thesis. Chapter 1 presents the cosmological background, including the standard ΛCDM model, the theory of structure formation, and statistical measures of clustering such as the two-point correlation function and higher-order moments. These concepts provide the physical interpretation of the large-scale structure analyses carried out in later chapters. Chapter 2 introduces photometric redshift estimation methods, with particular emphasis on importance of photo-z estimation, machine learning approaches. It outlines the principles of supervised learning, neural networks, and convolutional architectures, which form the basis of the models developed in this work. Together, these chapters bridge the gap between cosmological theory and data-driven methodologies, setting the stage for the subsequent analysis.

The first main contribution of this thesis (Chapter 3) is the development and validation of the deep learning framework `Hybrid-z` for estimating photo-zs of galaxies and quasars in the Kilo-Degree Survey Data Release 4 (KiDS-DR4; [70]). The model combines 4-band image data ($ugri$) with 9-band photometric magnitudes ($ugriZYJHK_{\rm s}$) using a hybrid architecture in which Convolutional Neural Networks (CNNs) extract features from images, while Artificial Neural Networks (ANNs) process tabular data. An Inception-based CNN further enhances feature extraction. The `Hybrid-z` model has been released as an open-source tool for the community[1], enabling reproducible and extensible photometric redshift estimation. The resulting photo-z catalogue for the KiDS-DR4 bright sample is also publicly available[2].

[1] https://hybrid-z.readthedocs.io/en/latest/
[2] https://cdsarc.cds.unistra.fr/viz-bin/cat/J/A+A/698/A276

For galaxies, the incorporation of image data leads to a significant improvement in performance, reducing the photo-z scatter by approximately 20% compared to the previous method, `ANNz2`[104], based solely on photometric inputs. The model performs particularly well for blue galaxies than red, likely due to the presence of morphological features that can be effectively captured by CNNs.

A key methodological result of this work is the identification of the impact of training data distribution on model generalisation. Using spectroscopic redshifts from the GAMA survey as training data introduces imprints of its non-uniform redshift distribution into the predictions. To address this, we introduce a smoothing strategy that mitigates these effects. This improves generalisation to unseen data and represents an important contribution to data-driven cosmological inference.

The second part of this thesis (Chapter 4) focuses on photometric quasars in KiDS-DR4. We evaluate the performance of the `Hybrid-z` framework for point-like sources and find that, unlike for galaxies, the inclusion of image data does not significantly improve redshift estimation, reflecting the lack of resolved morphological features in quasars. Instead, the quality and representativeness of the training dataset play a dominant role. By training on spectroscopic samples from DESI-DR1 and SDSS-DR17, we achieve a high level of accuracy, with a scatter of $\sim 0.04(1+z)$ compared to that of the previous method [91].

Using these improved photo-z estimates, we construct a quasar catalogue of approximately $157k$ objects and perform a tomographic angular clustering analysis. The inferred quasar bias increases from $b \sim 1.6$ at $z \sim 0.6$ to $b \sim 4.0$ at $z \sim 2.2$, consistent with a quadratic redshift evolution. The results indicate that quasars reside in dark matter haloes of mass $\log_{10}(M_{\rm eff}/h^{-1}M_{\odot}) \sim 12.7$–$12.9$, and that they trace increasingly rare density peaks at higher redshift, as quantified by the evolution of the effective peak height $\nu_{\rm eff}$. These findings are consistent with the hierarchical growth of structure outlined in the introductory cosmology chapter. We also investigate systematic effects in clustering measurements and show that the assumed redshift distribution has a significant impact on bias estimation, highlighting the importance of photo-z calibration. In contrast, stellar contamination is found to have a negligible effect on the results.

The final Chapter 5 represents a natural continuation of the earlier works: while Chapter 3 focused on improving the redshift estimation of the KiDS-DR4 bright galaxy catalog, Chapter 4 explored the angular clustering and Gaussian information of underlying matter distribution, Chapter 5 uses the improved redshifts of KiDS-DR4 bright galaxies to investigate the higher-order, non-Gaussianity of matter distribution. Using the Count-in-Cells (CiC) formalism, we measured the averaged connected moments $\overline{\omega}_J(\theta)$ and reduced cumulants $S_J(\theta)$ for the KiDS-DR4 bright sample. The results show a strong dependence of higher-order clustering on angular scale, redshift, colour, and stellar mass. In particular, red and high stellar mass galaxies exhibit stronger clustering and higher non-Gaussianity

compared to blue and lower stellar mass galaxies, consistent with the picture in which they reside in more massive dark matter haloes.

The CiC measurements further reveal a transition toward Gaussian behaviour on large angular scales, where the second-order moment progressively dominates over higher-order contributions. In addition, a non-physical upturn is observed at scales $\theta \gtrsim 1^\circ$, highlighting the sensitivity of higher-order statistics to survey geometry and observational systematics. This chapter demonstrates that CiC statistics provide a powerful and complementary probe of the non-Gaussian structure of the galaxy distribution, and establish a foundation for future studies of galaxy bias, and tests of gravity using current and forthcoming wide-field surveys.

Overall, this thesis demonstrates that combining a solid theoretical foundation in cosmology with advanced machine learning techniques enables more accurate photo-z estimation and, consequently, measurements of large-scale structure clustering. They are directly applicable to current and upcoming wide-field surveys, where precise photo-zs and higher-order clustering measurements will be essential for advancing our understanding of cosmology and galaxy evolution.

Bibliography

[1] Ait Ouahmed R., Arnouts S., Pasquet J., Treyer M., Bertin E., 2024, A&A, 683, A26

[2] Alam S., et al., 2021, jcap, 2021, 050

[3] Alarcon A., et al., 2021, MNRAS, 501, 6103

[4] Arnouts S., Cristiani S., Moscardini L., Matarrese S., Lucchin F., Fontana A., Giallongo E., 1999, MNRAS, 310, 540

[5] Baugh C. M., et al., 2004, MNRAS, 351, L44

[6] Baum W. A., 1957, AJ, 62, 6

[7] Baum W. A., 1962, in McVittie G. C., ed., IAU Symposium Vol. 15, Problems of Extra-Galactic Research. p. 390

[8] Baumann D., 2022, Cosmology. Cambridge University Press, doi:10.1017/9781108937092

[9] Benítez N., 2000, ApJ, 536, 571

[10] Bernardeau F., Colombi S., Gaztanaga E., Scoccimarro R., 2002, physrep, 367, 1

[11] Beutler F., et al., 2013, MNRAS, 429, 3604

[12] Bilicki M., et al., 2018, A&A, 616, A69

[13] Bilicki M., et al., 2021, A&A, 653, A82

[14] Birrer S., et al., 2020, A&A, 643, A165

[15] Bishop C. M., 2007, Pattern Recognition and Machine Learning (Information Science and Statistics), 1 edn. Springer, http://www.amazon.com/Pattern-Recognition-Learning-Information-Statistics/dp/0387310738%3FSubscriptionId%3D13CT5CVB80YFWJEPWS02%26tag%3Dws%26linkCode%3Dxm2%26camp%3D2025%26creative%3D165953%26creativeASIN%3D0387310738

[16] Bolzonella M., Miralles J. M., Pelló R., 2000, A&A, 363, 476

[17] Bouchet F. R., Davis M., Strauss M., 1992, in Mamon G. A., Gerbal D., eds, Distribution of Matter in the Universe. pp 287–300

[18] Brammer G. B., van Dokkum P. G., Coppi P., 2008, ApJ, 686, 1503

[19] Bruzual G., Charlot S., 2003, MNRAS, 344, 1000

[20] Burger P., Friedrich O., Harnois-Déraps J., Schneider P., 2022, A&A, 661, A137

[21] Carlson J., White M., Padmanabhan N., 2009, prd, 80, 043531

[22] Chisari N. E., et al., 2019, apjs, 242, 2

[23] Coles P., Jones B., 1991, MNRAS, 248, 1

[24] Coles P., Lucchin F., 2002, Cosmology: The Origin and Evolution of Cosmic Structure, Second Edition

[25] Collister A. A., Lahav O., 2004, PASP, 116, 345

[26] Colombi S., Bouchet F. R., Schaeffer R., 1994, A&A, 281, 301

[27] Connolly A. J., Csabai I., Szalay A. S., Koo D. C., Kron R. G., Munn J. A., 1995, AJ, 110, 2655

[28] Croton D. J., et al., 2004, MNRAS, 352, 1232

[29] Croton D. J., Norberg P., Gaztañaga E., Baugh C. M., 2007, MNRAS, 379, 1562

[30] Csabai I., et al., 2003, AJ, 125, 580

[31] Cybenko G., 1989, Mathematics of control, signals and systems, 2, 303

[32] DESI Collaboration et al., 2025, arXiv e-prints, p. arXiv:2503.14745

[33] D'Isanto A., Polsterer K. L., 2018, A&A, 609, A111

[34] Dark Energy Survey Collaboration et al., 2016, MNRAS, 460, 1270

[35] Desjacques V., Jeong D., Schmidt F., 2018, physrep, 733, 1

[36] Dodelson S., 2003, Modern Cosmology

[37] Drozda P., Hellwing W. A., Bilicki M., 2022, prd, 106, 043513

[38] Drozda P., Hellwing W. A., Bilicki M., 2025, Phys. Rev. D, 112, 063546

[39] Dvornik A., et al., 2020, A&A, 642, A83

[40] Efstathiou G., Kaiser N., Saunders W., Lawrence A., Rowan-Robinson M., Ellis R. S., Frenk C. S., 1990, MNRAS, 247, 10P

[41] Einasto J., Suhhonenko I., Liivamägi L. J., Einasto M., 2018, A&A, 616, A141

[42] Einstein A., 1916, Annalen Phys., 49, 769

[43] Einstein A., 1936, Journal of the Franklin Institute, 221, 349

[44] Ellis G. F. R., Maartens R., MacCallum M. A. H., 2012, Relativistic Cosmology

[45] Freedman W. L., et al., 2019, ApJ, 882, 34

[46] Fukushima K., 1980, Biological Cybernetics, 36, 193

[47] Gaztanaga E., 1994, MNRAS, 268, 913

[48] Gaztanaga E., Bernardeau F., 1998, aap, 331, 829

[49] Geman S., Bienenstock E., Doursat R., 1992, Neural Computation, 4, 1

[50] Goodfellow I., Bengio Y., Courville A., 2016, Deep Learning. MIT Press

[51] Harnois-Déraps J., et al., 2024, MNRAS, 534, 3305

[52] He K., Zhang X., Ren S., Sun J., 2015, arXiv e-prints, p. arXiv:1502.01852

[53] He S., Yu J., Rocher A., Forero-Sánchez D., Kneib J.-P., Zhao C., Burtin E., Hou J., 2025, arXiv e-prints, p. arXiv:2508.21182

[54] Hellwing W. A., Juszkiewicz R., 2009, prd, 80, 083522

[55] Hellwing W. A., Juszkiewicz R., van de Weygaert R., 2010, prd, 82, 103536

[56] Hildebrandt H., et al., 2021, Astron. Astrophys., 647, A124

[57] Hornik K., 1991, Neural Networks, 4, 251

[58] Hoyle B., 2016, Astronomy and Computing, 16, 34

[59] Hubble E., 1929, Proceedings of the National Academy of Science, 15, 168

[60] Huber P. J., 1964, The Annals of Mathematical Statistics, 35, 73

[61] Ilbert O., et al., 2006, A&A, 457, 841

[62] Ishak M., 2019, Living Reviews in Relativity, 22, 1

[63] Jalan P., et al., 2024, Astron. Astrophys., 692, A177

[64] John William A., Jalan P., Bilicki M., Hellwing W. A., Thuruthipilly H., Nakoneczny S. J., 2025, A&A, 698, A276

[65] Johnston H., et al., 2021, A&A, 648, A98

[66] Juszkiewicz R., Bouchet F. R., Colombi S., 1993, ApJL, 412, L9

[67] Kaiser N., 1984, ApJL, 284, L9

[68] Kerr R. P., 1963, Phys. Rev. Lett., 11, 237

[69] Koo D. C., 1985, AJ, 90, 418

[70] Kuijken K., et al., 2019, A&A, 625, A2

[71] Kurki-Suonio H., 2024, Cosmological Perturbation Theory I, https://www.mv.helsinki.fi/home/hkurkisu/cpt/CosPer.pdf

[72] LSST Science Collaboration et al., 2009, arXiv e-prints, p. arXiv:0912.0201

[73] Lancaster T., Blundell S., 2025, General Relativity for the Gifted Amateur, doi:10.1093/oso/9780192867407.001.0001.

[74] Landy S. D., Szalay A. S., 1993, ApJ, 412, 64

[75] Laur J., et al., 2022, A&A, 668, A8

[76] LeCun Y., Bengio Y., Hinton G., 2015, nature, 521, 436

[77] Lecun Y., Bottou L., Bengio Y., Haffner P., 1998, Proceedings of the IEEE, 86, 2278

[78] Li R., et al., 2022, A&A, 666, A85

[79] LoVerde M., Afshordi N., 2008, prd, 78, 123506

[80] Loh E. D., Spillar E. J., 1986, ApJ, 303, 154

[81] Longair M. S., 2008, Galaxy Formation

[82] Ma C.-P., Bertschinger E., 1995, ApJ, 455, 7

[83] Maraston C., 2005, MNRAS, 362, 799

[84] Masci F., SWIRE Team 2006, in Armus L., Reach W. T., eds, Astronomical Society of the Pacific Conference Series Vol. 357, The Spitzer Space Telescope: New Views of the Cosmos. p. 271 (arXiv:astro-ph/0503170), doi:10.48550/arXiv.astro-ph/0503170

[85] Mcculloch W., Pitts W., 1943, Bulletin of Mathematical Biophysics, 5, 127

[86] Mecke K. R., Buchert T., Wagner H., 1994, A&A, 288, 697

[87] Merz G., et al., 2025, The Open Journal of Astrophysics, 8, 40

[88] Misner C. W., Thorne K. S., Wheeler J. A., 1973, Gravitation

[89] Mo H., van den Bosch F. C., White S., 2010, Galaxy Formation and Evolution, doi:10.1017/CBO9780511807244.

[90] Murphy K. P., 2012, Machine learning: a probabilistic perspective. MIT press

[91] Nakoneczny S. J., et al., 2021, A&A, 649, A81

[92] Newman J. A., Gruen D., 2022, ARA&A, 60, 363

[93] Norberg P., Gaztañaga E., Baugh C. M., Croton D. J., 2011, MNRAS, 418, 2435

[94] Padmanabhan N., et al., 2007, MNRAS, 378, 852

[95] Pasquet J., Bertin E., Treyer M., Arnouts S., Fouchez D., 2019, A&A, 621, A26

[96] Peacock J. A., 1999, Cosmological Physics

[97] Peebles P. J. E., 1980, The large-scale structure of the universe

[98] Planck Collaboration et al., 2020, A&A, 641, A9

[99] Planck Collaboration et al., 2021, A&A, 652, C4

[100] Prechelt L., 2012, Early Stopping — But When?. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 53–67, doi:10.1007/978-3-642-35289-8_5, https://doi.org/10.1007/978-3-642-35289-8_5

[101] Riess A. G., et al., 2022, ApJL, 934, L7

[102] Ryden B., 1970, Introduction to cosmology. Cambridge University Press, doi:10.1017/9781316651087

[103] SDSS Collaboration et al., 2025, arXiv e-prints, p. arXiv:2507.07093

[104] Sadeh I., Abdalla F. B., Lahav O., 2016, PASP, 128, 104502

[105] Salvador A. I., et al., 2019, MNRAS, 482, 1435

[106] Salvato M., Ilbert O., Hoyle B., 2019, Nature Astronomy, 3, 212

[107] Schwarzschild K., 1999, arXiv e-prints, p. physics/9905030

[108] Scoccimarro R., 1997, ApJ, 487, 1

[109] Shuntov M., et al., 2025, A&A, 704, A339

[110] Slipher V. M., 1913, Lowell Observatory Bulletin, 2, 56

[111] Smith R. E., et al., 2003, MNRAS, 341, 1311

[112] Soo J. Y. H., et al., 2017, Monthly Notices of the Royal Astronomical Society, 475, 3613

[113] Srivastava N., Hinton G. E., Krizhevsky A., Sutskever I., Salakhutdinov R., 2014, Journal of Machine Learning Research, 15, 1929

[114] Szapudi I., Szalay A. S., 1998, ApJL, 494, L41

[115] Szegedy C., et al., 2015, in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp 1–9, doi:10.1109/CVPR.2015.7298594

[116] Takahashi R., Sato M., Nishimichi T., Taruya A., Oguri M., 2012, ApJ, 761, 152

[117] Tibshirani R., 1996, Journal of the Royal Statistical Society Series B: Statistical Methodology, 58, 267

[118] Wald R. M., 1984, General Relativity. Chicago Univ. Pr., Chicago, USA, doi:10.7208/chicago/9780226870373.001.0001

[119] Wechsler R. H., Tinker J. L., 2018, ARA&A, 56, 435

[120] White S. D. M., 1979, MNRAS, 186, 145

[121] Woods D., Fahlman G. G., 1997, ApJ, 490, 11

[122] Wright A. H., et al., 2025, A&A, 703, A158

[123] Yan Z., et al., 2025, A&A, 694, A259

[124] Zehavi I., et al., 2002, ApJ, 571, 172

[125] Zehavi I., et al., 2011, apj, 736, 59

[126] Zeldovich I. B., Einasto J., Shandarin S. F., 1982, Nature, 300, 407

[127] de Jong J. T. A., Verdoes Kleijn G. A., Kuijken K. H., Valentijn E. A., 2013, Experimental Astronomy, 35, 25

[128] van Uitert E., et al., 2016, MNRAS, 459, 3251

Appendix A

Scalar Perturbation

The Einstein equations for the background universe are,

$$G^{\mu}{}_{\nu}[\bar{g}_{\mu\nu}] = 8\pi G \overline{T}^{\mu}_{\nu}$$

$\overline{T}^{\mu}_{\nu}$ is the energy-momentum tensor and $\bar{g}_{\mu\nu}$ is the unperturbed FLRW metric in the background universe. Assume that the spacetime deviates only slightly from an idealized, perfectly homogeneous and isotropic spacetime, which is defined as the background. We introduce the perturbation, and the full metric is decoupled into background and perturbation ($\delta g_{\mu\nu}$) parts. The total perturbed metric is,

$$g_{\mu\nu} = \bar{g}_{\mu\nu} + \delta g_{\mu\nu}$$

The perturbed energy-momentum tensor is

$$T^{\mu}_{\nu} = \overline{T}^{\mu}_{\nu} + \delta T^{\mu}_{\nu}$$

The functional $G^{\mu}{}_{\nu}[g_{\mu\nu}]$ can be Taylor expanded up to linear order,

$$G^{\mu}{}_{\nu}[g_{\mu\nu}] = G^{\mu}{}_{\nu}[\bar{g}_{\mu\nu}] + \delta G^{\mu}{}_{\nu}[\delta g_{\mu\nu}],$$

Inserting $T^{\mu}{}_{\nu} = \bar{T}^{\mu}{}_{\nu} + \delta T^{\mu}{}_{\nu}$ into the Einstein equation $G^{\mu}{}_{\nu}[g_{\mu\nu}] = 8\pi G\, T^{\mu}{}_{\nu}$ and subtracting the background equation $G^{\mu}{}_{\nu}[\bar{g}_{\mu\nu}] = 8\pi G\, \bar{T}^{\mu}{}_{\nu}$ yields the linearised Einstein equations:

$$\delta G^{\mu}{}_{\nu} = 8\pi G \delta T^{\mu}_{\nu} \tag{A.1}$$

We can decompose the metric and stress-energy tensor perturbation by using the scalar-vector-tensor (SVT) decomposition theorem [71]. In linear order, the three sectors decouple, and for the study of density perturbations, only the scalar part is relevant. The source terms for these equations arise from the stress-energy tensor, which in the background takes the perfect fluid form specified by the total energy density and pressure

summed over all species. Perturbations of this tensor introduce fluctuations in the density, divergence of fluid velocity (θ), pressure, and anisotropic stress (σ), with the precise contribution depending on the physical nature of each component: cold dark matter and baryons contribute mainly through density and velocity perturbations, while relativistic species such as photons and neutrinos also provide significant pressure perturbations and anisotropic stress.

In synchronous gauge and in Fourier space ($\mathbf{k} = k\hat{k}$, and k is the wavenumber of Fourier mode), the scalar perturbations of the spatial metric are described by the two fields $h(\mathbf{k}, \tau)$ and $\eta(\mathbf{k}, \tau)$. Here, τ is the conformal time. In terms of h and η, the linearized Einstein equations in $\mathbf{k}$-space yield four relations corresponding to the following. The time-time component gives the Poisson constraint, relating the metric potentials to the perturbation of the energy density. The time-space component provides the momentum constraint, linking the time derivative of η to the velocity divergence of the fluid. The trace of the spatial components produces the dynamical equation for h, sourced by the isotropic pressure perturbation, while the traceless spatial part supplies the evolution equation for the combination of h and η, sourced by the anisotropic stress. scalar perturbation equations of the linearized Einstein equations are,

$$k^2\eta - \tfrac{1}{2}\frac{\dot{a}}{a}\dot{h} = 4\pi G a^2\, \delta T^0_0, \qquad \text{(00, Poisson constraint)}$$

$$k^2\dot{\eta} = 4\pi G a^2\, (\bar{\rho} + \bar{P})\, \theta, \qquad \text{(0i, momentum constraint)}$$

$$\ddot{h} + 2\frac{\dot{a}}{a}\dot{h} - 2k^2\eta = -8\pi G a^2\, \delta T^i_i, \qquad \text{(trace } ij\text{)}$$

$$\ddot{h} + 6\ddot{\eta} + 2\frac{\dot{a}}{a}(\dot{h} + 6\dot{\eta}) - 2k^2\eta = -24\pi G a^2\, (\bar{\rho} + \bar{P})\, \sigma. \qquad \text{(traceless } ij\text{)}$$

The conventions used in this section follow those of [82]. The Newtonian limit of these equations is discussed in Section.1.4.1.

Appendix B

Striding, Padding, and Pooling operations

In standard convolution, filters slide over the input with unit step size (stride = 1). CNNs may employ larger strides ($s > 1$) to downsample feature maps by skipping intermediate spatial positions. For stride s, the convolution output becomes:

$$\mathbf{F}_{ij} = \sum_{m=0}^{k_h-1} \sum_{n=0}^{k_w-1} \mathbf{W}_{mn} * \mathbf{X}_{i\cdot s+m,\ j\cdot s+n} + b \tag{B.1}$$

Increasing stride reduces feature map resolution, thereby decreasing computational cost and providing a controlled level of translational abstraction. $\mathbf{F}_{ij}$, is then passed through a non-linear activation function to introduce non-linearity into the model [50]. Padding is a technique used in CNNs to control the spatial dimensions of the output feature map. Without padding, the output size of a convolution decreases after each layer because the filter cannot extend beyond the boundaries of the input. By adding extra rows and columns (usually zeros) around the input, padding allows the filter to slide over edge elements more effectively [1].

Pooling layers further reduce spatial dimensionality while retaining salient information. The most common variant, max pooling, selects the maximum activation within a local spatial region:

$$\mathbf{P}_{ij} = \max_{(m,n)\in\Omega} \mathbf{F}_{i+m,\ j+n} \tag{B.2}$$

where Ω defines the pooling window (e.g., 2×2). Pooling introduces robustness to small translations and distortions while preventing overfitting by reducing representational capacity. Average pooling is an alternative that computes the mean over the region. Astronomical galaxy images often contain noise, and some pixels may have unusually high values compared to their neighbors. If max pooling is applied in such regions, the output will capture these high values, effectively amplifying the noise rather than representing the true structure of the galaxy. In contrast, average pooling computes the mean over the

[1] For the visualization of convolution operation visit this page https://github.com/vdumoulin/conv_arithmetic.

region, which smooths out such outliers and preserves the overall intensity and morphology.