UNIVERSITY OF GLASGOW

The Application of Supervised Machine Learning Techniques to Determine Photometric Redshifts

by

Lara Janiurek

A thesis submitted in fulfillment for the
degree of Masters by Research

in the

Institute for Gravitational Research
School of Physics and Astronomy

January 2023

——-

Declaration of Authorship

I declare that, except where explicit reference is made to the contribution of others, that this dissertation is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.

Signed: Lara Janiurek

Date: 29/08/2022

UNIVERSITY OF GLASGOW

Abstract

Institute for Gravitational Research
School of Physics and Astronomy

Masters by Research

by Lara Janiurek

The inference of the Hubble constant using gravitational wave data has allowed for a new way for the expansion of the Universe to be probed. The use of dark sirens, which are mergers of binary black hole systems, to measure the Hubble constant (H_0) may shed considerable light on the current Hubble tension. Galaxy redshift surveys are a key ingredient for the application of these dark sirens in the measurement of H_0. Most binary black hole merger events are not expected to have an associated electromagnetic counterpart, therefore measuring H_0 using these sirens requires the identification of the redshifts of potential host galaxies and marginalising over these host galaxy redshifts. Photometric redshift surveys often contain significant statistical or systematic errors which may impact adversely on the Hubble constant inference. Improving the performance of dark sirens in the future observing runs of the LIGO Virgo KAGRA(LVK) network requires a better understanding of the photometric redshift errors. The current redshift values used by the LVK for cosmological inference are assumed to have an associated Gaussian error, however a true quantification of the redshift posteriors would give a more accurate result in the overall inference of the H_0. Spectroscopic redshifts are difficult to obtain and many physical photometric techniques rely on cosmological models that could potentially introduce bias into the redshift measurements. Machine learning techniques are advantageous in that they don't rely on assumed cosmological models.

In this work, the random forest algorithm GALPRO is implemented to generate photometric redshift posteriors. It is initially calibrated using a truth dataset compiled by Zhou et al. The initial calibration is successful and analysis suggests that the redshift posterior distributions are largely non-Gaussian. This further reinforces the need for a reliable method to generate redshift posteriors to better represent these photometric errors in the inference of H_0.

Tests were run using the Zhou et al. dataset to determine how statistically similar the training and testing datasets from a survey must be for GALPRO to be applicable. It was found that the training and testing datasets must have similar redshift distributions and overlap by at least 90% in the band ranges to give accurate results. GALPRO was then trained using the Zhou et al. dataset and applied to a sample from the PanSTARRS survey to explore if GALPRO could be trained using a trusted dataset and applied to a general, new survey. It was shown that no matter how statistically equivalent the two surveys were, GALPRO could not produce accurate redshift posteriors for the new survey. The Zhou et al. and PanSTARRS surveys had very similar redshift distributions and overlapped in each inputted band by over 90%. Despite this, application of the algorithm still resulted in a catastrophic failure, indicating that there must be some underlying fundamental difference between the two surveys that causes the program to fail. This work serves as a cautionary tale in the application of random forests to new surveys when generating photometric redshift posteriors.

Acknowledgements

I would like to acknowledge and give my warmest thanks to Professor Martin Hendry, who made this work possible. His guidance and advice carried me through this project and I am very grateful. I could not have asked for a better supervisor. I would also like to thank Federico Stachurski for replying to my incessant emails and helping me with my endless computational issues, thank you so much!

I would like to thank my mum for all the support she has given me along the way. I want to express my thanks and deep gratitude to my dad, who read my entire thesis and sent me notes without getting any roid rage! You have both always encouraged and supported me in my academic journey and I will always be grateful for this. I would like to thank my grandma for giving me wise words of encouragement throughout this process and inspiring me to work hard to make her proud. I also want to send love to my granddad, who passed away midway through my Masters degree. He always wanted to be a physicist and was the one person in my family who understood what I was studying. I love you granddad and I will always work harder to make you proud since I know how much you wanted me to do well and become a scientist!

Thank you to the Kelvin building gang for making this degree fun and supporting me everyday whilst we study together. Finally, I'd like to thank Andrew White for his support and encouragement and bringing me a lovely lunch everyday. Without any of these people, I would not have produced this work or had as much fun during the process! Thank you to everyone involved, I will always be grateful and appreciative of your role in this year!

...

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Theoretical Foundations

1.1 The Expansion Of the Universe

It has been agreed upon for nearly a century that the Universe around us is expanding, however the quest to measure the rate of such expansion is still ongoing. The first inference that the universe may be expanding came in 1912, from Vesto Slipher's observations that the light emitted from distant galaxies was redshifted, and that in fact all of his observed spiral nebulae appeared to be receding from Earth [7]. This redshift was later explained by the fractional increase in the wavelength of emitted light from an object, due to the expansion of space through which the light is travelling [8].

In the early 1920s, Alexander Friedmann developed a theoretical prediction of the universe's expansion using Einstein field equations. The first observational evidence of expansion was published in 1924 by Kurt Lundmark as he worked on extragalactic distance measurements. Independently, in 1927, Georges Lemaître derived a theoretically similar proof to Friedmann, alongside providing observational evidence indicating a linear relationship between the distance to a galaxy and its recessional velocity [9, 10].

In 1929, these finding were confirmed by the observations of Edwin Hubble. Hubble focused on the measurement of distances out to extragalactic nebulae, which lead him to examine the relationship between the distances of the nebulae and their radial velocities.

Although his distance measurements were considerably underestimated, he still managed to determine a roughly linear relationship between the estimated distance out to a galaxy and its redshift. This relationship was later known as Hubble's law, which states that the larger the distance between a galaxy and the observer, the greater the recessional velocity of that galaxy. Hubble determined the distance and redshift values for 46 galaxies, leading him to deduce a value of 500 km s^{-1} Mpc^{-1} for the Hubble constant (H_0). This is much greater than the currently accepted value, due to the distance measurements used being considerably underestimated [11].

The scientific explanation for this observed redshift was yet to be clarified as many physicists, including Hubble himself, had dismissed the work of Lemaître. However it was later recognised that this redshift corresponds to the rate at which the universe itself is expanding. These early, inaccurate measurements of H_0 paved the way for the accurate results we see today, where the associated uncertainty has been reduced to a few percent [12, 13]. This has led to a fresh but challenging new problem, as different values of the Hubble constant indicate inconsistent results. Early universe measurements tend to infer lower values of H_0 whereas late-time, local measurements favour higher H_0 values [14, 15]. The explanation for this tension is not yet certain, with proposals generally falling into two groups, the first being that current estimations contain some underlying systematic error. If the potential error is corrected for, current results may align producing a single value of H_0. The other possibility is that our current cosmological model of the universe is incorrect and the evolution of the universe with time is not correctly described. This would imply that new physics is required to accurately describe our universe.

A measurement of H_0 calls for two variables: the distance out to an object which is caught in the Hubble flow, meaning that is recessional velocity is dominated by the expansion of the universe, and a measurement of said velocity. For cosmologically small distances, the redshift of the object is proportional to its recessional velocity. Determining accurate redshifts values for more distant galaxies poses a large problem when measuring H_0, which is discussed in great detail in later chapters. Obtaining accurate distance measures also presents many challenges as determining cosmological distances at varying scales calls for numerous methods of measurement. The cosmic distance ladder, seen in Figure 1.1, groups these various methods to measure cosmological distances at different scales. Standard candles, which are astronomical objects with a known

brightness, are a common tool used in distance measurements as their luminosity distance can be assessed using the object's observed and intrinsic brightness. However, the standard candle method does introduce some issues as it is difficult to know the object's intrinsic brightness with certainty. This brightness may potentially change, for example with redshift, which would therefore introduce an inherent bias to the measurement. Also, astronomical objects may seem fainter than they truly are, due to reasons such as interstellar gas obstructing them, meaning any decrease in luminosity may not be solely due to distance [16].

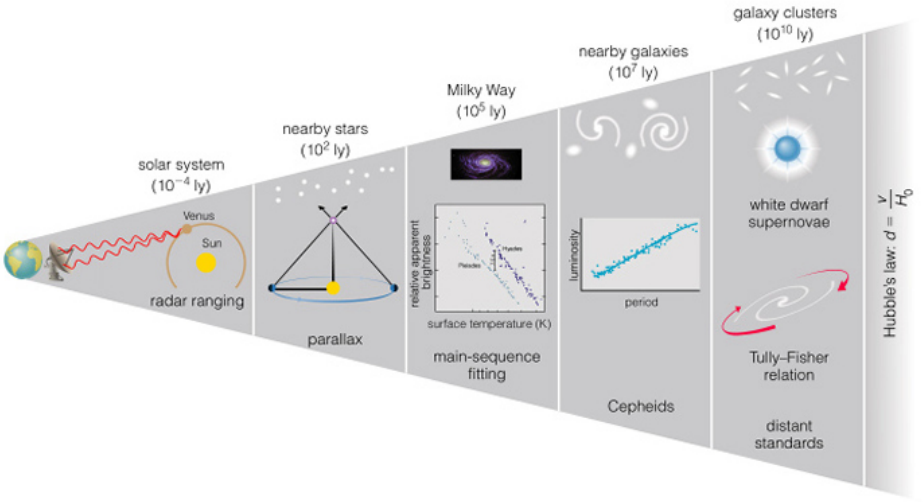


FIGURE 1.1: A diagram of the cosmic distance ladder. Figure credit: [1]

Gravitational waves, however, bypass many of these distant measurement issues due to their intrinsic properties. These waves form during extreme astronomical events, which cause ripples in the fabric of space-time. Examples of these events are compact binary coalescences (CBC), including the merging of black holes or neutron stars. The GWs formed by CBCs have a remarkable intrinsic property, as the luminosity distance out to a GW event is inversely proportional to the amplitude of the wave as it is measured on Earth. This is favourable, as no other form of calibration is required when taking distance measurements using GW data, although the measurement does still contain some uncertainty [17]. The GW sources are equivalent to standard candles in terms of an indicator of distance, and are therefore referred to as standard sirens. This property makes standard sirens an excellent tool in the measurement of H_0 .

The first ever GW signal to be detected on Earth was observed by Advanced LIGO in 2015 [18]. But it wasn't until 2017 when an extremely loud GW signal was detected with

an electromagnetic (EM) counterpart, leading to the host galaxy being identified and a direct measurement of H_0 [18]. This measurement of H_0 using the GW170817 BNS event gave an uncertainty of around 15% which is far too inaccurate to determine which previous H_0 measurements it may agree with. In order for the required accuracy to be achieved, we would need to observe many more similar events. However, the GW170817 was abnormally close by and well-localised, with the EM counterpart being identified extremely quickly. The probability of detecting multiple of these nearby events with EM counterparts is very low, and since then Virgo and Advanced LIGO have not observed another event which was similarly loud with an EM counterpar [19].

To tackle this issue, the code `gwcosmo` was developed by Rachel Gray in 2020, which is a software package used to estimate the Hubble parameter using gravitational wave observations. `gwcosmo` takes the theory, first proposed by Shutz in 1986, which estimates H_0 using GW detections with no EM counterpart, and implements it using a Bayesian framework that accounts for incomplete galaxy catalogues and selection effects. Typically, the EM counterpart of a GW signal would be used to determine the redshift of the host galaxy, however `gwcosmo` allows for the redshift to be provided by galaxy catalogues and the uncertainty relating to the true identity of the host galaxy can be marginalised over. Many mock data analyses, alongside real measurements using Advanced LIGO and Virgo data from the first three observing runs have been successful in combining multiple GW detections to measure H_0 [16].

With the upcoming fourth Advanced LIGO, Virgo and KAGRA observing run on the horizon, there is hope of many more GW event detections and therefore further constraining measurements of H_0. However, there is a lack of accurate redshifts describing the host galaxies of the events, yet this redshift value is key in determining the Hubble constant. Spectroscopic redshifts are difficult to obtain due to sparse spectral coverage and limited signal-to-noise ratios, and many physical photometric techniques rely on cosmological models which introduce inherent bias into the measurements. The current redshift values used by Advanced LIGO for cosmological inference are assumed to have an associated error which is Gaussian. If this assumption is incorrect, then the redshift errors must be quantified to lead to better constraints on the H_0 measurement. A true quantification of the redshift posteriors would give a more accurate result in the overall inference of H_0.

This thesis applies the software package GALPRO, developed by Sunil Mucesh in 2020 to generate photometric redshifts using machine learning techniques. The main goal of this thesis is to determine whether GALPRO is applicable to new surveys containing photometry data to generate reliable and accurate photometric redshifts for galaxies where no spectroscopic data is available. These redshifts may then be used for the quest to constrain H_0 using GW events. Redshifts generated by machine learning do not rely on cosmological models and therefore avoid any bias or artificial structures introduced by these models. GALPRO is made up of a random forest algorithm which can be trained using known photometry and spectroscopic redshifts to learn the mapping between the two. Once the mapping is learnt, it can then be applied to galaxy surveys where there is a lack of spectroscopic data to generate photometric redshifts. This thesis explores how reliable the photometric redshift estimates may be when applying the GALPRO software, trained on a specific dataset. By applying GALPRO to a new, different survey which is separate to the one it was trained on, the accuracy of the results can be determined, alongside any restrictions or limits on the software's capability. The photometric redshift posteriors are also assessed to determine whether the assumption that the redshift errors can be modelled as Gaussian is valid. The ability to generate reliable redshift estimates for surveys that lack spectroscopic data plays a key role in further constraining H_0 measurements, and this work assesses whether GALPRO may be used to compute these redshifts.

The first chapter details the context of this work, while the second chapter introduces the GALPRO software package and details its calibration and validation using known data samples. The third chapter explores the application of GALPRO to an unknown survey which only provides photometry data, and assesses its performance when encountering new surveys which may contain galaxies with different properties than those it was trained on. The goal of this work is to evaluate whether GALPRO can be reliably applied to new surveys to generate photometric redshifts, which may then populate the galaxy catalogues utilised by `gwcosmo` in the inference of H_0.

1.1.1 The Standard Model of Cosmology

The standard model of cosmology follows the cosmological principle, which is the notion that the universe is isotropic and homogeneous on a large scale. This implies that the

density of matter through the universe is around the same no matter where you are, and that it looks the same in all directions. The observation that all galaxies seem to be receding away from Earth is compatible with the cosmological principle if indeed the universe is expanding, as it means that no matter where you stand in the universe you would observe the same thing.

Hubble and Lemaître both independently came to the same conclusion, that a galaxy's recessional velocity from Earth is approximately proportional to the distance out to that galaxy. This lead to the Hubble-Lemaitre law [10, 11], which is written as:

$$v = H_0 d \tag{1.1}$$

with d being the proper distance from the observer to the object (Mpc). v is the recessional velocity of the object (km s^{-1}). H_0 is a constant of proportionality, named the Hubble constant, with units km s^{-1} Mpc^{-1}.

Obviously, the expansion of the universe implies that the universe was in fact previously much smaller than we see today, which is in agreement with the Big Bang theory. The standard model of cosmology, the Λ-cold-dark-matter (ΛCDM) model, starts with the Big Bang which happened around 13.8 billion years ago [20]. This was the birth of the universe, where a point of singularity began an explosive expansion leading to the universe being filled with cold dark matter and ordinary matter. As the universe initially began to expand, it filled with photons and matter in a dense hot plasma [16]. A slight variation in the density of this plasma was caused by quantum fluctuations which were amplified by Baryonic Acoustic Oscillations (BAO), similar to sound waves, as small perturbations propagated through the plasma [21]. The universe began to cool as it expanded, causing atoms to form. This meant that photons could now travel through space and are no longer absorbed and re-emitted. This period in which light and matter decoupled is known as recombination, and the photons at this time can still be observed today as the Cosmic Microwave Background (CMB). The perturbations in the plasma ceased to oscillate at this time, but still etched themselves on the distribution of matter, which can be seen today as small variations in the CMB temperature. As the universe further expands, the areas of over-density are contracted by gravity and increase in temperature, inducing the formation of stars and galaxies. This then leads

to the formation of the large scale structure of the universe, as galaxy clusters form and are joined by filaments. This method of formation means that the imprint of the BAO is still visible today through the web of matter in our universe [16].

The term *cold* dark matter refers to matter which is non-relativistic and does not interact with photons. Gravity should, and would, slow or halt the expansion of the universe if it weren't for the cosmological constant, Λ, which specifies the presence of dark energy in the universe [21]. This dark energy dominates as time progresses, and fights against the force of gravity thus accelerating the expansion of the universe. It is important to note the expansion of the universe does not cause the separation of gravitionally bound objects, but expands the empty space around them [20].

Discussion of the expansion of the universe often calls for the introduction of a *metric*. A metric is a mathematical description of the separation of events in space-time. The Friedmann-Lemaitre-Robertson-Walker (FLRW) metric is the most appropriate in this case as it is applicable to a universe which obeys the cosmological principle. The Einstein field equations of general relativity detail how the presence of matter affects the curvature of space-time, and the exact solution to these equations is given by the FLRW [9, 10, 22–27]. An important take away from the FLRW is that it includes a term which describes how the spatial distance between two events may evolve with time. The metric may be written as:

$$ds^2 = -c^2dt^2 + a^2(t)ds_3^2 \tag{1.2}$$

with ds^2 being the space-time interval between two events, $-c^2dt2$ represents the time dependent part of the separation, and $a(t)^2$ is a scale factor. $a^2(t)ds_3^2$ is the spatial part of separation, which is dependent on the geometry of the universe and whether it is assumed to be spherical, hyperbolic or flat.

An expanding universe may be described using comoving coordinates, bearing in mind that every single point in the universe is moving away from each other. For now, we note that we are not taking gravitionally bound objects into account. The comoving coordinates expand as the universe expands, and a scale factor $a(t)$ is used to describe how the comoving coordinates translate to proper distance (the distance between two

objects at a fixed time). The scale factor we see today, $a(t_0)$, is defined as 1. Generally speaking, we can say:

$$d(t) = a(t)d_0 \tag{1.3}$$

with $d(t)$ being the proper distance of two separated objects and d_0 being the distance at time t_0.

Equation 1.3 can be easily rearranged to give the Hubble Lemaitre Law. Firstly, the time derivative of each side is taken

$$\dot{d}(t) = \dot{a}(t)d_0 \tag{1.4}$$

It can be noted that Equation 1.3 can be rearranged as $d_0 = d(t)/a(t)$. This is substituted into Equation 1.4, giving

$$\dot{d}(t) = \frac{\dot{a}(t)}{a(t)}d(t) \tag{1.5}$$

This now gives the Hubble-Lemaitre law, as $\dot{d}(t)$ is the recessional velocity of an object with respect to the observer, $d(t)$ is the proper distance to object and $\dot{a}(t)/a(t)$ represents the Hubble parameter [16].

This is useful, but the FLRW is not able to describe how matter and energy may be affected by the universe. However, the FLRW can be assumed while solving the Einstein field equations, which produces the Friedmann equations [9, 22]. These equations do describe how the scale factor $a(t)$ may be affected by the geometry, pressure and density of the universe. The Friedmann equations are written as

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{kc^2}{a^2} + \frac{\Lambda c^2}{3} \tag{1.6}$$

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\rho + \frac{3p}{c^2}\right) + \frac{\Lambda c^2}{3} \tag{1.7}$$

Here, a is the scale factor as we use the shorthand notation for ease. $\dot{a}$ is therefore the time derivative of a, ρ represents the density of universe with p being the pressure, k the curvature of the universe and Λ is the cosmological constant.

All three of the parameters, a, p and ρ evolve with time, while π, c, G and Λ remain constant. The parameter k represents the curvature of the universe, which is indeed a continuum.

There are two common choices for the values of a and k. Firstly, the value of k may be equal to 1, 0 or -1, depending on whether the universe has spherical, flat or hyperbolic geometry respectively. The scale factor a is therefore rescaled to account for this. Secondly, the more modern definition defines $a = 1$ and a positive value of k gives a hyperspherical universe. If k is negative, then the universe is hyperbolic and $k = 0$ gives a flat universe 3.1. These two definitions both describe the same physics, but the first definition, where a is rescaled to account for the geometry of the universe, is used in this derivation.

The Hubble parameter, H (or $H(t)$) by definition can be given as

$$H \equiv \left(\frac{\dot{a}}{a}\right) \tag{1.8}$$

and the value of H at the current day is the Hubble constant, H_0. To determine the relationship between the Hubble constant, redshift and luminosity distance, one must first introduce the density parameter, Ω_m. This parameter describes the density of the universe with respect to the density if the universe were flat and therefore having no cosmological constant Λ, called the critical density. This can be defined by solving Equation 1.6 with k and $\Lambda = 0$:

$$\Omega_m = \frac{\rho}{\rho_c} \tag{1.9}$$

where

$$\rho_c \equiv \frac{3H^2}{8\pi G} \tag{1.10}$$

Coming back to Freidmann's equation, Equation 1.6 and rearranging:

$$\frac{8\pi G}{3H^2}\rho - \frac{kc^2}{a^2H^2} + \frac{\Lambda c^2}{3H^2} = 1 \tag{1.11}$$

The curvature and dark energy density parameters may then be defined as

$$\Omega_k = -\frac{kc^2}{a^2H^2} \quad \text{and} \quad \Omega_\Lambda = \frac{\Lambda c^2}{3H^2} \tag{1.12}$$

Substituting Ω_m and Equations 1.12 in Equation 1.11, we find

$$\Omega_m + \Omega_k + \Omega_\Lambda = 1 \tag{1.13}$$

Since the Hubble parameter varies with time, the density parameters containing H are also time dependant. The present day parameter values are defined as:

$$\Omega_{k,0} = -\frac{kc^2}{a_0^2H_0^2} = -\frac{kc^2}{H_0^2}, \quad \Omega_{\Lambda,0} = \frac{\Lambda c^2}{3H_0^2} \quad \text{and} \quad \Omega_{m,0} = -\frac{\rho_0}{\rho_{c,0}} = \frac{8\pi G}{3H_0^2}\rho_0 \tag{1.14}$$

If we now return to Equation 1.6, dividing through by H_0^2 and using present day parameters leads to

$$\frac{H^2}{H_0^2} = \frac{\rho}{\rho_0}\Omega_{m,0} + \frac{1}{a^2}\Omega_{k,0} + \Omega_{\Lambda,0} \tag{1.15}$$

Since the goal is to define the relationship between H_0, distance and redshift, it is now time to introduce cosmological redshift into the (literal and metaphorical) equation. The cosmological redshift, z, depends only on the scale factor, following:

$$1 + z = \frac{a(t_0)}{a(t_e)} \tag{1.16}$$

with $a(t_0)$ being the size of the universe at the time the light from the object was observed and $a(t_e)$ being the size of the universe when light was emitted. This can be easily converted to a present-day time scale as $a(t) = 1/(1+z)$.

We can now use the present day relationship between the scale factor and redshift alongside the present day density of the universe with respect to the density at time t, $\rho/\rho_0 = 1/a(t)^3$, in Equation 1.15 to give:

$$E(z) = \frac{H(z)}{H_0} = \sqrt{\Omega_{m,0}(1+z)^3 + \Omega_{k,0}(1+z)^2 + \Omega_{\Lambda,0}} \tag{1.17}$$

Where $E(z)$ is the dimensionless Hubble parameter. Currently, measurements give $\Omega_{m,0} \approx 0.3$, $\Omega_{k,0} \approx 0$ and $\Omega_{\Lambda,0} \approx 0.7$, which can be substituted into the above equation [12]. It is broadly assumed the universe is close to or completely flat, meaning $k = 0$, which also simplifies Equation 1.17.

Now, attention turns to the luminosity distance d_L, and its relation to H_0 and z. This is the distance out to an object if the inverse square law is retained across the entire universe, which of course is not true. If the universe is assumed to be flat, then the luminosity distance is given in terms of the redshifted comoving distance, $(1 + z)D_c$. From this, the relationship between d_L, z and H_0 can be written as:

$$d_L = \frac{c(1+z)}{H_0} \int_0^z \frac{dz'}{E(z')} \tag{1.18}$$

The comoving volume V_c is also useful to define at this point. As the universe expands, it is assumed that isotropy and homogeneity still stands and therefore matter has a uniform comoving volume distribution. This however, isn't entirely true due to galaxy clustering but still holds when viewed on a large enough scale. How the comoving volume changes with redshift is given by:

$$\frac{dV_c}{dz} = \frac{c^3}{H_0^3} \frac{1}{E(z)} \Big(\int_0^z \frac{dz'}{E(z')} \Big)^2 \tag{1.19}$$

The above equation would hold even if the distribution was not comoving. The dependence on H_0 can be dropped when normalised [16].

Focus is shifted back to the redshift of galaxies, z, which is given by:

$$z = \frac{\lambda_{obs} - \lambda_{emit}}{\lambda_{emit}} \tag{1.20}$$

with λ_{emit} being the redshift of light when emitted from object and λ_{obs} as the observed redshift.

It can be assumed that the galaxy's recessional velocity, v, is much much less than c and we can approximate:

$$z \approx \frac{v}{c} \tag{1.21}$$

This shows that for very small redshift values, v is directly proportional to z. At these small redshift values it can also be assumed that the proper distance and luminosity distance are of a very similar length, and so this finally leads to the Hubble relation:

$$cz \approx H_0 d_L \tag{1.22}$$

It is important to remember here that z represents cosmological redshift and does not account for peculiar motion. Any motion of the object which is not directly due to the Hubble flow can often impact the measurement of redshift, particularly at low redshifts where peculiar motion dominates.

Local H_0 measurements are often made using Equation 1.22, as d_L may be reliably measured using standard candles and z can be determined via spectroscopic or photometric techniques. This method is advantageous in that it requires no further assumptions or parameters, and was used in the first measurements of H_0 by Hubble and Lemaitre. Although these initial measurements contained severe measurement uncertainties due to greatly inaccurate distance measurements, they paved the way for current value measurements of $H_0 \approx 70$ km s^{-1} Mpc^{-1} [28]. However, there is still much tension arising from today's H_0 measurements, which will be discussed in detail in the next section.

1.1.2 The Hubble Tension

Different types of observation have given rise to a tension in the measured values of H_0. Although the ΛCDM model encapsulates current observations well with regards to our universe, early and late-time measurements of H_0 result in differing H_0 values. Early-time measurements refer to measurements made at high redshifts, which looks back into

the early universe and propagates forward to determine the current H_0 value, whereas late-time measurements make use of much lower redshifts to measure the current day H_0 value [14, 15]. Figure 1.2 demonstrates this tension in a graphical manner.

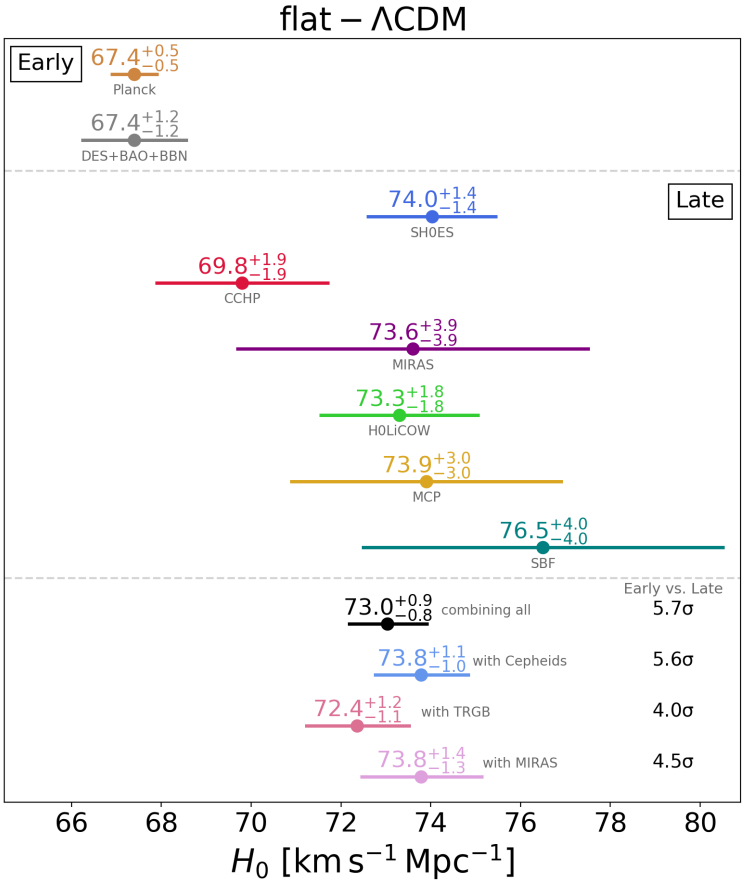


FIGURE 1.2: A graphical display of the tension between early and late time measurements of the Hubble constant. Figure credit: [2]

The value of H_0 can be determined using early-time measurements via the cosmic microwave background (CMB). The universe is not completely isotropic and homogeneous, therefore the CMB is not completely uniform, but contains minor density fluctuations as a result of early quantum fluctuations amplified by the BAO. When light and matter decoupled at the epoch of recombination, photons could freely travel through the universe and fluctuations in the density of the plasma of the early universe was imprinted on the structure of the universe, which birthed the large-scale structure of the universe that we currently see.

The temperature fluctuations in the CMB, caused by the presence of quantum fluctuations as the universe formed, have an angular dependence that can be used to determine

the value of H_0. Anisotropies in the CMB have been detected in measurements by both the Planck collaboration and the Atacama Cosmology Telescope + Wilkinson Microwave Anisotropy Probe, which gave values of $H_0 = 67.36 \pm 0.54$ km s^{-1} Mpc^{-1} and $H_0 = 67.6 \pm 1.1$ km s^{-1} Mpc^{-1} respectively [12, 29].

The sound horizon, r_s, is defined as the distance that sound waves would have travelled in the period of time before recombination. Both CMB anisotropies and theory produce agreeing values of r_s. From the sound horizon, a length scale may be determined and this length scale will change with time as the universe expands. Measuring the variance of this length scale allows for a determination of the rate of expansion. Studying the clustering of galaxies leads to low redshift BAO measurements, which allow for a joint constraint on r_s and H_0, however CMB or Big Bang Nucleosynethsis (BBN) measurements are required to break degeneracy. BBN and CMB measurements of H_0 are independent but both results agree nicely with the Planck value and others [30, 31].

Type 1a supernovae can be used to produce late-time H_0 measurements as their intrinsic luminosities can be found from their light curves, meaning they act as standard candles. Cepheid variable stars of known distances may be used to calibrate type 1a supernovae by measuring the distance between type 1as and Cepheid variables using parallax estimates and identifying the two object's host galaxy. The Supernovae H_0 Equation of State of Dark energy (SH0ES) collaboration produced a measurement of H_0 in 2019 by surveying 70 cepheids contained in the host galaxy of Type 1a supernovae. This gave a value of $H_0 = 74.03 \pm 1.42$ km s^{-1} Mpc^{-1} [32]. After recalibrating the distances out to the cephied, an improved measurement of $H_0 = 74.2 \pm 1.3$ km s^{-1} Mpc^{-1} was published. The Planck 2018 result, which assumes ΛCDM, differs from this late-time measurement by a value of 4.2σ, demonstrating the tension between early and late-time measurements.

There have been many more late-time measurements such as the Maser Cosmology Project which used distances out to megamaser-hosting galaxies, giving a result of $H_0 = 73.9 \pm 3.0$ km s^{-1} Mpc^{-1} [33]. The HOLiCOW collaboration measured a value of $H_0 = 73.3\ ^{+1.7}_{-1.8}$ km s^{-1} Mpc^{-1} [13] using 6 gravitationally lensed quasars, producing a value which is independent of the cosmic distance ladder yet agrees with the SH0ES measurement. Other measurements include a method which calibrates the distance to the type 1a supernovae using the tip of the red giant branch, which gave a H_0 value that sits in the centre of the tension region [34]. However, once this method was reproduced

by a separate research group, the measured value was much higher and agreed with other late-time measurements [35].

It is clear in the results outlined above that there exists a strong disconnect between early and late time measurements, and currently there is no sufficient explanation for this. There are two main hypotheses that may explain this tension. Firstly, the systematics in the measurements may introduce bias to the measured value. In the case of late-time measurements, many observations are usually combined to give a single H_0 value and the physics behind certain astronomical events is not well known. However, we currently have some late-time H_0 measurements that are independent of the cosmic distance ladder that still tend towards a higher value and we are not presently aware of a singular systematic in said ladder that could be accountable for this discrepancy.

Secondly, the ΛCDM model that is currently used to describe our universe may not be completely correct. The H_0 value derived by Planck using the CMB uses highly precise measurements with a decreased risk of any systematic error being introduced. However, due to the H_0 inference being model specific, if the ΛCDM model is incorrect then this measurement tension may be due to a breakdown in our cosmological model and not systematic effects. This would lead to the exploration of new physics to explain the Hubble tension. Any previous modifications to the ΛCDM model that may explain or reduce this tension tend to bring about tensions in other areas of cosmology. Examples of these possible modifications include modified gravity, non-zero curvature and dynamical dark energy explanations [36–41]. All of these methods do reduce the Hubble tension, however this usually favours broadening the H_0 posterior and not actually changing the central value.

It is widely agreed upon by many cosmologists that the ΛCDM model describes our universe very well but doesn't explain the full picture, much like Newton's theory of gravity in relation to Einstein's theory of relativity. It seems as though we may have reached the limit of this cosmological model, as demonstrated by the Hubble tension, however only time will tell if this is correct. The use of gravitational waves as a cosmological measure may give insight on this issue.

1.1.3 Gravitational Waves as standard sirens

As a result of his General Theory of Relativity published in 1915, Einstein was the first to predict GWs the following year in 1916 [42, 43]. His field equations described how space-time behaves when mass is present, which concluded that an accelerating mass would perturb space-time and cause ripples to form that spread outwards from the accelerating mass, travelling at the speed of light. It was highly debated whether these GWs actually exist until the 1950s, when it was shown by Pirani that GWs were not an artifact of the coordinate system but do in fact exist [44].

In the 1980s, indirect evidence of GWs came from the orbital decay of binary pulsars, however GWs had no direct evidence until 2015 [45]. Almost 100 years after they were first predicted, the Advanced LIGO detectors observed a signal from a binary black hole merger event, roughly 440 Mpc away [46]. GW observations from a CBC provide a good measure of cosmological distance independent of the distance ladder, making them extremely useful for astrophysical purposes.

GWs create strain as they distort space-time, in a way that is perpendicular to their direction of propagation. This is the basis of how GWs are detected using laser interferometers. GWs have the property of polarisation, which may be split into plus-polarised and cross-polarised components, h_+ and $h_\times$.

The properties of the GW source and the orientation of the source with respect to the observer are encoded in the strain of the GW signal, which is measured using a GW interferometer. The plus-polarised strain component is given as a function of time, assuming the quadrupole approximation [47]:

$$h_+(t) = \frac{2M_z}{d_L}\big(1+\cos^2(i)\big)(\pi M_z f)^{2/3} cos(\Theta + \Psi) \tag{1.23}$$

$$h_+(t) = \frac{2M_z}{d_L}\big(1+\cos^2(i)\big)\left(\frac{5}{256}\frac{M_z}{(T-t)}\right)^{\frac{1}{4}} \cos\left(-2\left(\frac{(T-t)}{5M_z}\right)^{\frac{5}{8}} + \Psi\right) \tag{1.24}$$

with d_L being the luminosity distance between the observer and the source, and assuming $c = G = 1$. The inclination of the binary pair relative to the observer is given as i, and f is the time-varying frequency of the signal. The phase of the signal at time $T = t$

is represented by Ψ, with T being the exact time of coalescene of the binary pair. M_z represents the redshifted chirp mass of the source:

$$M_z = (1+z)\frac{(m_1 m_2)^{\frac{3}{5}}}{(m_1+m_2)^{\frac{1}{5}}} \tag{1.25}$$

with z being the redshift of the binary system with respect to the observer and m_1 and m_2 being the primary and secondary masses of the system.

The cross-polarisation can be written as:

$$h_\times(t) = \frac{4M_z}{d_L} cos(i)(\pi M_z f)^{2/3} sin(\Theta + \Psi) \tag{1.26}$$

$$h_\times = \frac{4M_z}{d_L} cos(i)\left(\frac{5}{256}\frac{M_z}{(T-t)}\right)^{\frac{1}{4}} sin\left(-2\left(\frac{(T-t)}{5M_z}\right)^{\frac{5}{8}} + \Psi\right) \tag{1.27}$$

Both equations 1.24 and 1.27 demonstrate that the GW strain is inversely proportional to the luminosity distance of the source.

The GW strain can be given as a linear combination of the cross and plus polarisation components:

$$h(t) = F_+ h_+ + F_\times h_\times \tag{1.28}$$

with F_+ and $F_\times$ being the antenna response function of the interferometer. These factors depend on the source's sky position with respect to the detector and also the polarisation of the GW [48]. The detector-frame (redshifted) mass of the CBC and its distance relative to the detector can then be determined as the frequency and the frequency evolution over time will break the degeneracy between d_L and M_z. Any uncertainty arising due to the inclination may be marginalised over, giving a luminosity distance measurement that accounts for the inclination uncertainty.

The use of multiple detectors to measure a single GW is advantageous in that it lowers the signal to noise ratio and also provides a means to partially break the inclination and distance degeneracy, giving a tighter constrain on d_L [49]. The measurement of the sky

localisation of an event, Ω, is also improved when using multiple interferometers because of the time delay of the GW at each detector [50].

The GW signal does contain some redshift information, however it is is degenerate with the source-frame mass and only appears in the M_z parameter. This is important to note, as it calls for either a method to break the degeneracy or, more feasibly, an independent method to measure the redshift of the GW source. To accurately measure H_0, GW signals provide us with an excellent distance measurement, but the need for reliable and accurate redshift estimates of the GW source is still required. This is discussed in much greater detail in later sections.

1.1.4 Cosmology using Gravitational Waves

The first proposal outlining the use of GWs to measure H_0 was given in 1986 by Schutz [51]. He described two separate ways to do this using BNS merger events. The first method describes an event which has a well localised sky location, meaning its EM counterpart can be measured and can be unmistakably linked to the host galaxy. The host galaxy's redshift can then be identified and a measurement of H_0 can be taken using the distance estimate from the merger event. However, if the EM counterpart cannot be identified, then the second method treats every single galaxy that falls within the localisation volume as a possible host. Since the galaxies that lie within the localisation volume are randomly distributed, any measurement of H_0 using these galaxy redshift estimates and the distance estimates from the merger event would statistically average out, while the true host redshift would give the 'real' H_0 measurement.

A Bayesian derivation of Schutz's methods as outlined above was given by Del Pozzo in 2012, which included both the EM counterpart and galaxy catalogue cases [52]. However, this derivation did not include the situation where the catalogue does not account for all of the galaxies within the localisation volume, meaning that the true host may not even be contained in the catalogue for all events. This calls for a methodology which expresses the likelihood of a GW event with a host that is not present in the galaxy catalogue to ensure an unbiased measurement of H_0.

BNS merger events are extraordinarily violent, and alongside GWs they also release powerful EM emission across a wide range of wavelengths which is known as a kilonova

[53, 54]. Evidence also suggests that these BNS events also produce short gamma-ray bursts (GRBs), giving off bright flashes of gamma rays which last up to 2 seconds [55]. A GW event may have its location pin-pointed in the sky by the observation of a short GRB or a kilonova with an associated GW signal, meaning that the host galaxy could be easily identified and therefore its redshift can be found and used to estimate H_0 [56–59].

The merger event, GW170817, was revolutionary in that it was the first BNS merger detected by the Advanced Ligo and Virgo detectors with an observed EM counterpart. It was detected on the 17th of August 2017 [18], with the GW signal being measured first and then 1.7 seconds later, a GRB was observed. An immense observational campaign began, with many telescopes all around the world observing across the range of the EM spectrum leading to an initial distance estimate of 40 Mpc, localised within 31 deg^2 on the sky. The host galaxy, NGC4993, was then identified after many observations of the BNS's optical transient [60, 61]. The first ever measurement of H_0 using GW standard sirens was made using the EM and GW information, giving a value of $H_0 = 70^{+12}_{-8}$ km s^{-1} Mpc^{-1} [19]. Due to the event's strong EM counterpart, it is an excellent demonstration of the use of GW standard sirens for cosmological exploration, and will always be pivotal moment for GW cosmology.

Obviously, there is still a significant associated error with this measurement no matter how liberated it is from the cosmic distance ladder. The peculiar motions had a great impact on the redshift measurement of the event due to it being so nearby. Many have addressed this issue, for instance [19], and much work has gone into exploring how the peculiar velocities impact the final H_0 measurement [62, 63]. This highlights a considerable need for reliable galaxy redshifts estimates, not only when using the galaxy catalogue method but even when there is an EM counterpart observed. Attention has also been placed on how the luminosity distance could be further constrained by breaking the d_L-inclination degeneracy using the information provided from the EM counterpart on GW170817's inclination, which would further constrain H_0 [64].

Thus far, GW170817 is the only GW event observed with an associated EM counterpart as measured by LIGO and Virgo. However, there is a high probability of a future detection with a similar EM counterpart, which would lead to further constraint on H_0 using this method. It has been estimated that $O(100)$ GW events with associated counterparts

would have to be measured to produce a H_0 value with around 1% uncertainty [65, 66], allowing for comparison with early and late time measurements.

The galaxy catalogue method was then the main cause for concern as it may be applied to any GW event confirmed by LIGO and Virgo, including previously measured observing runs 1-3 (O1-3) mergers. A paper, [65], published in 2018 explored the use of BNSs to constrain H_0 with no observed counterpart which assumed that galaxy catalogues were complete. They also outlined the prospect of using well-localised BBHs for the same thing. Although these forecasts assumed that the catalogues contained all potential host galaxies, a method was described which used a completeness fraction as a weight for the GW likelihoods as to whether the host could be found inside or outside the catalogue, thus accounting for an incomplete galaxy survey. The GW170817 event was then used as a test for this methodology if, hypothetically, the EM counterpart was not observed. A fully Bayesian methodology was independently derived in [67], which used standard sirens to determine H_0 in the situation of an incomplete galaxy catalogue, modelling the restriction of the galaxy surveys using a magnitude threshold. After rigorous assessment using mock data anaylsis (MDAs), this method was successful in achieving an unbiased measurement H_0 with incomplete galaxy surveys.

Advanced LIGO and Virgo published their first GW transient catalogue containing eleven detections, ten BBHs and one BNS [68]. One particular BBH, GW170814, had a localisation volume which was contained within a dense cluster of galaxies and therefore was very informative, leading to the Dark Energy Survey (DES) collaboration, alongside with the LIGO and Virgo collaborations, to produce a H_0 value from the lone dark siren. The LIGO and Virgo collaborations also produced a H_0 measurement using all of the O1 and O2 events and methodology from [67] and public surveys, to give the first H_0 value using a combination of many events, as $H_0 = 68.7^{+17.0}_{7.8}$ km s^{-1} Mpc^{-1} [69].

Since then, the third observing LIGO and Virgo run has led to many more GW events being recorded [70], however the GW170817 event is still the only observation with a measured EM counterpart. This means that the elusive 1% uncertainty of the H_0 value is out of sight for now, highlighting the need for analyses that do not require EM counterparts. The future constraint of H_0 calls for the development of cosmological analysis using the much more abundant *dark* sirens. Dark siren events can be observed out to much greater distances due to the mass of BBHs being much larger than BNSs, which

opens the door for cosmological inference beyond that of just the H_0 parameter. This calls for the development of catalogue methods, which in turn requires a development of redshift estimate methods contained within said catalogues.

The degeneracy between the source-frame mass and redshift may be broken for GWs using methods that don't rely on either counterparts and galaxy surveys. For a BNS merger event, the phase evolution of the merger is dependent on tidal effects between the two masses, which can be used to break this degeneracy [71]. Also, the BNS is required to have a redshift which translates the detector-frame mass to the range of acceptable source-frame masses [72–74], and due to a narrow mass distribution allowed for the BNS, the degeneracy may be somewhat broken. However, the observation of the GW190425 event is an example of an exceptionally heavy BNS system which somewhat weakens this argument [75]. Similarly, the mass-redshift degeneracy may also be broken for BBHs using the sharp features in their mass distributions. This could lead to a competitive H_0 measurement with $O(10{,}000)$ BBH detections [76].

A key takeaway of this section is that events which lack much catalogue support can be somewhat uninformative, yet the population of these events overall provides useful information when inferring cosmological parameters. Current and future developments relating to GW cosmology are exciting, with the increasing sensitivity of detectors with every observing run, leading to higher SNRs and an increasing number of detections. An addition to the GW detector network in time for the fourth observing run, the KAGRA detector in Japan, will give rise to more coincident detections of GWs, thus increasing the SNR and giving better localisation of the event. This increased localisation will not only reduce the number of potential hosts for dark sirens, meaning their contribution will be more informative to the measurement of H_0, but will also increase the likelihood of identifying any associated EM counterparts with bright sirens.

Now, attention will turn to `gwcosmo`, a code which implements Schutz's methodology for using galaxy catalogues in conjunction with dark siren events to give H_0 estimates. This method has and will be used to measure H_0 using galaxy catalogues methods and is a key component for cosmological inference for the upcoming fourth observing run. An outline of this code will be given in the following section to provide context as to why producing reliable and accurate photometric redshift estimates and their corresponding errors is imperative in the inference of H_0.

1.2 `gwcosmo`

The code `gwcosmo` uses a Bayesian framework to combine many GW events, and is able to consider both when an associated EM counterpart is observed for each event, and when there is not, calling for the use of galaxy catalogues to fill in any missing redshift information. The code has been rigorously testing using both mock data analysis (MDAs) [67] and using data from previous observing runs [16]. It has also been independently tested in [77], whereby `gwcosmo` was implemented to measure H_0 using the GW190814 and GW170817 events, giving a value of $H_0 = 70^{+29.0}_{-18.0}$ km s^{-1} Mpc^{-1} . With the approaching fourth observing run, it is hoped that `gwcosmo` can be applied to new GW signal measurements to further constrain the measured value of H_0 thus potentially shedding light on the Hubble tension issue.

The entirety of the methodology and mathematical expressions described in this section has been derived originally in [16] and so the following section can be entirely credited to this work.

1.2.1 Bayesian Framework

Firstly, the Bayesian framework used to combine the information from many GW signals resulting in a H_0 measurement is introduced. The situation in which there is an EM counterpart observed in not discussed here, as the scenario where there is no EM counterpart and the galaxy catalogue method is used is the one which is pertinent to this work. This thesis focuses on reliable methods to generate photometric redshifts which may be used to populate these galaxy catalogues of potential host galaxies, therefore the EM counterpart method is, although interesting, not relevant.

Parameters used in the discussion below are listed in Table 1.1. From $N_{\rm det}$ GW events, the posterior probability on H_0 can be given as:

$$p(H_0|\{x_{GW}\},\{D_{GW}\},I) \propto p(H_0|I)p(N_{det}|H_0 I)\prod_i^{N_{det}} p(x_{GWi}|D_{GWi},H_0,I) \qquad (1.29)$$

Parameter	Definition
H_0	Hubble constant
N_{det}	number of events detected during observation period
x_{GW}	the GW data associated w some GW source s
D_{GW}	denotes that the GW signal was detected (ie. x_{GW} passed some detection statistic threshold)
g	denotes that galaxy is (G) or is not ($\bar{G}$) contained within galaxy catalog
x_{EM}	EM data associated w EM counterpart
D_{EM}	denotes the EM counterpart was detected (ie. x_{EM} passed some threshold)
I	additional info not explicitly stated eg. underlying cosmological model

TABLE 1.1: A summary of the parameters used in the discussion of the Bayesian methodology.

with x_{GW} being the set of GW data which corresponds to N_{det} detections. x_{GWi} must have passed some given threshold and been officially detected as an event, which is represented by D_{GWi}. The prior on H_0 is given by $p(H_0|I)$. Over the observational period, the probability of detecting N_{det} events for a given H_0 value is $p(N_{det}|H_0, I)$. This is dependent on the intrinsic rate of events within the source frame, $R = \frac{\partial N_s}{\partial V \partial T}$, with N_s being the number of sources [16]. The units of R are the number of events per unit comoving volume per unit time, as measured in the observer frame. $N_{det} = R\langle VT\rangle$ denotes the number of expected detections, with $\langle VT\rangle$ being the average of the observation time multiplied by the surveyed comoving volume. The H_0 dependence is removed by selecting the prior on the rate, $p(R|I) \propto 1/R$. Any other terms left factorise to give individual likelihoods for each GW event [16]. A single GW event, labelled i, may then be expressed by the following, where we omit the i subscript for ease of notation:

$$\begin{aligned} p(x_{GW}|D_{GW}, H_0, I) &= \frac{p(D_{GW}|x_{GW}, H_0, I)p(x_{GW}|H_0, I)}{p(D_{GW}|H_0, I)} \\ &= \frac{p(x_{GW}|H_0, I)}{p(D_{GW}|H_0, I)} \end{aligned} \tag{1.30}$$

Here, $p(D_{GW}|x_{GW}, H_0, I) = 1$ due to the fact that any analysis is only performed when x_{GW} has passed some defined detection threshold. During this analysis, it is assumed that every event has passed this threshold and therefore can be deemed as a detected event. To calculate $p(D_{GW}|H_0, I)$, all realisations of the detectable GW event must be integrated over [16]. The integral must therefore be performed over all values of x_{GW} that exceed the detection threshold, as so far for which $p(D_{GW}|x_{GW}, H_0, I) = 1$. This then gives:

$$\begin{aligned} p(D_{GW}|H_0, I) &= \int p(D_{GW}|x_{GW}, H_0, I)p(x_{GW}|H_0, I)dx_{GW} \\ &= \int^{x_{GW\,det}} p(x_{GW}|H_0, I)dx_{GW} \end{aligned} \tag{1.31}$$

This calls for information on the detector configuration, detection threshold, sensitivity and the GW source population [16].

1.2.1.1 The Galaxy Catalogue Method

As previously stated, this work focuses on the use of the galaxy catalogue method to calculate H_0, as this is the method which performs analysis using EM information provided by galaxy surveys. This gives context for the requirement of reliable photometric redshift estimates, as these surveys must be accurate and preferably complete to give the best constraint on the final H_0 value. The apparent magnitudes in multiple bands and the sky location of the galaxies are also provided by the catalogues. The GW event may not be contained within the galaxy catalogue, and this possibility must also be considered in this analysis. A single GW event can be expressed in terms of its likelihood, which marginalises over both the cases where the host galaxy is in the galaxy catalogue, and when it is not, which is represented by G and $\tilde{G}$ respectively:

$$p(x_{GW}|D_{GW}, H_0, I) = \sum_{g=G,\tilde{G}} p(x_{GW}|g, D_{GW}, H_0, I)p(g|D_{GW}, H_0, I) \tag{1.32}$$

$$\begin{aligned} p(x_{GW}|D_{GW}, H_0, I) &= p(x_{GW}|G, D_{GW}, H_0, I)p(G|D_{GW}, H_0, I) + \\ &= (x_{GW}|\tilde{G}, D_{GW}, H_0, I)p(\tilde{G}|D_{GW}, H_0, I) \end{aligned} \tag{1.33}$$

When the host galaxy can be found within the catalogue, (G), then any EM information can be used to modify the sky location, magnitude and galaxy redshift priors. However, in the case where the host galaxy is not contained within the catalogue, the term $\tilde{G}$ includes information on the limits of the galaxy survey [16]. The catalogue is firstly modelled as having an apparent magnitude threshold, so that the observed apparent

magnitude of any given galaxy will determine whether or not it is included in the catalogue. A detailed derivation of the terms in Equation 1.33 can be found in [16], however only the relevant component of the derivation is described below which pertains to the redshift and its associated uncertainty.

When considering the redshift in the likelihoods of a GW event, the prior on the potential host galaxies is the focus and not the prior on all the galaxies in the survey. For host galaxies of a GW event, the redshift, absolute magnitude, apparent magnitude (m) and sky location prior may be expressed as:

$$\begin{aligned} p(z, \Omega, M, m|s, H_0, I) &= p(m|z, \Omega, M, s, H_0, I)p(z, \Omega, M|s, H_0, I) \\ &= p(m|z, \Omega, M, s, H_0, I)p(z|s, I)p(\Omega|I)p(M|s, H_0, I) \end{aligned} \tag{1.34}$$

with z, Ω and M, the absolute magnitude, is assumed to be conditionally independent given s, H_0 [16]. The s term here denotes that a GW signal has been emitted. The m term may be directly found if z, M and H_0 are known.

$$\begin{aligned} p(z, \Omega, M, m|s, H_0, I) &= \delta(m - m(z, M, H_0))p(z|s, I)p(\Omega|I)p(M|s, H_0, I) \\ &= \delta(m - m(z, M, H_0))\frac{p(s|z, I)p(z|I)}{p(s|I)}p(\Omega|I)\frac{p(s|M, I)p(M|H_0, I)}{p(s|H_0, I)} \end{aligned} \tag{1.35}$$

The $p(s|H_0, I)$ and $p(s|I)$ terms cancel out, meaning their exact form can be dismissed. Ω has no dependence on s if it is assumed that the universe is isotropic, meaning that no one sky location is more likely to contain a GW event than any other, so the equation does not contain a $p(s|\Omega, I)$ term [16].

The probability that a GW event is hosted by a galaxy with magnitude M is given by $p(s|M, I)$, given by:

$$p(s|M, I) \propto \begin{cases} L(M), & \text{if GW hosting is proportional to luminosity} \\ \text{constant}, & \text{if GW hosting is independent to luminosity} \end{cases} \tag{1.36}$$

The merger rate is dependent on redshift, which is represented by the term $p(s|z, I)$:

$$p(s|z, I) \propto \begin{cases} \frac{1}{1+z} R(z), & \text{if rate evolves with redshift} \\ \frac{1}{1+z}, & \text{if rate does not evolve with redshift} \end{cases} \tag{1.37}$$

Due to the time delay between the emission of the GW signal and its observation on Earth, the factor $1/(1 + z)$ is present in the above as the universe has expanded in the time between the GW being emitted and observed. The term converts between the source and detector frames even if the intrinsic merger rates of the CBCs are constant with time and therefore independent of redshift [16].

1.2.1.2 Likelihood when the host is in the galaxy catalogue

By expanding Equation 1.30, the likelihood of GW data when the host is contained within the galaxy survey can be expressed by marginalising over the redshift, absolute and apparent magnitudes and the sky location:

$$p(x_{GW}|G, D_{GW}, s, H_0, I) = \frac{p(x_{GW}|G, s, H_0, I)}{p(D_{GW}|G, s, H_0, I)} \tag{1.38}$$

Since x_{GW} is independent of G, m and M, the numerator can be expanded and the above equation can be written as

$$p(x_{GW}|G, s, H_0, I) = \iiiint p(x_{GW}|z, \Omega, s, H_0, I) p(z, \Omega, M, m|G, s, H_0, I) dz d\Omega dM dm \tag{1.39}$$

G being on the right hand side of this equation specifies that z, m, M and Ω are given by the galaxy catalogue, with any uncertainties being ignored for now [16]. The absolute magnitude can be easily calculated using m, z, H_0 and M, leading to further factorisation:

$$p(x_{GW}|G, s, H_0, I) = \frac{p(s|z, M(z, m, H_0), I) \delta(M - M(z, m, H_0)) p(z, \Omega, m|G, I)}{p(s|G, H_0, I)} \tag{1.40}$$

Ω has been dropped from the right side of the $p(s|z, M(z, m, H_0), I)$ term, and the dependence of s on m and H_0 has been recognised to only enter through their relationship with M [16]. Equation 1.39 may now be integrated over absolute magnitude to give:

$$p(x_{GW}|G, s, H_0, I) = \frac{1}{p(s|G, H_0, I)} \iiint p(x_{GW}|z, M(z, m, H_0), I) \times p(z, \Omega, m|G, I) dz d\Omega dm \tag{1.41}$$

Henceforth, the prior on z, m and Ω for all of the galaxies within the catalogue is given by $p(z, \Omega, m|G, I)$, whereas $p(x_{GW}|z, M(z, m, H_0), I)$ denotes the probability that galaxy with z and M values would host a GW source [16]. It may be assumed that all of the galaxies within the survey can be approximated as delta functions on z, m and Ω, meaning the integral may be converted into a sum of the number of galaxies, N, in the survey:

$$\begin{aligned} p(x_{GW}|G, s, H_0, I) &= \frac{1}{p(s|G, H_0, I)} \frac{1}{N} \sum_{i=1}^{N} p(x_{GW}|z_i, \Omega_i, s, H_0, I) p(s|z_i, M(z_i, m_i, H_0), I) \\ &= \frac{1}{p(s|G, H_0, I)} \frac{1}{N} \sum_{i=1}^{N} p(x_{GW}|z_i, \Omega_i, s, H_0, I) \frac{p(z_i, M(z_i, m_i, H_0)|s, I) p(s|I)}{p(z_i, M(zi, m_i, H_0)|I)} \\ &= \frac{1}{p(s|G, H_0, I)} \frac{1}{N} \sum_{i=1}^{N} p(x_{GW}|z_i, \Omega_i, s, H_0, I) \frac{p(z_i|s, I) p(M(z_i, m_i, H_0)|z_i, s, I) p(s|I)}{p(z_i|I) p(M(z_i, m_i, H_0)|z_i, I)} \end{aligned} \tag{1.42}$$

The application of Bayes theorem to the terms $p(z_i|s, I)$ and $p(M(z_i, m_i, H_0)|z_i, s, I)$ produces the following equation [16]. Henceforth, the dependence of $p(M(z_i, m_i, H_0)|z_i, s, I)$ on z_i is contained in the $M(z_i, m_i, H_0)$ term by definition, meaning that z_i may be removed from the right hand side.

$$\begin{aligned} p(x_{GW}|G, s, H_0, I) &= \frac{1}{p(s|G, H_0, I)} \frac{1}{N} \sum_{i=1}^{N} p(x_{GW}|z_i, \Omega_i, s, H_0, I) \\ &\times \frac{p(s|z_i, I) p(z_i|I) p(s|M(z_i, m_i, H_0), I) p(M(z_i, m_i, H_0)|I) p(s|I)}{p(s|I) p(s|I) p(z_i|I) p(M(z_i, m_i, H_0)|I)} \\ &= \frac{1}{p(s|G, H_0, I) p(s|I)} \frac{1}{N} \sum_{i=1}^{N} p(x_{GW}|z_i, \Omega_i, s, H_0, I) p(s|z_i, I) p(s|M(z_i, m_i, H_0, I)) \end{aligned} \tag{1.43}$$

Most of the terms have now cancelled, and the term $1/p(s|I)$ can be brought outside of the sum.

Now, we may return to Equation 1.38 and focus on the denominator. The $p(D_{GW}|G, s, H_0, I)$ term is expanded by marginalising over z, Ω, m and M which gives a similar term to Equation 1.43 but with $p(D_{GW}|z_i, \Omega_i, s, H_0, I)$ instead of $p(x_{GW}|z_i, \Omega_i, s, H_0, I)$. This is subsituted into Equation 1.43 alongside Equation 1.38 to produce the likelihood for the case where the host galaxy is in the catalogue. Here, the factors $p(s|G, H_0, I)$ and $p(s|I)$ cancel, giving:

$$p(x_{GW}|G, s, H_0, I) = \frac{\sum_{i=1}^{N} p(x_{GW}|z_i, \Omega_i, s, H_0, I)p(s|z_i, I)p(s|M(z_i, m_i, H_0), I)}{\sum_{i=1}^{N} p(D_{GW}|z_i, \Omega_i, s, H_0, I)p(s|z_i, I)p(s|M(z_i, m_i, H_0), I)} \quad (1.44)$$

The above equation is suitable for a very simplistic case, where galaxies found in the catalogue are delta-function-like and contain no uncertainties [16]. However, this is typically not the case, as most surveys contain uncertainties with redshift uncertainties being particularly significant. In the situation where redshift errors are provided (whereby the redshift distribution of the ith galaxy is given as $p(z_i|I)$), they may be included as:,

$$\mathrm{p}(\mathrm{x}_{GW}|G, s, H_0, I) = \frac{\sum_{i=1}^{N} \int p(x_{GW}|z_i,\Omega_i,s,H_0,I)p(s|z_i,I)p(s|M(z_i,m_i,H_0),I)p(z_i|I)dz_i}{\sum_{i=1}^{N} \int p(D_{GW}|z_i,\Omega_i,s,H_0,I)p(s|z_i,I)p(s|M(z_i,m_i,H_0),I)p(z_i|I)dz_i} \quad (1.45)$$

This means that the uncertainty associated with the redshift of each galaxy is marginalised over [16].

The above equation describes the likelihood of the GW data in the situation where the host galaxy is contained within the galaxy survey. Differing H_0 values cause the galaxies to 'pick out' different parts of the likelihood. Depending on how well the galaxy's redshift is consistent with the luminosity distance of the GW event will determine how much that galaxy contributes to the final H_0 value. This is also weighted by the astrophysical properties of the galaxy, as this determines how likely it is to host a GW event. For more detail on the derivation of Equation 1.32 and how the probability of the host being inside or outside the catalogue is evaluated see [16]. The key take away from this section

is that the likelihood of a GW event detection depends on the redshift and its associated uncertainty, which are both marginalised over.

1.2.1.3 Implementation

The previous paragraphs discussed the derivation involved to use galaxy catalogues to make a measurement of H_0 when the exact host galaxy is unknown. The following subsection details how this is actually implemented in the `gwcosmo` code to provide more context for the importance of accurate galaxy redshifts in the quest to constrain H_0.

When using GW detection for cosmological inference, more often than not the event undergoes parameter estimation, giving posterior samples which cover many parameters such as inclination, sky location and luminosity distance [16]. Here we note that the posterior samples have already had a prior applied in their generation. Beginning with:

$$p(x_{GW}|d_L(z,H_0),\Omega,I) = \frac{p(d_L,\Omega|x_{GW},I)p(x_{GW}|I)}{\pi(d_L|I)\pi(\Omega|I)} \tag{1.46}$$

Where $p(d_L,\Omega|x_{GW},I)$ denotes the posterior samples on d_L and Ω for the event x_{GW}. The $p(x_{GW}|I)$ term can be disregarded as a normalisation constant as it does not depend on H_0. The priors on d_L and Ω are given by $\pi(d_L|I)$ and $\pi(\Omega|I)$ and were used during the parameter estimation of x_{GW} [16].

The equation is expanded to separate d_L and Ω:

$$\begin{aligned} p(x_{GW}|d_L(z,H_0),\Omega,I) &= \frac{p(d_L|\Omega,x_{GW},I)p(\Omega|x_{GW},I)p(x_{GW}|I)}{\pi(d_L|I)\pi(\Omega,I)} \\ &\approx \frac{p(d_L|x_{GW},I)p(\Omega|x_{GW},I)p(x_{GW}|I)}{\pi(d_L|I)\pi(\Omega,I)} \end{aligned} \tag{1.47}$$

$p(d_L|x_{GW},I)$ is used to generate the d_L posterior samples, however this function needs to be smoothed. This is done by using kernal density estimation (KDE) to generate a 1-dimensional function over d_L, $f_{KDE}(d_L) \approx p(d_L|x_{GW},I)$ so that the function can be integrated over M and z and be evaluated at delta-like-function galaxies [16].

The $p(\Omega|x_{GW}, I)$ term is evaluated using skymaps rather than represented by a KDE on the RA and declination posterior samples. These coordinates are spherical in nature, meaning that issues could potentially arise when analysing events that cross the prior boundary for RA and wrap around (for example, 2π to 0). The skymaps are advantageous, as they avoid this issue and contain the 2d GW sky probability, split in equally sized pixels [16]. These pixels are equivalent to $p(\Omega|x_{GW}, I)$. When integrated over the whole sky, this gives:

$$\int p(\Omega|x_{GW}, I) = \sum_{k=1}^{N_{pix}} P(k|x_{GW}, I) = 1 \tag{1.48}$$

Ω_i lies within a pixel k, and the probability contained within said pixel is given by$P(k|x_{GW}, I)$. The probability density at sky location Ω_i can be expressed as:

$$p(\Omega_i|x_{GW}, I) = P(k_{\Omega_i}|x_{GW}, I)\frac{N_{\text{pix}}}{4\pi} \tag{1.49}$$

with $P(k_{\Omega_i}|x_{GW}, I)$ being the probability contained within the pixel that Ω_i can be found in.

The desired final equation for $p(x_{GW}|z, \Omega, H_0, I)$ to be implemented by `gwcosmo` requires the acknowledgment that any normalisation constant of the GW likelihood is unimportant on the condition that is does not depend on H_0, z or Ω [16]. This constant is the same as that present in Equation 1.45, meaning the term $p(x_{GW}|z, \Omega, H_0, I)$ can be expressed as:

$$p(x_{GW}|z, \Omega, H_0, I) \approx C\frac{f_{KDE}(d_L}{\pi(d_L}\frac{p(\Omega|x_{GW}, I)}{\pi(\Omega|I)} \tag{1.50}$$

with C being some function that is constant with respect to H_0.

However, a quantification of the redshift uncertainty still needs to be included in the code. Galaxy catalogues often provide redshift uncertainties which contribute to the overall H_0 measurement error, which regularly come in the form of photometric estimates, as spectroscopic values are much harder to obtain. This is discussed in great detail in the following sections. Large surveys call for photometric estimates which usually have significant associated uncertainties. Photometric redshifts also tend to decrease

in reliability out to larger redshifts, and risks some particular structure or bias being introduced to the catalogue. The true structure present is due to galaxy clustering, yet the redshift estimates could introduce some other structure. In general, catalogues provide a mean value of the redshift and some error, normally a 1σ value or the upper and lower 1σ bounds in the case of an asymmetric distribution. When `gwcosmo` has previously been used, it has been assumed that the redshift uncertainty distribution is Gaussian, due to the lack of detailed catalogue information. Assuming a Gaussian distribution does blur out any particular structure introduced by the redshift estimates, but a true quantification of the redshift posteriors gives a more accurate result in the overall inference of H_0 [16]. If posterior samples are available then they can be directly used for a Monte Carlo marginalisation over the redshift uncertainty. A method of accurately generating photometric redshift posterior samples would allow this marginalisation to be performed for all galaxies in the catalogue, therefore increasing the precision and accuracy of any results. Currently, gwcosmo has been applied to O2 data using 6 GW signals (GW150914, GW151226, GW170104, GW170608, GW170809 and GW170814) to give a value of $H_0 = 68.8^{+15.9}_{-7.8}$ km s^{-1} Mpc^{-1} [16]. This result is still far from the desired 1% uncertainty, however many advancements over the last year within the LIGO cosmology team will seek to achieve to a smaller uncertainty, including the availability of accurate photometric redshift posteriors.

Now that the use of redshift in cosmological inference has been broadly discussed and it is clear that there is a need for accurate redshift estimates and posteriors in the quest to constrain H_0, it is now time to turn attention to how these redshifts and their associated errors may be acquired to populate galaxy surveys.

Chapter 2

Determining Galaxy Redshifts

2.1 Redshift

Before being able to infer meaningful cosmological parameters from astronomical sources, the redshift of the source must first be acquired in the process of building a galaxy survey. In the previous section, the motivation to obtain accurate photometric redshift posteriors to feed into the `gwcosmo` code was detailed. The light emitted from an object is redshifted as the object's spectrum is displaced towards longer wavelengths. The three primary astronomical and cosmological reasons behind this redshift are: gravitational redshift, whereby the emitted photons move towards an object in flatter spacetime and therefore a weaker gravitational potential, which causes them to increase in wavelength. Secondly, the relativistic Doppler effect is the lengthening in the wavelength of light due to the source receding from the observer. This phenomenon is attributed to special relativity and time dilation [78]. The final cause of electromagnetic redshift is cosmological redshift, which occurs as photons travel through expanding space. As the universe itself expands, the light emitted from the source is shifted to longer wavelengths as it travels through expanding space. This redshift is therefore due to the expansion of space itself and is not caused by the motion of an object or observer [79]. The further the object is from the observer, the higher the redshift and the greater its recessional velocity, as described by the Hubble Law. It is possible for photons to be redshifted by both cosmological redshift and Doppler shift; for example, a distant binary star. The cosmological redshift would be due to the expansion of space and be dependant on the distance from the earth to the binary system, and the Doppler shift would be caused by the individual

motions of the stars in the binary. The current work focuses on measuring cosmological parameters such as the Hubble constant and so the term redshift will henceforth describe cosmological redshift.

The redshift of extragalactic objects is a difficult property to measure, and may be determined via spectroscopy or photometry. Spectroscopic redshifts are advantageous in that they introduce much less random and systematic error than photometric techniques, however they are much harder to measure [80]. Spectrometry requires the distinct electromagnetic spectral energy distribution (SED) of each galaxy to determine the redshift, which is made up of a continuum and emission/absorption lines. The SED is redshifted by the expansion of the Universe to longer wavelengths by the relation:

$$\lambda_{\text{em}} = \frac{\lambda_{\text{obs}}}{1+z} \tag{2.1}$$

Where z is the redshift measured, λ_{em} is the original wavelength of the spectral feature and λ_{obs} is the observed wavelength we measure [81].

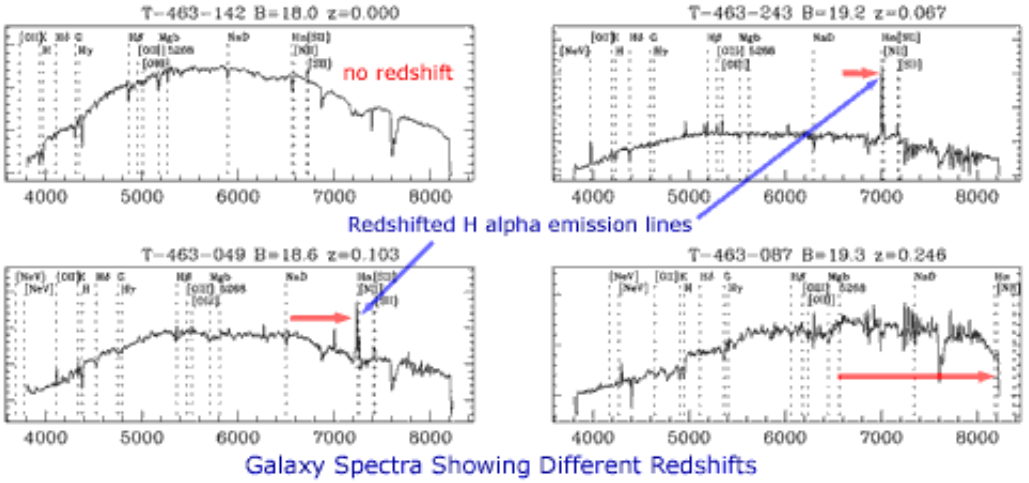


FIGURE 2.1: Four different galactic spectra from the 2dF Galaxy Redshift Surveys with the differing redshifts of the galaxies. Demonstrating the redshift of the Hα emission lines. Figure credit: [3]

Figure 2.1 shows the galactic spectra of four galaxies taken from the 2dF Galaxy Redshift Survey, with redshifts of 0.000, 0.0067, 0.103 and 0.246. The red arrows indicate the Hα spectral feature on each galactic spectra. These lines are more prominent in the $z = 0.067$ and 0.103 spectra, suggesting that these galaxies are actively forming [3]. As can clearly be seen in the image, the Hα spectral feature is shifted beyond the rest frame

value of 6563 Åas the redshift increases from zero. The difference in values between rest frame Hα wavelength and the redshifted wavelength can be inserted into Equation 2.1 to determine the redshift of the galaxies. This is a general example of how the redshift of a galaxy may be determined using spectroscopic techniques.

Spectroscopic redshifts are much more accurate, however these are often hard to obtain and therefore more expensive than other photometric redshift techniques. Difficulty in spectroscopic redshift estimations arise when determining characteristic features in the spectra and quantifying how much these features have been shifted. Even if the SED has a good signal-to-noise ratio, only a small percent of sources detected by deep imaging surveys have a sufficient resolution to give meaningful emission/absorption spectra [82]. Characteristic features in the SED also include the Balmer break below 4000 Å, which is accredited to the absorption of photons that are more energetic than the Balmer limit; and the Lyman break below 1216 Å, due to the fact that radiation above the Lyman limit of 912 Åis almost completely absorbed by the intergalactic medium along the line of sight [83]. Using broad filters to measure the flux of an object gives a sparse sampling of the SED which constrains the continuum shape and allows for identification of broad features such as the Balmer and Lyman breaks, which may then be used to calculate the photometric redshift. These distinctive features may be identified in the SED and used to measure the redshift of the object by measuring how much the breaks have been shifted in wavelength. Absorption and emission spectra analysis gives rise to accurate spectroscopic redshifts, whilst photometric analysis produces low resolution redshift estimates. Due to the multi-object spectrographs limited spectral coverage and limited SNR in spectra for faint objects, spectroscopic redshifts may only be determined for up to 50-70% of deep galaxy surveys [79].

2.1.1 Methods for Estimating Photometric Redshift

Due to spectroscopic techniques being expensive and difficult to determine for many galaxies, attention has been turned to focus on photometric redshift methods. Photometric redshift techniques measure the redshift by observing the brightness of galaxies observed through broadband photometric filters and making use of information about how features in the galaxy spectra move through those filters, and so change the relative brightness of the galaxy in the different bands. Photometric redshifts are advantageous

in that they allow for the derivation of redshift estimates for any source identified by an imaging survey, however the precision of the redshift inference is lowered by a factor of 10-100 times compared to a low-resolution spectrograph. This lack of precision is due to photometric filters being sensitive to wavelength range and relying on cosmological assumptions and assumptions on the nature of the spectrum at the source [79]. The use of photometric redshift surveys in cosmological inference calls for the accurate assessment of photometric redshift performance, which requires deep spectroscopic samples that highlight the complimentary aspects of photometric and spectroscopic redshift surveys [83].

Photometric methods have long been a popular method of redshift determination, with the first instance being accredited to Baum in 1962. He developed a technique using a photometer and 9 bandpasses to observe the SED of 6 elliptical galaxies and compared this to other known elliptical galaxy SED distributions. From this, he was able to derive the difference between the energy distributions and therefore the redshift of the galaxies. These measurements were fairly accurate, however they were dependant on a large 4000 $\mathring{A}$ break feature and so were only applicable to spectral galaxies [84]. The CfA Redshift Survey was the first systematic survey of its kind in 1977, measuring the photometric redshifts of around 2200 galaxies. Since the spectrum of each galaxy was measured one at a time, early redshift surveys were very limited in size. Once multi-slit and fibre-optic spectrographs became available in the early 90s, many spectra may be observed at once and the size of redshift surveys increased dramatically. The Hubble Deep Field (HDF) observations, published in 1996, led to peaked interest in photometric redshift techniques as the data was ideal for photometric redshift applications, as it contained very deep multiband information spanning the visible range [85]. This provided a huge advancement in high redshift Universe studies, and lead to many attempts in increasing the robustness of photometric redshift techniques. Due to its depth, it is impossible to measure spectra for the majority of the galaxies in the survey. Since then, many spectroscopic and photometric redshift surveys have been published using HDF objects, which advanced the development of photometric techniques. These catalogs include Lanzetta, Yahil, Fernandez-Soto (1996), Sawicki, Lin, Yee (1997), Wang, Bahcall, Turner (1998), Lanzetta, Yahil (1999) and Furusawa et al. (2000) [83]. Photometric redshifts are also extremely useful in the identification of incorrect spectroscopic redshifts, which was made apparent using the HDF catalog. By plotting

the photometric versus spectroscopic redshift for many galaxies contained in the HDF data, it was observed that some discordances were due to mistakes associated with the spectroscopic redshifts themselves [79]. The DEEP2 Galaxy Redshift Survey is currently densest and highest-precision redshift catalogue to date out to z $\approx$ 1, which covers 2.8 deg^2 and contains over 50,000 galaxies [86].

The core principles of photometric redshift techniques lies in mapping between a range of colours (or fluxes) and redshift. Comparing the mapping with observed fluxes of a studied source leads to the redshift Probability Distribution Function (PDF) which finally gives the redshift solution. Mapping may be performed by a multitude of methods, including using machine learning methods, where a representative training sample is used with known redshift and photometry, or using template-fitting methods, whereby the redshift-colours mapping is based upon previous scientific knowledge. The use of additional priors improves the performance of both methods [83].

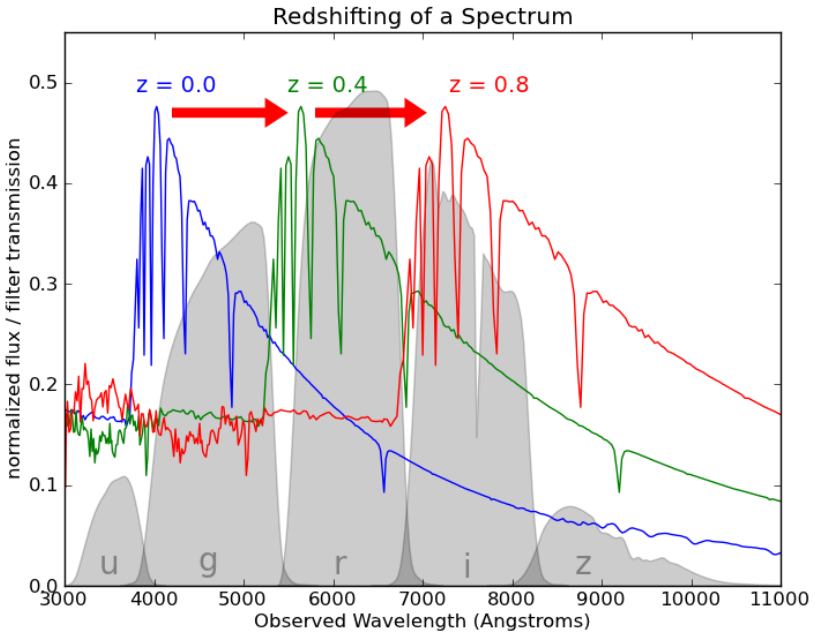


FIGURE 2.2: The spectrum of the star Vega (α-Lyr) at three different redshifts. The SDSS ugriz filters are shown in gray for reference. Figure credit: [4]

By localising the spectral fingerprint of a particular element or process, the photometric redshift of an object can be estimated. This shift of the spectrum, due to the expansion of the universe as the light is travelling to the observer, means that identical sources at different redshift values will have a different colour through each filter. Figure 2.2 shows

the spectrum of the α-Lyr star at three different redshift values. It is clear from the image that at a redshift of 0.0, the spectrum is bright in the u and g bands and much dimmer in the i and z filters. As the redshift increases, the spectrum shifts through the bands. Once the star has reached a redshift of z = 0.8, it is much brighter in the i and z bands and very dim in the g and r bands. This shift in spectral features into different observed wavelength bands provides the basis of photometric redshift techniques [4]. This same methodology may be applied to galaxy spectra.

2.1.1.1 Physically motivated methods

Template fitting methods are a popular method of estimating photometric redshifts as they can be easily and quickly applied to new data. This method compares known galaxy SEDs with those derived from templates at different redshifts to pinpoint spectral features and thus generate redshift estimates. Template methods are generally reliable, however like most photometric techniques they require a representative sample of spectroscopic data. Uncertainties arise due to colour-redshift degeneracies, incorrect fitting of colours or magnitudes to the template spectra and other measurement issues [82].

In order to apply the photometric redshift technique, one must first isolate the wavelength position of the redshifted spectral features. As a spectral break is caused by the rapid increase in continuum flux from the blue to red part of the SED, it may be identified by observing the flux through adjacent filters which are selected to encompass key features at the redshift range of interest. It is vital to ensure that the multi-wavelength coverage is broad enough to limit the risk of photometric redshift degeneracies, which is done by combining a range of several colours of filters [87].

The mapping between the redshift and observed flux is easily predicted for astronomical objects by accounting for the physical processes regulating the light emission. This means that the SED templates must be defined by scientific theory or observations. Stellar population synthesis models provide the basis for theoretical models, however these templates rely heavily on astrophysical assumptions [80]. Observational templates are drawn from observed galaxy spectra over an entire wavelength range. The quality of the photometric redshift is dependant on both the type of template and their optimal coverage of the colour-redshift space. The reliance on astrophysical and cosmological

assumptions makes physically driven photometric methods unfavourable, as we cannot be sure of the accuracy of our assumptions.

Templates should also consider nebular emission lines emitted by HII regions, which may contribute to the improvement of photometric redshift accuracy by a factor of 2.5 [83]. Dust attenuation reddens the SED continuum as the dust residing in the interstellar medium absorbs and scatters light, mainly affecting the UV section of the SED. This must be accounted for by modelling the dust attenuation as free parameters when computing photometric redshift template-fitting codes to ensure redshift values are accurate for $z > 1$ [80]. Dust from the Milky Way also attenuate light along the line of sight, however this is usually corrected for by galaxy catalogs using dust extinction maps and may be disregarded in template-fitting codes. Modelled fluxes are then made by integrating the redshifted templates through filter transmission curves. Charged Couple Device transmission, the optics of the telescope, the efficiency of the filter curves and the impact of the Earth's atmosphere may all modify the light distribution and tend to be integrated into a single transmission curve and stored within the code libraries. One should note that this discussion applies to the modelling of extra-galactic sources, however stellar templates must also be considered in template fitting as we cannot automatically assume the Galactic nature of the studied sources. Usually, stellar templates are fitted independently and the galactic nature of the source is decided a posteriori [79].

Currently, there are many succesful template fitting algorithms, such as (e.g. Benítez 2000; Bolzonella, Miralles Pelló 2000; Csabai et al. 2003; Ilbert et al. 2006; Feldmann et al. 2006; Assef et al. 2010) which either use empirical (e.g. Coleman, Wu Weedman 1980; Assef et al. 2010) or synthetic spectral templates (e.g. Bruzual Charlot 2003) [83].

2.1.1.2 Data driven methods

The increasing number of multi-wavelength surveys and public access to well tested photometric redshift codes have made photometric techniques popular, and comparison with spectroscopic redshift surveys out to very faint galaxies provides reassuring confidence [87]. Photometric redshifts are used in a range of scientific applications, such as the search for primordial galaxies, cluster identification, exploring the relationship between

galaxy properties and their dark energy halos and, of course, for inference of cosmological parameters such as the Hubble constant. They may also be used in weak lensing tomography, which requires less stringent photometric redshift precision in comparison to studying galaxy evolution [83].

Photometric redshifts may also be determined using data driven approaches, which is advantageous in that they don't rely on physical models. Machine Learning (ML) is a popular method whereby ML algorithms take training samples and learn the mapping between colour and redshift. This may be done by supervised learning, which calls for reliable spectroscopic redshift data samples and photometry, or unsupervised learning which only requires photometry. ML photometric redshift estimations are performed by finding a function which maps between the multi-dimensional photometry space and the training sample's redshift values. This learned function is then used to perform function approximation, whereby the photometry of a source with unknown redshift is localised in the multi-dimensional photometry space and paired with a corresponding redshift value or probability distribution. Supervised ML methods are interpolative in their nature, and so the space of photometric properties of the sample for which predictions will be made must be well sampled by the training data to ensure the algorithm does not lose accuracy [87]. ML methods have been chosen over template fitting as they consistently out-perform template fitting estimations [83].

Random Forests (RF) and Neutral Networks (NN) are two commonly used supervised ML algorithms. RFs make use of decision trees to group the training sample's properties into cells, which are defined to minimise the spectroscopic redshift dispersion of the data in each cell. Each cell is then assigned the average value of the spectroscopic redshift. The photometric properties of a source sample may then be passed through each tree and averaged over to give a predicted redshift value. The theory behind RF algorithms is explored in detail in Section 2.9. NNs perform complex non-linear matrix transformations of the input properties to provide an output [88]. The output, for example, redshift, is defined by the user and training data is used to tune the transformations to minimise the residuals between the true and predicted redshift values. Deep Machine Learning (DML) is based upon NNs, but uses thousands of neurons hidden in each layer. It may be used to estimate photometric redshifts directly from galaxy images which avoids having to compute photometric redshifts from catalogs which use different surveys and therefore may use different measures. ML techniques to determine redshift

were first implemented in 1995 then further developed in 1997, which used a training sample to find a mapping between colours and redshift using a polynomial function [82]. This process has much been improved upon, with more recent algorithms including NNs made by Collister and Lahav (2004) and Oyaizu et al. (2008), Quasi-Newton Algorithms (e.g. Cavuoti et al. 2012), and support vector machines (e.g. Wadadekar 2005). In 2016, Beck et al. utilised machine learning methods to compute photometric redshifts using local linear regression for over 2 million galaxies in the SDSS Data Release 12 catalogue [89]. However, many of these algorithms are not publicly available and moreover, only provide redshift point estimates and not the elusive redshift PDFs [82]. The DELIGHT machine learning algorithm is a ML algorithm which avoids the need for a representative spectroscopic training sample by using a large collection of latent SED templates alongside a template SED library which is used as a guide for mapping the model. However, this algorithm parameterises all PDFs as a simple Gaussian, which may introduce much bias as the Gaussian parameterisation may not be able to capture the intricacy of the redshift posterior [90].

Supervised ML demands both spectroscopic redshifts and photometry as training data, so naturally the quality of the training data dictates the precision of the output. Galaxies with high redshifts or faint luminosities do not have as much accurate spectroscopic support, and so alternative methods are necessary to predict their photometric redshifts. Unsupervised ML doesn't require spectroscopic redshifts as a basis for training. Instead, it identifies similar objects by performing clustering in the input data space. Although these methods don't require representative training data, unsupervised techniques must be further developed before producing competitively accurate photometric estimates. Both supervised and unsupervised photometric redshift ML methods can produce PDFs, with varying levels of sophistication. The most basic method is performed by sampling randomly from values and their associated errors for each input parameter, and associating the normalised distribution of predictions as a PDF. Many physical photometric methods only produce redshift point estimates, however PDFs have become increasingly important in the complete characterisation of redshift uncertainties [87].

Data augmentation, feature importance and anomaly detection techniques allow ML methods to identify problematic data, allowing the algorithm to extrapolate further than when just provided with a training sample [88]. ML may also consider additional

information in its calculations, such as a prior (eg, reweighting the training data) which will further improve on photometric redshift estimates.

2.1.2 Random Forests

Random forest algorithms are a supervised ML technique made up of an ensemble of decision trees to make a prediction. The trees make a series of decisions based upon yes/no questions, splitting up data into two groups at each step until some defined threshold is reached. The first RF algorithm was developed by Tin Kam Hi in 1995. In the following decades the algorithm has been advanced by Adele Cutler and Leo Breiman, who went on to trademark the term 'Random Forests' in 2006 [91, 92]. Random forest techniques, much like the one implemented in this work, were first used to predict photometric redshifts in 2008 by Carliles et al. They used a sub-sample of the SDSS Data Release 6 to predict redshifts using a regression tree package in *R*. This work provided evidence that the RF technique is more than suitable for photometric redshift estimations, as their results were comparable to many other ML methods. This was then extended by Kind and Brunner (2013) [82], and then Sunil Mucesh (2021) [93], who developed RF algorithms which compute photometric redshift PDFs and point estimates using decision trees. A simplified diagram which visually represents the basic process of an RF algorithm can be seen in Figure 2.3.

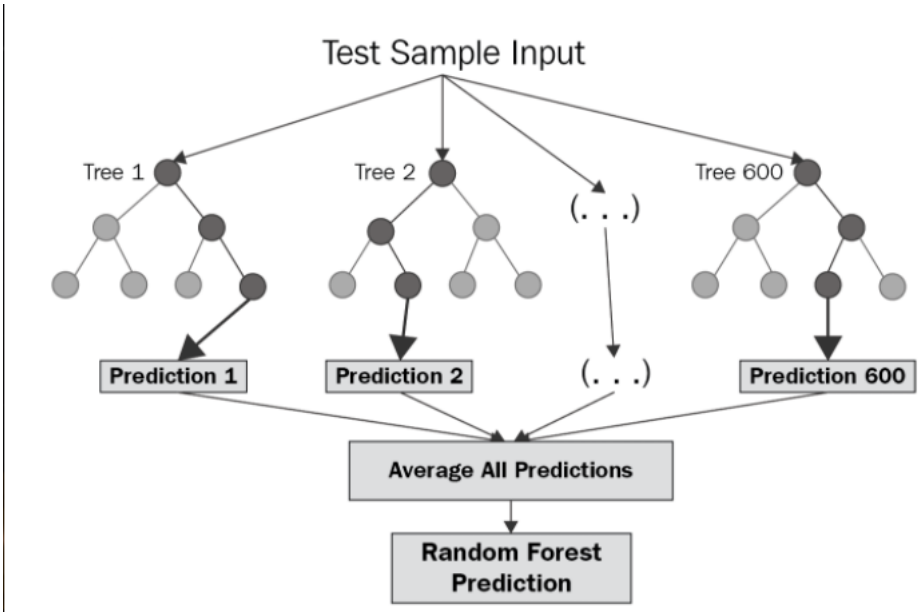


FIGURE 2.3: A demonstration of a simple random forest algorithm. Figure credit: [5].

The decision tree is made up of the root node, which determines the first split, a series of decision nodes which define the following splits, and the leaf nodes which describes the final groups. Each split groups the data with increasingly similar properties, with the leaf node then containing a small subsample of the data with very similar properties. At each step in the Random Forest process, all possible splits are assessed in all dimensions of the input feature space. Data is divided in such a manner that the average values of the target variable are representative of the groups [82].

The decision tree can then be built and the leaf node will contain new data which gives a prediction in the form of the mean of this node. The decision tree method is popular due to its simplicity, however it is susceptible to overfitting, as its performance is based solely on the training data provided. The RF methodology accounts for this by combining multiple decision trees and making adjustments. An example of these adjustments would be feature bagging, which introduces randomness by using only a subset of the training data and features when building the decision trees. Feature bagging is beneficial in that it allows the RFs to be better suited when making predictions on data it has not previously encountered. A trade-off between minimal variance and a low bias is achieved by this combination of decision trees and feature bagging [88].

The process of building an RF is as follows:

1. Randomly sample from training data with replacement to generate bootstrapped data.
2. Construct a decision tree by selecting a random subset of input features using the bootstrapped data.
3. The process is repeated to create multiple decision trees, which make up the 'forest'.

The mean of all the values predicted by the trees are then collated. The RFs require parameters to be set before the ML process may begin, which are termed hyperparameters. The hyperparameters may be tuned to optimise the RFs performance and are defined as [93]:

- n_estimators - The RFs effectiveness is determined by the number of decision trees used to construct it, with each tree being built using a subset of training data. If

the number of trees used is too little, it is likely that the training data will not give complete coverage. However, an increase in number of trees comes at a cost of computational time. A successful trade-off must be reached between the training time and performance in order to optimise the RF.

- max_features - The flexibility of the RF is determined by the correlation between the individual decision trees, which in turn is controlled by the maximum number of features considered at each step when building the trees. The maximum number of features that is generally sufficient to build each tree is $\sqrt{N}$, with N being the total number of input features.

- max_depth - The maximum depth parameter describes the number of levels in the decision tree. This dictates how coarsely or finely the training data are grouped. A depth which is too low or too high may lead to under or over-fitting respectively. It essentially acts as a stopping criterion, as it essentially represents the depth of each tree in the forest and determines how many times each tree is allowed to split.

2.1.2.1 Introducing spatial information

Redshifts may also be estimated using the correlation of source positions, as galaxies tend to reside in large scale structures and are therefore not randomly distributed in space. The spectroscopic redshift of a reference sample may be used to determine the redshift distribution of an unknown sample by maximising the spatial cross-correlation signal between the two. This method is mainly used in weak lensing applications to estimate mean photometric redshifts of selected samples, which requires the unknown sample to be preselected in a narrow redshift range. If the sample of preselected galaxies is narrow enough, individual redshift measurements may even be possible [87].

Another photometric redshift measurement method uses the cosmic web to derive the estimated redshift from the product of the density field, cosmic web and colours. This method calls for an accurate characterisation of void regions, clusters and filaments throughout the sky area of interest as well as accurate photometric redshifts (better than 1%) from colours. This is required in order to avoid the association between photometric redshift and the wrong cosmic web structure. Redshift estimates with an accuracy of $10^{-5} - 10^{-4}$ were possible for 3% of the SDSS multicolour catalog [83].

The decision to use ML or template fitting to measure photometric redshifts highly depends on the scientific application and the quality of spectroscopic redshift samples available for training. ML is favoured in cases where there is sufficient spectroscopic redshift support whereas template fitting code is useful when, for example, studying galaxy populations with limited spectroscopic coverage. When accounting for biases, template-fitting methods may also be used to model uncertainties in the absolute calibration [88]. However ML algorithms are insensitive to photometry biases depending on magnitude or colours, meaning they're favourable when limiting biases over a large area that is well supported by spectroscopic data. ML is also favourable in the fact that it is computationally much faster than template-fitting and can easily accommodate a much greater volume of data [87]. Many template fitting and spatial techniques only provide point estimates of redshifts, so for the purposes of this work it seems that ML methods are most appropriate to generate redshift PDFs for cosmological inference. Supervised ML methods are preferential as they tend to have much greater accuracy, however the level of representation of training data when using supervised ML may become an issue.

2.1.2.2 State of the art

The accuracy of the photometric redshift result is determined by the type of the redshift code. However, it is also independently affected by the quality of the input data and wavelength coverage.

The CANDLES survey provides some of the deepest photometric redshift samples available with the precision, σ_{z_p} (See Section 3.1.2 for thorough definition) increasing from 8% to 28% between H < 24 and $26 <$ H < 28. Medium band data with filter widths of 400 Åhave provided a breakthrough in precision in the last 10 years by improving SED resolution [94]. The first survey which produced a photometric redshift catalogue using medium band imaging was COMBO-17, which gave a bright source σ_{z_p} of 0.02 for bright sources. The ALHAMBRA, COSMOS and SHARDS surveys further improved this precision to 0.01 out to redshifts of $z \approx 1.5$ by using deeper medium-band photometry and utilising emission bands into their templates [83].

Photometric redshift studies that extend beyond the *redshift desert*, following the Lyman break ($z > 8$) and the Balmer break ($1.6 < z < 4$), have been made possible by the increase in sensitivity of near-infrared detectors. The identification of galaxies across

thousands of deg^2 has been made possible using photometric redshifts, with the DES survey reaching out to $z \approx 24$ over 5000 deg^2. This survey has an expected photometric redshift accuracy of 0.08 [83].

Conducting photometric redshift surveys which cover a large sky area presents many challenges, including acquiring spectroscopic training samples which are homogeneously distributed over the whole sky area, or the calibration of photometric noise precision over large observational periods with varying sky conditions. The potential degradation of instrument quality may also affect the calibration of photometric noise. The way in which the photometry is actually extracted from an image may also pose difficulties. The SExtractor is commonly used in source extraction, however one must be careful to ensure that the region within which the galaxy flux is measured is small enough that the SNR of faint sources isn't compromised [82].

2.1.2.3 The Future of Redshift Techniques

The quality and availability of spectroscopic surveys greatly impacts the evolution of photometric redshift methods. The future is bright for the next generation of spectroscopic surveys, such as the HSC imaging survey which will use the multi-object Prime Focus Spectrograph to gather redshift data over a wavelength range of 3800-13000 $\mathring{A}$ for millions of galaxies at $0.8 < z < 6$. The Euclid survey will be published alongside spectroscopy performed by the NISP instrument, which hopes to conduct slitless spectroscopy for over 50 million galaxies at $0.7 < z < 2$ [83].

A new generation of imaging surveys dedicated to cosmological study will be published over the next decade. The DES Survey covers 5000 deg^2 in 5 bands from 300 million sources, which is the largest galaxy survey currently available [86], and this number of galaxies is on track to increase by a factor of 10 in the next generation. The LSST, which is due to begin operating in 2023 plans to cover 18000 deg^2 of the sky and record data from 4 billion sources over the next 10 years [83]. The increase in volume of these imaging surveys does pose a computational problem when determining photometric redshifts, as it will increase the demand on computational power needed for calculations and the storage of the PDFs will become difficult over such a large sky area. For example, the LSST survey requires band-to-band calibration errors less then 0.005 mag, with the

variation across the sky being no more than 0.01 mag. It is essential that this calibration is maintained to ensure homogeneous performance, however much computational power is needed to compute photometric redshifts and store the PDFs to the required quality. The LSST also hopes to repeat observations of each sky location over 10 years in up to 6 bands, which would make the variable Universe more accessible, with measurements in the a band breaking degeneracy between high and low redshift solutions. Multi wavelength coverage is vital when defining the redshift range of interest so further advancements in photometric redshift will be driven by the Euclid survey, which will use three filters between 9200-20000 $\mathring{A}$. These filters should ensure precise photometric redshifts for $z > 1.3$, however the survey will require backing from ground based data in optical wavelengths. The PAU and J-PAS surveys are the first two cosmological imaging surveys to be performed using medium bands. J-PAS plans to cover 8500 deg^2 with 54 narrow band filters, which will study 300 million galaxies with a photometric redshift error of 0.3%. The James Web Space Telescope was launched in 2021, and contains an efficient near-infrared NIRCAM camera, which will lead to a new burst of activity with surveys being conducted up to mag_{AB} $30 - 31$ and produce photometric redshifts up to $z \approx 20$ [87]. .

This advancement in spectroscopic surveys, alongside steps forward in ML and template fitting techniques will provide a huge leap forward in future photometric redshift measurements. In the context of measuring cosmological parameters using GW data, this evolution of redshift techniques will decrease associated measurement uncertainties and contribute to revealing the solution to the Hubble tension issue.

The next chapters will explore whether these ML algorithms can be applied to galaxy surveys, and whether the RF may be trained on a sample from one galaxy catalogue and then applied to another catalogue which potentially may not be statistically equivalent. If it is possible to train the RF with a known spectroscopic sample and apply it to any general photometry data then this method would be a very useful tool in the generation of accurate photometric redshifts. Not only this, but the current assumption that the redshift uncertainties are Gaussian is accurate can be assessed using the RF algorithm. If this is not the case, then the availability of individual galaxy redshift posteriors is key in quantifying these errors, leading to more precise cosmological inference.

Chapter 3

GALPRO Calibration

3.1 GALPRO

Section 2.1 explored different photometric redshift estimation techniques and highlighted that supervised ML softwares appear to be the most appropriate option for this work, due to their accuracy and ability to produce redshift PDFs. In the following section the software GALPRO is selected as the supervised ML algorithm used for redshift PDF estimation and more insight is provided into how GALPRO works and how its performance may be assessed. This calibration and assessment of performance using known training and testing samples acts as a sanity check before GALPRO may be applied to new surveys and its limits are explored.

GALPRO is a Python software package, developed in 2020, which may be used to compute galaxy properties such as redshift, star formation rate, stellar mass, etc. It utilises a RF algorithm to form a supervised ML program [93]. RFs may be used for the regression or classification of large datasets, and here GALPRO is used for regression purposes to compute the photometric redshift of galaxies.

GALPRO has previously been used to compute photometric redshifts for the DESI Legacy survey [6]. The training sample provided by Zhou et al. (2021) used to generate these redshifts is the same as that used in this thesis, indicating that the package and training sample are trusted by the wider scientific community and further motivating its use in this work. Other motivation for this methodology includes the paper published in 2021 by Palmese et al. which uses GW signals from the first three LIGO/Virgo

observing runs to constrain H_0 to a value of $72.77^{+11.0}_{-7.55}$ km s^{-1} Mpc^{-1} [95]. Here, they used the galaxy catalogue method and populated the surveys with redshifts generated by GALPRO using the same training sample compiled by Zhou et al. as that used for the DESI Legacy survey. This further enforces that applying GALPRO using this training sample is applicable for computing redshifts to constrain H_0 using dark standard sirens. Since this Zhou et al. sample has been previously successfully used with GALPRO, it is an obvious choice to initially calibrate GALPRO and explore its functions, which is detailed in the following sections. More detailed information on the training data used is provided in Section 3.2.1.

3.1.1 GALPRO Random forests

In the case of estimating photometric redshifts, the RF algorithm determines a function which maps between the spectroscopic redshift values of the training sample and the multi-dimensional photometry space, ie. the fluxes and colours of each galaxy. It samples a random subset of the training data, which comprises fluxes, colours and spectroscopic redshifts, to build the decision trees. Every cell in the decision tree is then successively assigned an average spectroscopic redshift value. A photometric redshift estimate may then be made by inputting the photometric properties of a galaxy, in this case the flux and colours, and passing these properties through each tree. An average is then taken over all of the obtained redshifts to produce the final photometric redshift estimation of the individual galaxy [6].

GALPRO uses regression trees, which is a type of prediction tree that produces a continuous prediction, in this case, a photometric redshift PDF. The colours of galaxies are used as input variables to find the probability that an object may or may not lie in a specific redshift bin. Initially, the first node contains all of the data in the sample, which then splits recursively in such a way that maximises the information about the desired variable. The optimal dimension of these splits is determined by the minimisation of the sum of the squared colour errors. For node T, this is shown by [82]:

$$S(T) = \sum_{m \in values(M)} \sum_{i \in m} \left(z_i - \hat{z}_m\right) \tag{3.1}$$

where z_i are the target variable values, in this case redshift, $\hat{z}_m$ is the prediction model used, and m are the possible values of the dimension M. When using the arithmetic mean, $\hat{z}_m = \frac{1}{n_m} \sum_{i \in m} z_i$, with n_m being the members on branch m, we are able to rewrite equation 3.1 as:

$$S(T) = \sum_{m \in values(M)} n_m V_m \tag{3.2}$$

with V_m being the variance of the estimator $\hat{z}_m$.

The point estimates and PDFs generated by GALPRO implement a similar method to Kind and Brunner (2013) [82], in which photometry is perturbed by summing the flux from each band to a random value from a Gaussian distribution. The standard deviation of this distribution is assigned using the photometric error. This only applies to the fluxes, as the morphological parameters do not have an associated error.

GALPRO builds 50 individual decision trees in the RF, using the methodology described above, with each tree returning a photometric redshift estimate for a galaxy [6]. The tree is built by splitting the nodes as it follows equation 3.2, leaving every terminal leaf with only a few sources to be used for prediction. At each node, all dimensions are scanned to result in a split which minimises $S(T)$, and the dimension, M, and thus the minimum result is chosen as the splitting direction. Once a threshold in S is reached, the splitting ceases and a result may be determined. These estimates are then perturbed 20 times, resulting in an RF which spans the whole spectroscopic range. The photometric redshift point estimate and its associated error is determined by the mean and standard deviation of the resulting (50×20) estimates from each tree. The set of estimates given by all the trees and perturbations are then converted to a given resolution and normalised to the total number of objects returned, resulting in the redshift PDF estimate of the galaxy [93].

GALPRO has already been used to successfully generate photometric redshift estimates that are accurate enough to be included in the DESI Legacy survey [6]. This provides reassuring confidence that GALPRO can give reliable results when trained using spectroscopic data from the same survey. However, this thesis hopes to determine whether GALPRO can be trained using a specific sample of training data and then applied to

a brand new survey which contains no spectroscopic data and still produce accurate redshift estimates.

3.1.2 Point Estimate Performance Assessment

It is vital to thoroughly characterise the photometric redshift performance for every catalogue of interest. This is usually done by comparing the photometric redshift point prediction (eg. mean or mode of a PDF, z_{phot}) and the spectroscopic redshift (z_{spec}). Note that any spectroscopic redshift values used in ML training or prior construction should be discarded from the validation sample.

Assessing photometric redshift point estimate performances usually entails measuring the following:

- Precision (σ_{NMAD}): This describes the 68^{th} percentile width of the bias distribution about the median, and is defined as either the standard deviation of $(z_{\text{phot}} - z_{\text{spec}})/(1 + z_{\text{spec}})$ or $1.48 \times \text{median}(|z_{\text{phot}} - z_{\text{spec}}|)/(1 + z_{\text{spec}})$.
- Bias: This describes the average separation between the true and predicted redshifts and is defined as $\langle z_{\text{phot}} - z_{\text{spec}} \rangle$
- Outliers fraction: Gives the fraction of anomalous sources with unexpectedly large error values, defined by $|z_{\text{phot}} - z_{\text{spec}}|/(1 + z_{\text{spec}}) > 0.15$

These measures are not unique, but they are commonly used to allow for easy comparison between galaxy surveys. The direct comparison between photometric and spectroscopic redshifts as a way to assess photometric redshift performance is less informative if the spectroscopic redshift coverage isn't fully representative of the redshift coverage of the photometric samples. In this case, one must define a statistical mapping between the photometric parameter space covered by the two samples. If a region of photometric space doesn't have great spectroscopic support, alternative methods may be utilised to determine photometric redshift performances. The galaxy closed pairs technique is an example of this, in which the fact that neighbouring galaxies have a high probability of being associated and therefore having similar redshifts is used[93].

3.1.3 PDF Performance Assessment

Determining the full photometric redshift PDF performance often highlights the need for more training samples or templates, or sub-optimal redshift prediction routines. The under-estimation of photometric uncertainties is a common problem when dealing with PDFs, as this causes the PDF peak to be too narrow. Usually, PDFs are validated by ensuring that 1% of the spectroscopic redshift lies within the first percentile of their CDFs and so on. [90].

GALPRO is designed such that it generates joint posterior distributions by default (See Section 3.2.3.2) and it must be specified in the code that there is only one target variable, the photometric redshift. Most literature detailing the use of GALPRO, including from the creator of the software (See [93], [95]) refers to the PDFs of a single target variable, as opposed to joint posteriors, as 'marginal' PDFs and for continuity this work will do the same.

The probability integral transform (PIT) is an extremely useful tool for validation, as it indicates if any bias has been introduced to the PDFs. The PIT is used to assess probabilistic calibration and is defined as the cumulative distribution function (CDF), evaluated at the true redshift:

$$PIT = \int_{-\infty}^{\tilde{y}} f(y)dy \tag{3.3}$$

Here, $f(y)$ represents the redshift PDF and $\tilde{y}$ is the 'true' redshift value [93].

The spectroscopic redshifts should be random draws from their respective distributions, meaning that the CDF of the estimated photometric redshift, evaluated at the spectroscopic redshift, would not have a preferred value but will be uniformly distributed between 0 and 1. The PIT plot demonstrates how uniform the CDF accumulation is compared to an expected uniform PIT from 0 to 1. If the PIT values are from a uniform distribution then percentile of the CDF, 2% of the galaxies will have their 'true' redshifts found within the second percentile of their CDF, and so on. If the PIT is not uniform, then the 'true' redshifts are not random draws from their respective distributions, indicating some bias present in the redshift PDFs [6]. The shape of the PIT contains valuable information regarding the marginal PDFs. The PIT distribution will display a

convex shape if the marginal PDFs are too broad, as less objects will have true redshifts within the tails of their PDF. Conversely, if the PIT is a concave shape, the PDFs are overly narrow and many objects contain the true redshift within the tail of the PDF.

The PIT distribution must be uniform in order for the marginal PDFs to be valid, although a uniform PIT may also contain some bias [96]. This calls for another method of validation to ensure that no bias is present in the PDFs.

The equality of the true and predicted redshift distributions may be assessed using marginal calibration. This entails comparing the true empirical CDF ($\tilde{G}_I$) with the average predictive CDF ($\hat{F}_I$):

$$\hat{F}_I(y) = \frac{1}{n}\sum_{i=1}^{n} F_i(y) \tag{3.4}$$

$$\tilde{G}_I(y) = \frac{1}{n} 1\{\tilde{y}_i \leq y\} \tag{3.5}$$

Where n represents the number of test galaxies, F_i is the predicted CDF and $\tilde{y}_i$ is the galaxy's true redshift. The 1 is an indicator function which may be defined as:

$$1\{\tilde{y}_i \leq y\} = \begin{cases} 1 & \text{if True} \\ 0 & \text{if False} \end{cases} \tag{3.6}$$

To accurately describe the redshift PDFs as 'marginally calibrated', the average predictive and true empirical PDFs must be completely or almost equal. The marginal calibration plots shown in future sections of this work represent the difference between the average predictive and true empirical PDFs at regular intervals. If this variation is greater than 0.01 then the PDFs are considered not marginally calibrated.

Quantile-quantile (Q-Q) plots may be used to visually assess the uniformity of the PIT distributions, which highlights any deviations. The quantiles of the PIT plot and a uniform standard distribution (U(0,1)) are plotted against one another to demonstrate any deviation between the PIT and a uniform distribution. If the quantiles match perfectly, then the two distributions are identical and lie along the diagonal. GALPRO is able to produce PIT plots which also show the quantile deviation. The plot also shows

the value of each metric test (defined in Section 3.1.3.1) and percentage of catastrophic outliers. In this case, a catastrophic outlier is defined as a galaxy for which the true spectroscopic redshift is completely outside the support of its marginal PDF [83].

3.1.3.1 Uniformity of the Probability Integral Transform

There are many methods of qualitatively measuring the uniformity of the PIT, such as the Kullback–Leibler divergence, Kolmogorov–Smirnov test and Cramér-von Mises tests [93]. These metric tests all determine the similarity between the PIT and U(0,1) in slightly different ways and are very useful in quantifying how much the posteriors deviate from a Gaussian distribution.

The KL divergence is defined as:

$$KL \equiv \int_{-\infty}^{\infty} p(x) \log \frac{\mathrm{p(x)}}{\mathrm{q(x)}} \mathrm{dx} \tag{3.7}$$

where $p(x)$ is the the reference U(0,1) and $q(x)$ is the target PIT.[97].

The KS test determines the maximum distance between the CDF, $F(x)$, of the reference distribution and the empirical distribution function, $F_n(x)$, which is defined as:

$$KS \equiv sup_x |F_n(x) - F(x)| \tag{3.8}$$

with sup_x being the supremum of the set of distances. This test takes into account any random noise present in the PDF [98]. The closer the KS test value is to zero, the more uniform the PDFs from 0 to 1.

The CvM test is much like the KS test, however it is more sensitive to the edges of the distribution and is defined as [99]:

$$CvM \equiv \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 dF(x) \tag{3.9}$$

Where $F(x)$ represents the uniform reference distribution U[0,1].

All of these tests result in a value of zero if there is a perfect match between the two distributions. The larger the absolute value of each test, the larger the deviation of the PIT from a uniform distribution.

The marginal calibration may also be assessed graphically by plotting the difference between true empirical and average predictive redshift CDFs. If the marginal calibration is successful, there should be only minor fluctuations about the zero line [6]. Using the PIT, the individual galaxy PDFs may be probabilistically and marginally calibrated to ensure that the PDFs are valid and may then be used for cosmological inference.

However, it is imperative at this point to acknowledge that these metrics are far from perfect when assessing individual redshift PDFs. In the absence of true redshift posterior samples, it is obviously difficult to establish metrics which evaluate the PDF performance, which remains an ongoing issue in the scientific community [90].

The redshift posterior distributions generated by GALPRO could potentially provide much more detailed information than previously used by the `gwcosmo` code, which would lead to better constraints on H_0.

3.2 Results

3.2.1 Redshift Truth Table

GALPRO requires a training dataset, describing the photometry of the galaxies and their spectroscopic redshifts, and also a 'truth' dataset to assess calibration. The truth dataset contains the flux and colours of the galaxies, which is used by GALPRO to predict their photometric redshifts and generate redshift PDFs for each galaxy in the subset. Once all of the objects in the truth dataset have been assigned a photometric redshift point estimate, the spectroscopic redshift of each truth galaxy can be compared to the photometric values to assess the performance of the RF mapping.

GALPRO firstly requires a 'truth table' of detailed and reliable spectroscopic and photometric redshifts for training and validation purposes.

The sample used to train the RF algorithm was compiled by Zhou et al. [6], which includes data from the 2dFLenS, DEEP2, COSMOS15, AGES, GAMADR3, Oz-DES,

Survey	Full Dataset	Downsampled dataset
BOSS	678 370	224 345
SDSS	449 386	186 666
WiggleZ	122 907	47 334
GAMA	109 790	55 990
COSMOS2015	53 973	53972
VIPERS	44 175	44 175
eBOSS	23 549	23 459
DEEP2	15 994	15 994
AGES	11 235	11 235
2dFLenS	8102	8102
VVDS	5490	5490
OzDES	1407	1407

TABLE 3.1: The number of objects from each survey used in the truth dataset [6].

SDSS DR14, VIPERS, VVDS and WiggleZ surveys [93]. The RF takes input features of r-band magnitude, $r-z$, $g-r$, z - $W1$ and $W1-W2$ colours. It also requires the ratio between the semi-minor and semi-major axes, the half light radius, and a model weight, which assesses how well a galaxy is fit by an exponential light profile versus a de Vaucouleurs profile [95]. Soo et al (2017) found that including these three morphological parameters significantly reduces the outlier fraction and scatter of the test sample when training a RF algorithm on grz photometry. Table 3.1 shows the number of objects from each survey in the dataset. All of the imaging catalogues used overlap with the DECaLS catalogue footprint and already have K corrections applied [100], [6].

The majority of the galaxies in the truth dataset are taken from the SDSS, BOSS, GAMA and WiggleZ surveys, which are limited to shallow magnitudes and apply colour selections. This causes sharp peaks in the redshift distribution which can introduce systematic bias as a result of the non-uniform training sample, as the RF algorithm may favour redshift estimates that are over-represented in the truth dataset. To avoid this, galaxies from the four surveys are downsampled to create a more uniform training dataset. This is done by firstly determining the object density in the 2D r-band magnitude space versus redshift, using bin sizes of $\Delta r_{mag} = 0.01$ and $\Delta Z = 0.01$. Each survey is assigned a different threshold for the target number of galaxies contained within the r_{mag}-redshift bins, which are 20, 70, 400 and 400 for the WiggleZ, GAMA, BOSS and SDSS catalogues respectively. To preserve an accurate sampling of galaxies over the full

luminosity range, the objects are downsampled randomly until the r_{mag}-redshift density reaches the threshold level. This method also ensures that the rare most luminous objects are preserved in the sample. This downsampled dataset will now be referred to as the truth catalogue. A plot of the redshift distribution of the full truth dataset and the downsampled truth dataset prepared by Zhou et al. can be seen in Figure 3.1. The truth table sample contains galaxies out to a maximum redshift of 4.29, however the number of galaxies with redshifts above 1.5 is very small.

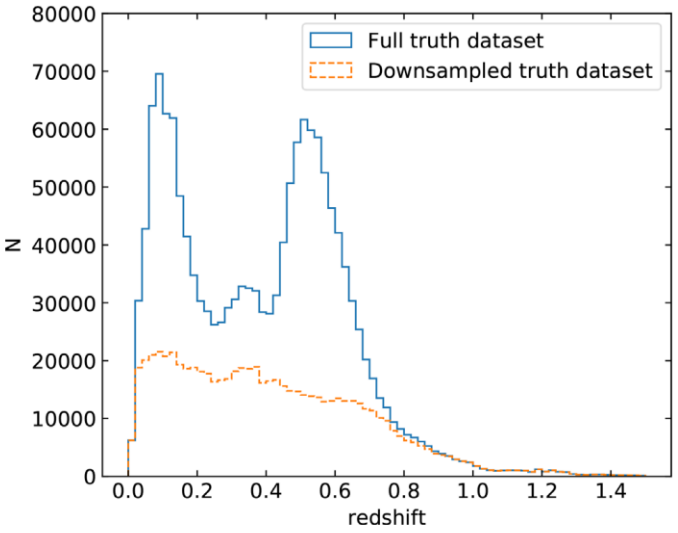


FIGURE 3.1: Redshift distribution of the redshift truth data set. N(z) gives the total number of objects in each z = 0.02 bin. The SDSS and BOSS surveys contribute to the sharp peaks at z = 0.1 and z = 0.5 respectively. These peaks are downsampled to avoid bias. Figure credit: [6]

The truth table contains some fainter galaxies with fluxes in the grz bands being negative, which causes their magnitudes to be undefined. This is accounted for by converting the fluxes in luptitudes (μ), using the formula:

$$\mu = \mu_0 - a \sinh^{-1}\left(\frac{f}{2b}\right) \tag{3.10}$$

with $\mu_0 = m_0 - 2.5 \log b$. $a = 2.5 \log e$, where m_0 is the zero point magnitude, f is the flux and b is an arbitrary softening parameter [93]. The associated errors in each flux, σ_{mu} are also converted as follows:

$$\sigma_\mu = a\left(\frac{\sigma_f}{f}\right) \tag{3.11}$$

with σ_f being the error given on the flux. Converting the objects' flux to luptitudes removes the need to discard galaxies with a negative flux, thus avoiding introducing another selection effect.

To increase the performance of the algorithm, the RF hyperparameters (discussed in Section 2.1.2) may be tuned. Following results described in [93], which used the same truth sample as this text, the optimal hyperparameters are found to be:

- `n_estimators: 100`
- `max_features: auto`
- `max_depth: none`
- `min_samples_leaf: 1`
- `min_samples_split: 2`
- `max_leaf_nodes: none`
- `min_impurity_decrease: 0.0`
- `min_impurity_split: none _weight_fraction_leaf: 0.0`

This selection of hyperparameters allows the decision trees to be fully grown, meaning the training data may no longer be split. The choice of max_features being auto guarantees that the RF algorithm has a sufficient amount of prior information, which is essential as we are only using a small number of photometric bands.

3.2.2 Applying GALPRO using the Zhou et al. Dataset

This section uses the above described truth table to calibrate GALPRO and explore its functions, before it is applied to any new survey. Since the sample, compiled by Zhou et al, has been previously used in [93] and [6], it is an excellent choice for training and testing as it is known to perform satisfactorily.

The $g-r, r-z, z-W1, W1-W2$ luptitudes and their associated errors are used as input variables for `GALPRO`, alongside the previously mentioned half-light radius, axis ratio, model weight and the spectroscopic redshift. Once the Zhou et al. truth table

has been converted to contain these parameters it is then randomly split, with 90% of the data being used to train the RF algorithm, and 10% being used for testing and validation. `GALPRO` takes 4 input arrays, x_{train} and x_{test}, which contains the luptitudes and parameters of the training and testing samples, and y_{train} and y_{test}, containing the corresponding spectroscopic redshifts (the target variables). The hyperparameters are set to those described in Section 3.2.1.

The run-time of GALPRO is dependent on the size of the training and testing samples and here, the training sample contains around 2.7 million objects and the testing sample contains around 300,000 objects. Both the training and calibration of GALPRO using samples of this size requires a large amount of memory usage and initially GALPRO was implemented using the University of Glasgow WIAY cluster. However, since memory usage is restricted to a certain threshold on this cluster, GALPRO continuously failed to run as it required much more computational power than was available. This posed a large problem and an attempt to use a dataloader to input the training and testing arrays to GALPRO was undertaken. However, it seemed that it wasn't the size of the samples that was the issue, but the large dimensions of the arrays created by the RF while it is training that was using a large amount of memory. After over a month of deliberations and attempts to reduce memory usage, access to the LIGO Caltech cluster was granted. This cluster has 1.5 Terabytes of memory and 72 CPU, so any issues regarding memory storage is completely avoided. Training GALPRO does require a large amount of computational power, but once trained, the RF stores the model. This means that GALPRO only needs to be trained once and then this model is saved and can be applied to any testing dataset. The calibration process performed on the testing sample also requires a lot of computational power, and using a testing sample of around 300,000, the process takes roughly 6 hours. This is larger than the time required for training with a sample of around 3 million objects, which usually takes 3 hours. The full calibration process produces redshift point estimates, PDFs, the PIT, marginal calibration and spectroscopic versus photometric redshift plots. GALPRO can also produce kendall calibration plots (see [93]), however these are used to assess the calibration of the joint posterior distributions, which aren't particularly relevant to this work. The production and storage of the redshift PDFs and point estimates only takes a short while, around 20 minutes in total, while the marginal and probabilistic calibration takes up the majority of the 6 hour period. This long calibration time may

seem impractical, but it is worth noting that this calibration only needs to be run once using a representative subsample of the testing data to ensure accurate results. Then, GALPRO can be applied to the entire testing sample and ran in a way which only produces the redshift PDFs, thus reducing the run-time significantly. The entirety of this work was performed on the LIGO Caltech LDAS-pcdev6 cluster remotely.

3.2.2.1 Redshift Point Estimate Results

Analysis was run by inputting the above described arrays into GALPRO. The photometric redshift point estimates are generated for every galaxy in the testing sample, which is given by an average of all the predictions from all of the decision trees in the RF model. The scatter of the distribution is also used to quantify the photometric redshift estimation performance, which is represented using the σ_{NMAD} and describes the observed scatter between predicted photometric redshifts and their corresponding spectroscopic (i.e. true) values (see Section 3.1.2).

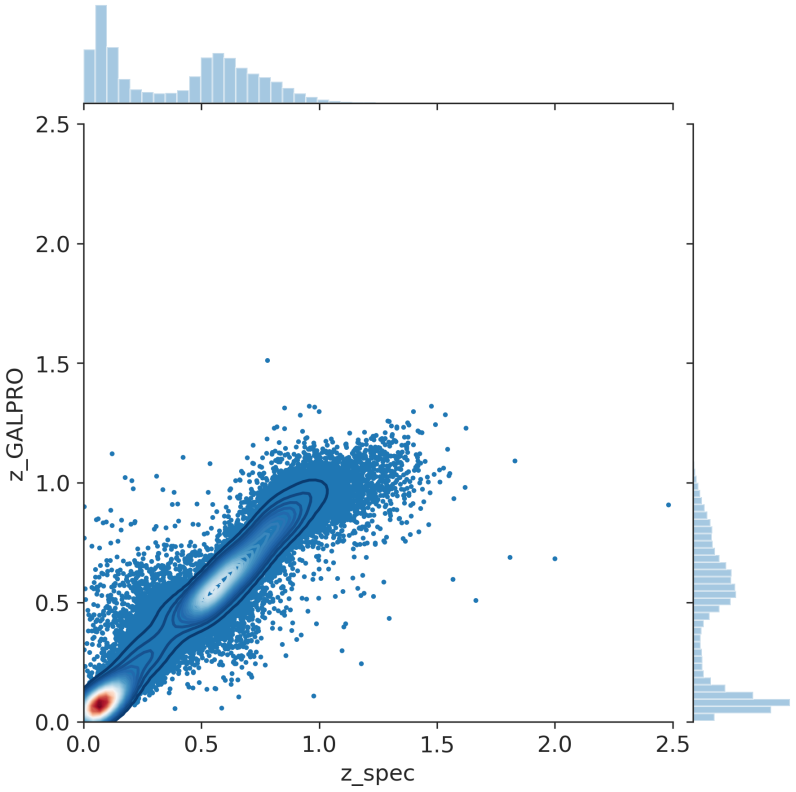


FIGURE 3.2: The spectroscopic versus photometric redshift point estimate plot produced when GALPRO is trained using a randomly sampled dataset containing 90% of the Zhou et al. dataset and tested using the other 10%.

The plot, shown in Figure 3.2, has a good correlation between spectroscopic and photometric values, and generally follows the diagonal. It gives a σ_{NMAD} value of 0.029, which is relatively low indicating that the scatter has a strong correlation. As we reach out to larger values of z, it can be seen that GALPRO tends to underestimate z values above $z_{\text{spec}} \geq 1.5$. Since the truth table doesn't contain many objects at these high redshifts, the mapping between the flux and these redshift values obviously isn't correctly learnt due to there being fewer representative objects at this range in the training dataset. This is to be expected, as the RF will not perform as well in areas that are lacking representative training data, highlighting the need for the training and testing samples to have a similar range of redshift distribution. The contours of the scatter plot represents the iso-proportions of the density, with 50% of the data-points lying withint the largest contour line. Overall, the photometric redshift point estimates are generally accurate, however the redshift posterior PDFs are of more interest to this thesis.

3.2.3 Posterior Distributions

3.2.3.1 Marginal Posteriors

The LIGO team, when implementing `gwcosmo`, currently assume that any photometric redshifts used to measure the Hubble constant have an associated error that is generally Gaussian. When determining whether this assumption is valid, a photometric redshift posterior is much more informative than an individual point estimate. The posterior shows the probability of the estimated redshift value, and so the shape of the posterior directly relates to the assumed shape of the error distribution.

GALPRO produces the marginal PDFs of the photometric redshifts for each galaxy in the testing sample. These PDFs are automatically stored by GALPRO in a specified file. An example of a redshift PDF, randomly selected from the testing sample is shown in Figure 3.3. The predicted photometric redshift value, given by the black dashed line, is the median of the PDF.

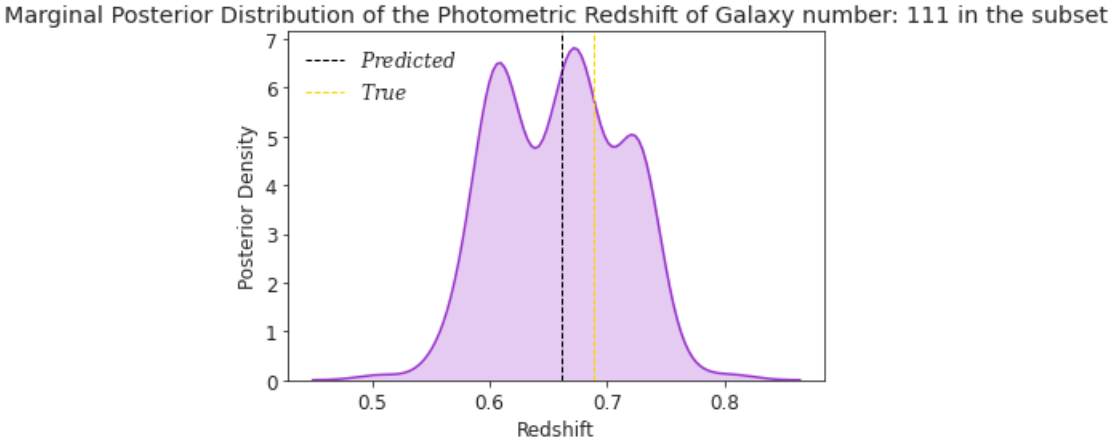


FIGURE 3.3: An example of a redshift PDF generated by GALPRO when trained and tested using the Zhou et al. truth sample.

The photometric redshift estimate predicted by GALPRO for this galaxy and the real, spectroscopic value can be seen on the plot. More detail on the 68th percentile region can be found later in Section 3.3. The RF generates PDF coordinate values and then uses kernel density estimation to plot the smooth PDF that can be seen in the figure. As previously stated, the computational cost of generating the PDF is actually not very high, which is beneficial when applying GALPRO to large photometry surveys.

3.2.3.2 Joint Posterior Distributions

The GALPRO package is also able to generate multivariate posterior distribution functions on-the-fly for its individual galaxy redshift estimates. This is implemented by using the RF algorithm to predict both the photometric redshift and rest frame r-band absolute magnitudes simultaneously. The redshifts and magnitudes may then be used to plot joint PDFs for each galaxy in the testing sample. This work focuses on the generation of photometric redshift PDFs, so the joint posterior distributions are not entirely relevant or useful for this thesis. However, the exploration of GALPRO's functions is interesting nonetheless and could be relevant to future work, hence why this small section is included.

The algorithm creates joint PDFs by creating a single model to predict both variables. The algorithm is much like that for an RF predicting a single variable, however now the average loss function (Equation 3.1) is minimised at each step. The $\tilde{z}_i$ and $\tilde{z}_m$ now

represent the vector of target variables and means. The scales of the target variables must first be transformed to be within similar ranges so that the variance of one variable does not dominate. The algorithm groups galaxies in an n-dimensional space according to the values of the input features and then minimises the loss function of these clusters. The leaf nodes then eventually contain the the r-band magnitudes and stellar masses of similar galaxies, which are determined simultaneously to preserve the correlation between the properties. The test galaxy properties may then be run through all of the decision trees to generate new photometric redshift and r-band magnitude estimates using the mean of all the predicted values.

The joint PDFs may be generated using GALPRO by simply adding the desired target feature to the y_{train} and y_{test} arrays. Here, the r-band magnitude was used as both an input in the x arrays and a target variable in the y arrays. The purpose of this is to both generate joint PDF plots and to assess how well GALPRO can predict properties, as having this magnitude in both the x and y arrays should lead to the GALPRO r-band magnitude predictions being almost identical to the true values. Any deviation from this would indicate that the RF algorithm has failed to learn the correct mapping.

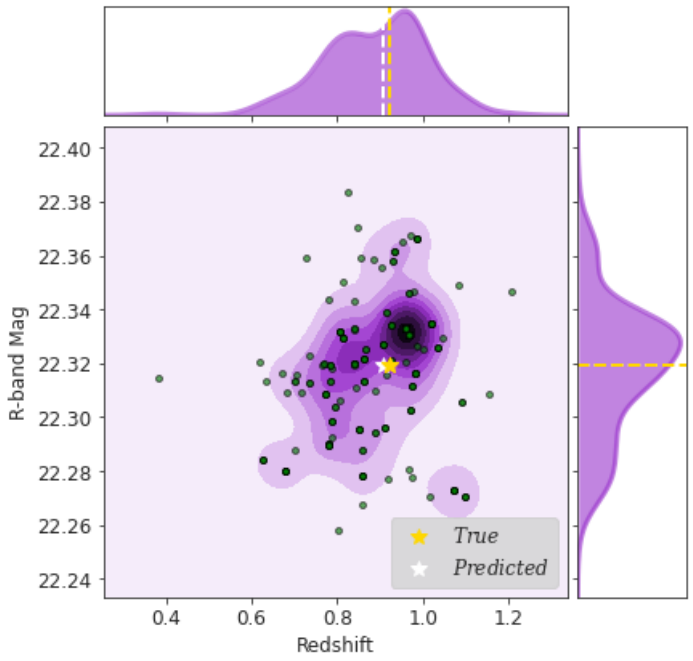


FIGURE 3.4: An example of a joint redshift and r-band magnitude PDF generated by GALPRO when trained and tested using the Zhou et al. truth sample.

An example of a joint PDF produced by GALPRO can be seen in Figure 3.4, which was randomly chosen from the testing sample. The marginal PDFs of the redshift and

r-band magnitude can be seen on the graph, alongside the true and predicted values of both variables. As expected, the predicted r-band magnitude value is identical to the real value, as this magnitude is both an input and a target variable. This reassures that GALPRO is performing correctly, as these values should be identical since it has already been fed the r-band magnitudes. The green circles represent the redshift and r-band magnitude values in the leaf nodes that are representative of the test galaxies. Kendal calibration is used to assess the performance of these PDFs.

The joint PDFs are not directly useful in the quest to generate redshift PDFs to be used in the constraint of H_0, and so no more detail will be discussed in this work. However, more detail of the production and validation of these plots can be found in [93]. Although not relevant to this work, these plots are rather aesthetically pleasing!

3.2.4 Marginal Calibration

The marginal calibration plot represents the difference between the average predictive and true empirical CDFs of redshift, at regular intervals. The marginal calibration plot produced when GALPRO is trained and tested using the Zhou et al. truth sample (as described above) can be seen in Figure 3.5. There are negligible fluctuations about the zero line with a peak around 0.0015. Any fluctuation greater than 0.01 indicates that PDFs are not marginally calibrated. This demonstrates that the PDFs generated using this testing sample are successfully marginally calibrated.

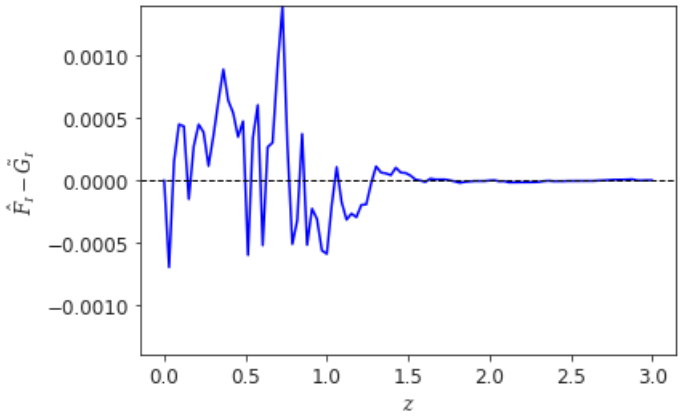


FIGURE 3.5: The marginal calibration plot generated by GALPRO when trained and tested using the Zhou et al. truth sample.

3.2.5 PIT

The validation of individual redshift PDFs is not possible due to the true distributions not being available, therefore the marginal PDFs must be validated as a whole. This validation is conducted by GALPRO using framework developed by [101]. The paper introduces three modes of calibration; exceedance, probabilistic and marginal calibration. These methods may be interpreted as characterising the statistical consistency between the distributions and the truth. The sharpness of the predictive distributions is defined as the concentration of predictive distributions, which is solely a property of the distributions, and may be maximised subject to calibration to validate the PDFs. This maximisation is the basis of the validation methods, however GALPRO focuses on the calibration to validate the PDFs rather than the sharpness. Section 3.1.1 has already detailed this assessment, which is used in this thesis to determine the accuracy of the generated PDFs.

The PIT produced by GALPRO when trained and tested using 90% and 10% of the Zhou et al. truth sample respectively can be seen in Figure 3.6.

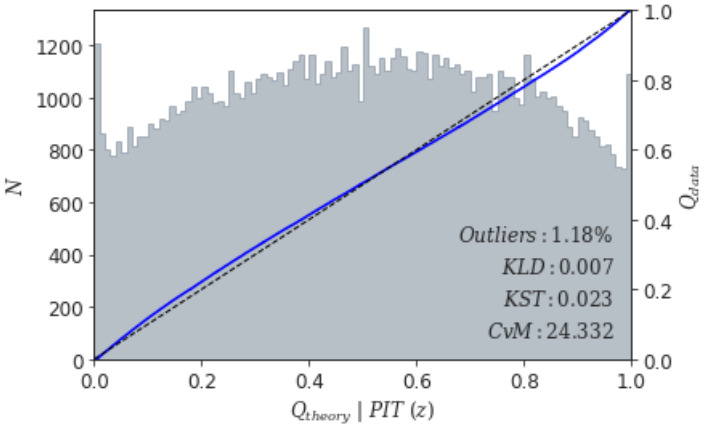


FIGURE 3.6: The probability integral transform produced by GALPRO using the Zhou et al. truth sample for training and testing.

The outlier fraction is given as 1.18% which is very reasonable, and the KLD and KST tests give values of 0.007 and 0.023 respectively. These values are close to zero, showing that the PIT is fairly uniform. The Q-Q plot represents the quantiles of a uniform distribution using the black dashed diagonal line, and the quantiles of the results are given as the blue line. The blue line doesn't deviate greatly from the black dashed line,

further demonstrating that the PIT is fairly uniform. However the CvM test gives a value of 24.332, which is reflective of the fact that the PIT is slightly convex.

3.2.5.1 Improving probabilistic calibration

Although the PIT is fairly uniform, this slight convex shape indicates that the marginal PDFs are somewhat overly broad. To improve the probabilistic calibration of the PDFs, the minimum number of samples that must be present in a leaf node (`min_samples_leaf`) is increased from 1 to 3 and 5. Figures 3.7, 3.8 and 3.9 show the results produced by GALPRO with `min_samples_leaf` = 3. It was found that increasing the minimum number of samples in the leaf node to 3 improved results slightly, with the outlier fraction reducing to 0.51% and the CvM test reducing to 21.055. The σ_{NMAD} of the scatter plot also reduces by 0.001. When min_leaf_ is increased to 5 however, the results deteriorate and the PIT becomes less uniform (See Appendix A.1). For the rest of this thesis, the `min_samples_leaf` hyperparameter will be set to 3.

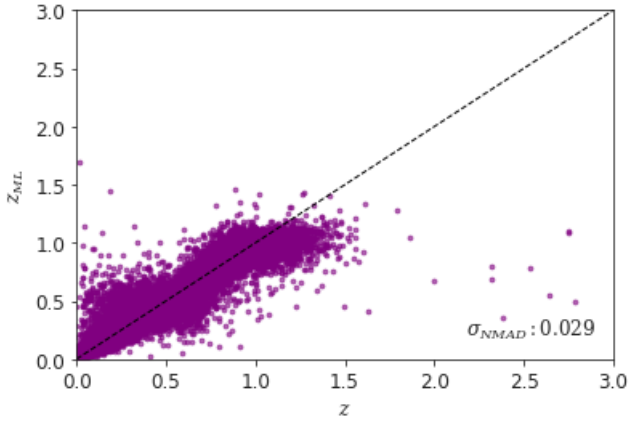


FIGURE 3.7: The spectroscopic versus photometric redshift plot produced by GALPRO when trained and tested using the Zhou et al. truth dataset with `min_samples_leaf` = 3.

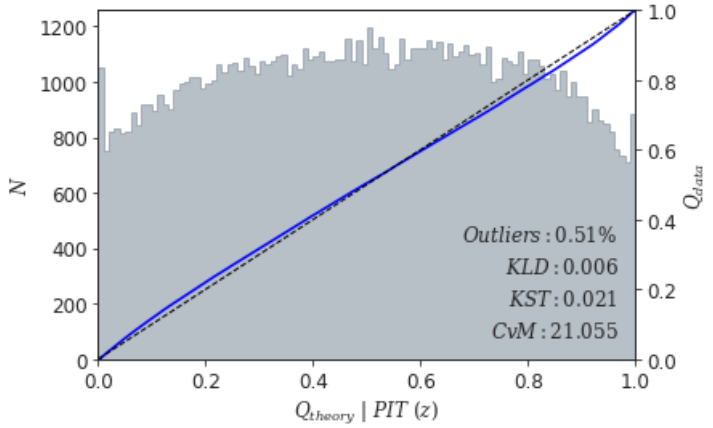


FIGURE 3.8: The PIT plot produced by GALPRO when trained and tested using the Zhou et al. truth dataset with `min_samples_leaf` = 3.

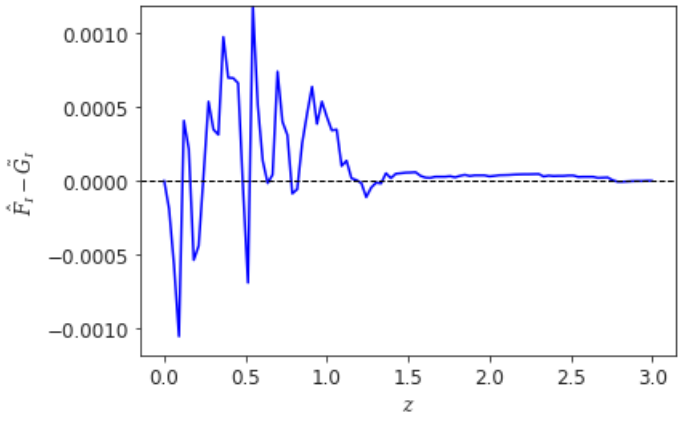


FIGURE 3.9: The marginal calibration plot produced by GALPRO when trained and tested using the Zhou et al. truth dataset with `min_samples_leaf` = 3.

Increasing the `min_samples_leaf` was successful in the attempt to slightly improve probabilistic calibration, but there is a second option which may further improve the uniformity of the PIT. This involves augmenting the training data by scattering the galaxies. The scattering is performed on the photometry in accordance with the photometric errors. Each galaxy has its magnitudes scattered by adding on a randomly selected value from a Gaussian distribution with a mean of 0 and a standard deviation of 1, multiplied by the photometric error, following:

$$S = M + (Z \times \sigma) \tag{3.12}$$

Where S is the scattered magnitude, M is the observed magnitude, σ is the associated photometric error and Z is the value randomly selected from the Gaussian distribution.

These magnitudes are scattered in each filter and then the scattered colours are computed. The photometric errors associated with these scattered galaxies remain the same and are not scattered.

This scattering process essentially creates mock "observed" magnitudes, meaning the value we might have observed in a parallel universe, or more mundanely if we'd simply observed the galaxy on a different date so that a different, random number of photons arrived from the galaxy. Each galaxy in the training set is scattered five times and the testing dataset is scattered once for consistency. GALPRO was again run with a randomly selected 90% sample used for training and the other 10% for testing, The resulting PIT plot can be seen in Figure 3.10. This plot shows a small improvement in the PIT as the CvM value has slightly decreased and the outlier percentage has decreased by nearly 1%. The marginal calibration remains the same and the scatter is very similar to the previous analysis with no scattering of the photometry (See Appendix A.2).

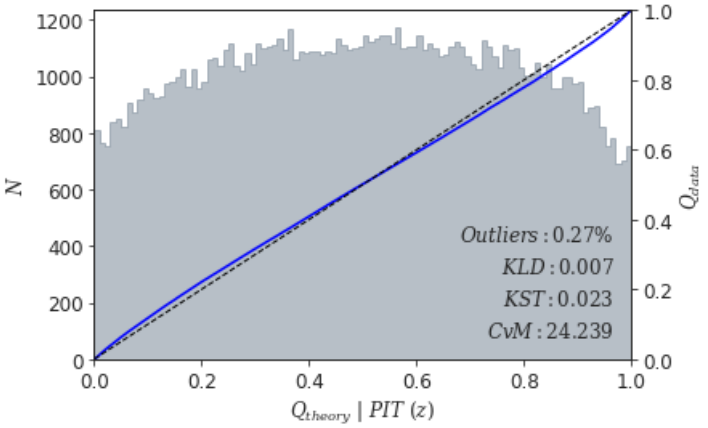


FIGURE 3.10: The PIT plot generated by GALPRO when trained and tested using the Zhou et al. truth sample where the photometry has been scattered.

To summarise, the marginal redshift PDFs produced by GALPRO when testing and training using the scattered Zhou et al. dataset are both probabilistically and marginally calibrated, giving confidence that they are valid. Henceforth, any reference to the use of the Zhou et al. dataset being used for testing or training will refer to the Zhou et al. truth dataset with min_leaf_sample set to 3, with the photometry scattered as described above and the hyperparameters described in Section 3.2.1, as this produces the optimal results with the most uniform PIT plot. The following section explores the errors associated with the photometric redshifts, and whether the current assumption in the application of `gwcosmo` that these errors are Gaussian is valid.

3.3 68^{th} Percentile Region

It is currently assumed by the LVK Collaboration that photometric redshift errors input to `gwcosmo` follow a normal distribution, which take a symmetric and unimodal shape. If X is a random, normal variable then the probability distribution of X may be described by a Gaussian curve as:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \tag{3.13}$$

with μ being the mean and σ as the standard deviation from the mean. The 1σ confidence interval encapsulates the area of the graph for which approximately 68% of the population should fall within. This may be mathematically expressed as:

$$F(\mu - 1\sigma \leq X \leq \mu + 1\sigma) \approx 68.27\% \tag{3.14}$$

For a normal distribution, the probability function and therefore the 68^{th} credible region is symmetric. It has been assumed thus far in most gravitational wave analyses that the photometric redshift error is indeed Gaussian and so has a symmetric 68^{th} confidence region. How the code `gwcosmo` utilises these photometric errors is described in Section 1.2.

The analysis performed by GALPRO using the Zhou et al. scattered training and testing samples, described in the above sections, has the ability to produce individual photometric redshift PDFs for each galaxy in the testing sample. This is very useful, as it allows for the assessment of the 68^{th} credible region of each redshift posterior to be evaluated. The quantiles are determined using NumPy package `percentile`, which takes a set of posteriors and returns an array containing the quantile values for each distribution. In this case, the photometric redshift posteriors were inputted to the code and the package returned the 16^{th} and 84^{th} quantile of each posterior, which is the bounds of the 68% credible region.

Figure 3.11 is a randomly selected example of a photometric redshift PDF chosen from the analysis run in Section 3.2.5.1. It can be clearly seen by inspection alone that the PDF is non-Gaussian and does not have symmetric 68^{th} percentile errors.

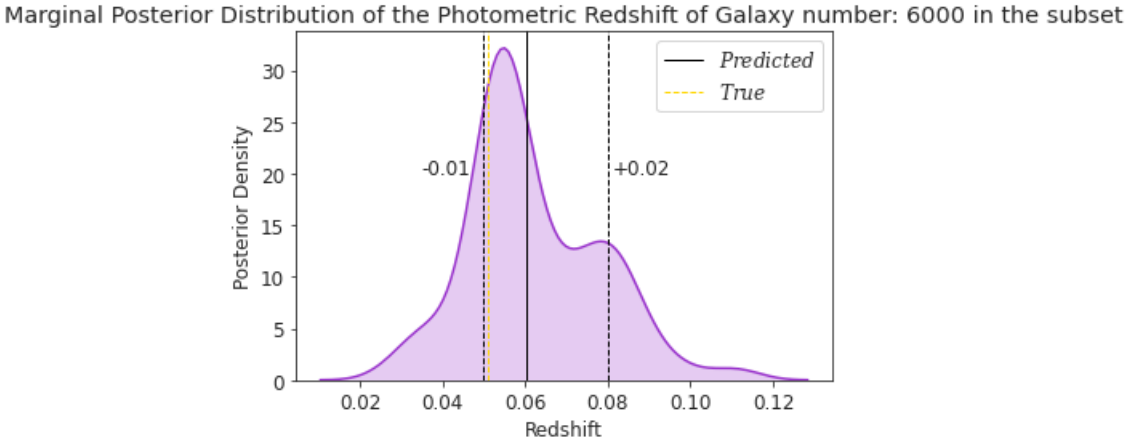


FIGURE 3.11: The photometric redshift PDF of a randomly selected galaxies from the Zhou et al. testing subsample.

A quantification of just how non-Gaussian the redshift PDFs are is required before making the assumption that the associated photometric errors generally do not follow a Gaussian distribution. This is done by applying the D'Agostino's K-squared test to each of the redshift posteriors in the testing sample. This test is designed to establish whether a function comes from a normally distributed population. When applied to the redshift PDFs generated from the Zhou et al. testing sample containing 101945 galaxies, it was found that 78241 galaxies had a non-Gaussian redshift PDF and 23704 were found to follow the Gaussian distribution. This demonstrates that the majority of the galaxies in the testing sample has redshift PDFs that did not follow a normal distribution.

The Zhou et al. truth sample is relatively representative of the redshift range and type of galaxies that would be analysed by the `gwcosmo` code, which are shown to have a generally non-Gaussian behaviour. This means that an inference of H_0 using this methodology may increase in accuracy if the redshift errors inputted to the program are precise, and not generalised to follow a normal distribution. This implies that GALPRO could be extremely useful in generating accurate photometric redshift PDFs that can be used to thoroughly characterise the associated errors that are inputted to `gwcosmo`. However, the accuracy and precision of GALPRO must be rigorously tested before it can be reliably used to generate these PDFs.

GALPRO has been successfully implemented using the Zhou et al. truth table to generate marginally and probabilistically calibrated redshift PDFs, which have been used

to probe the GALPRO functions and examine current LIGO assumptions. The following chapter explores whether GALPRO can be trained using the Zhou et al. truth dataset and applied to a new, unknown survey to generate accurate photometric PDFs. The PDFs generated by GALPRO may only be used alongside `gwcosmo` to measure the Hubble constant if the following analysis produces reliable results.

Chapter 4

Applying GALPRO to a New Survey

Chapter 3 demonstrated that `GALPRO` can be successfully trained using a dataset containing certain input parameters and then applied to a second dataset from the same survey to generate photometric redshift estimates and posteriors which can be marginally and probabilistically calibrated to provide reliable results. This is a useful sanity check, however it is hoped that `GALPRO` may now be applied to a completely new survey that doesn't necessarily contain spectroscopic data with which to directly calibrate the photometric redshifts, but that the photometric redshifts derived would still have equally reliable PDFs. The following chapter explores the application of GALPRO, trained using the previously described dataset, to a new survey and discusses the limitations or restrictions involved in this process. Essentially, therefore, this chapter investigates whether GALPRO is a suitable choice for the generation of redshift PDFs when applied to an unknown survey, and if these PDFs are reliable enough to be included in the inference of H_0 using gravitational wave data.

4.1 Applying GALPRO to the PanSTARRS Survey

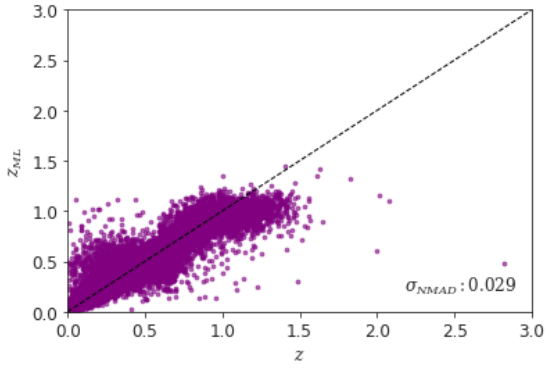
The aim of this thesis is determine if GALPRO can be reliably used to generate photometric redshifts when applied to a brand new galaxy catalogue with no spectroscopic calibrating data. From the fourth LIGO observing run and beyond, it is hoped to

add much more galaxy redshift data from various new catalogues, which can be used alongside new GW data to constrain H_0. However, these catalogues may contain few spectroscopic redshifts, so it is important to have an accurate method to generate photometric redshifts for the galaxies they contain.

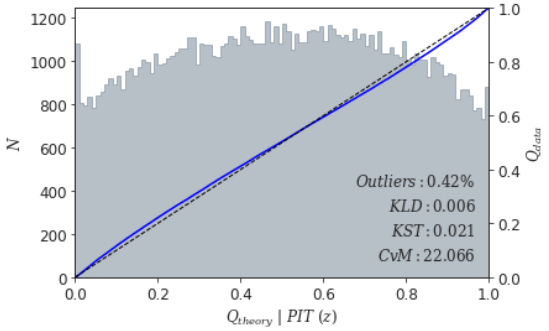
Here, the Panoramic Survey Telescope and Rapid Response System (PanSTARRS) survey is used to fill the role of a 'new' survey which hypothetically has no spectroscopic data. In fact, the PanSTARRS survey has been cross-matched to provide a large sample of spectroscopic redshifts which are used for validation [89], however it is used here as an example of an artificial survey which *doesn't* provide spectroscopic data, much like any catalogues that may become available in the future. The PanSTARRS telescope is located in Hawaii and accurately measures photometry of known objects by surveying for variable objects using telescopes, cameras and a large computer system. The latest PanSTARRS release is the largest volume of astronomical data ever published, and the ability to apply an RF algorithm to this very large catalogue would be advantageous for cosmological inference due to such a large number of objects gaining associated photometric redshift values.

The main aim of the PanSTARRS survey is to observe objects that are near to the Earth which may pose the threat of impact events, meaning the survey is relatively shallow. A sample of around 2 million galaxies included in the PanSTARRS survey have cross-matched spectroscopic redshifts, unlike the hypothetical new surveys, which is advantageous in that these spectroscopic redshifts can be used to assess the performance of the applied RF. In the previous chapter it was shown that the Desi Legacy sample compiled by Zhou et al. (2012) can be used to successfully train GALPRO and when tested with a sub-sample of its own data, can produce accurate photometric redshift estimates and PDFs. This makes the Zhou et al. sample an excellent choice of training data to determine whether or not the redshifts obtained by applying the trained RF to a new survey are accurate enough to be useful. Due to the PanSTARRS survey focusing on near-Earth objects, it is shallower than the Zhou et al. training sample, meaning the range of the redshifts and photometry bands differs slightly between the two catalogues. How this range difference affects the application of the RF algorithm will be explored in the following sections.

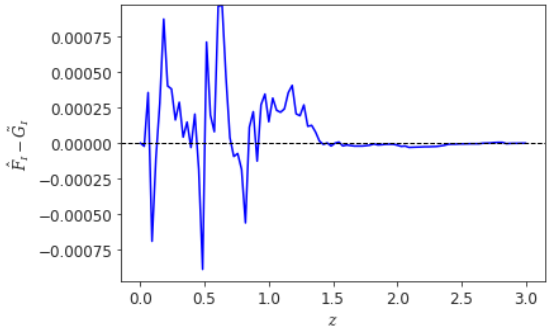
Before applying GALPRO, trained using the Zhou et al. dataset, to the PanSTARRS



(A) Spectroscopic versus Photometric redshift plot



(B) PIT plot



(C) Marginal Calibration plot

FIGURE 4.1: GALPRO results when trained and tested using the randomly selected 90% and 10% of the Zhou et al dataset respectively with the morphological parameters omitted from the training data arrays.

survey, it must first be ensured that both datasets have consistent variables. The PanSTARRS survey does contain the g, r, z, $W1$ and $W2$ magnitudes and their associated errors, however it does not contain the half light radius, model weight and axis ratio. This means that these morphological parameters must also be removed from the Zhou et al. training sample in order for GALPRO to be tested using the PanSTARRS data. To ensure that the removal of these three parameters from the training sample doesn't dramatically affect the RF algorithm itself, GALPRO was run in an identical manner as Section 3.2.2, but with these three columns omitted from the training data

that was used to generate the RF. GALPRO was run using the Zhou et al. sample with these three columns removed and 90% of the data randomly selected for training and 10% randomly selected for testing, as before. As shown in Figure 4.1, the removal of these columns only gives a very minor change in results. The scatter value only increases by 0.01 and the percentage of catastrophic outliers increases by 0.06%. Interestingly, the KLD, KST and CvM tests all decrease in value when the three parameters are removed which indicates that less bias has been introduced. This may be due to the fitting of the RF algorithm, as too many parameters can cause overfitting of the PDFs. It can be seen that the removal of the morphological parameters from the training and testing samples does not introduce any significant change or error in the outputted photometric redshifts. This reassures that omitting the three parameters from the Zhou et al. sample still leaves a sufficiently rich and diverse training dataset to be applied to other, new surveys, with negligible impact on the results.

Now, focus turns to applying the GALPRO-generated RF to the 'new' PanSTARRS survey. However, before the algorithm can be applied to the PanSTARRS survey, corrections must be applied to ensure that the training and testing datasets have compatible photometric band corrections and magnitude systems. These corrections are described in the following sections and then the RF algorithm is applied and analysis is performed.

4.1.1 Photometric Band Corrections

It may be expected that the RF algorithm, trained using one survey, cannot be straightforwardly applied to a new survey and give good results. There might exist some underlying differences in how the data for the two surveys are defined and calculated that need to be addressed. For example, it became apparent that the PanSTARRS survey defines the g, r and z bands slightly differently from the Zhou et al. survey meaning that the photometric systems are not consistent. The photometric bands are therefore not equivalent between the two surveys, also resulting in slightly different photometric band corrections and meaning that the data do not have statistically equivalent properties. This is not only true for the PanSTARRS and Zhou et al. training and testing samples, but *any* two surveys used to successfully train the RF and produce results must have compatible photometric systems and, consequently, compatible band corrections. The Zhou et al. dataset was compiled using the Desi Legacy Survey, which already

has the band corrections applied and provides the following equations to convert the PanSTARRS photometric bands to the Desi Legacy Survey band definitions [100].

The g, r and z bands usually are converted from the PanSTARRS band definition to the Zhou et al. definition using the below equations [102]:

$$g_{\text{Zhou}} = g_{\text{Pan}} + 0.00062 + 0.03604(g-i)_{\text{Pan}} + 0.01028(g-i)^2_{\text{Pan}} - 0.00613(g-i)^3_{\text{Pan}}$$
$$r_{\text{Zhou}} = r_{\text{Pan}} + 0.00495 + 0.08435(g-i)_{\text{Pan}} + 0.03222(g-i)^2_{\text{Pan}} - 0.01140(g-i)^3_{\text{Pan}}$$
$$z_{\text{Zhou}} = z_{\text{Pan}} + 0.02583 + 0.07690(g-i)_{\text{Pan}} + 0.02824(g-i)^2_{\text{Pan}} - 0.00898(g-i)^3_{\text{Pan}}$$

The PanSTARRS photometric bands had been converted to the same definition as the Zhou et al. sample, however the magnitude systems of the two surveys must now be taken into account before any analysis can be performed.

4.1.2 Ensuring Compatible Magnitude Systems

A comparison between the Zhou et al and PanSTARRS photometric bands is necessary to ensure that the two surveys have compatible magnitude systems. Figure 4.2 shows the CDFs of the population distributions in the $g, r, z, W1$ and $W2$ bands for both the PanSTARRS and Zhou et al. datasets. The KS test, as previously described in Section 3.1.1, is performed to quantify the difference between the two CDFs for each band. This provides a good measurement of how different or similar the CDFs are. The larger the value of the KST statistic, the smaller the probability that the null hypothesis, that the two CDFs are sampled from the same underlying distribution, is true.

Firstly, the g, r and z bands have very similar CDFs, which is reflected in the low KST scores of 0.041, 0,065 and 0.084 respectively. The distributions of each band can be seen for both surveys in Figure 4.4. Therefore, the distributions of g,r and z band photometry have consistent range and shape in both the PanSTARRS and Zhou et al samples.

The $W1$ and $W2$ bands, however show a significant difference in their distributions, as characterised by their KST results being at least an order of magnitude larger than the other bands. The $W1$ band gives a KST value of 0.826, and the $W2$ has a score of 0.878

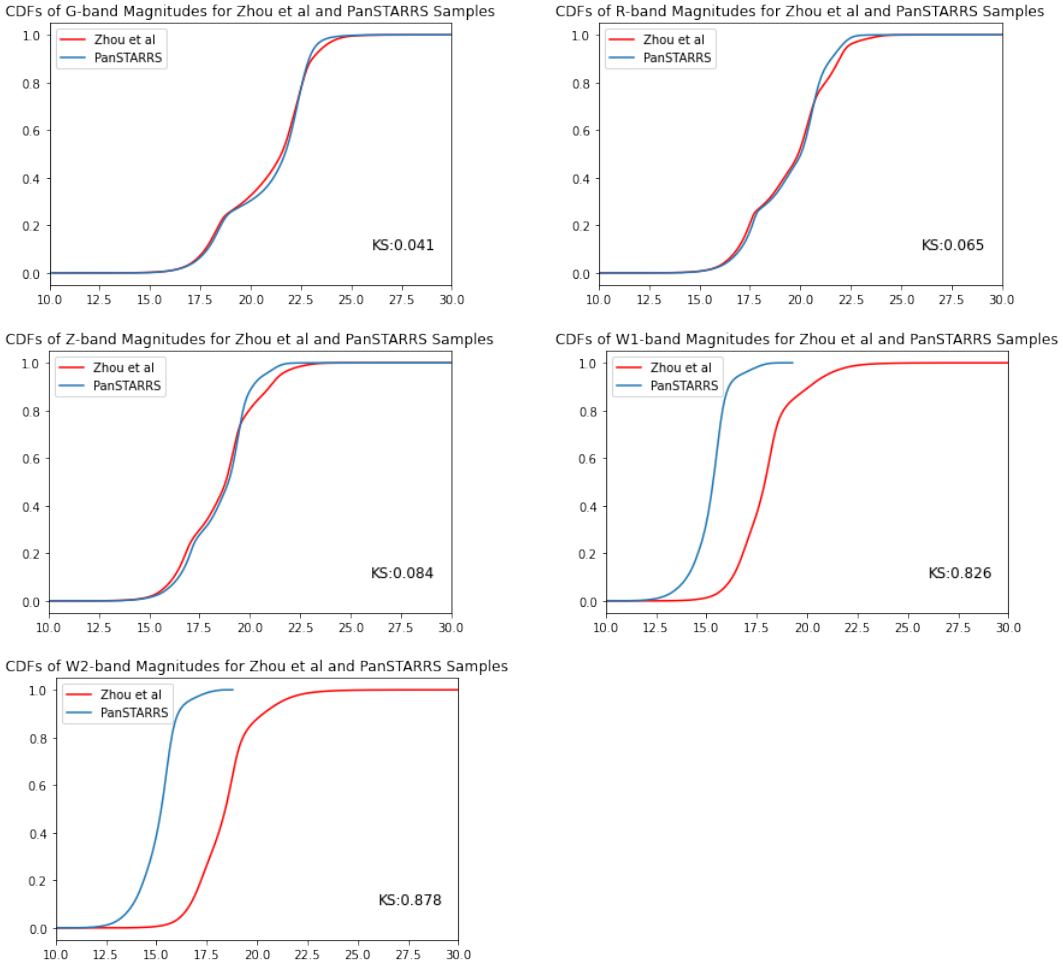


FIGURE 4.2: The CDFs of each photometry band in the PanSTARRS and Zhou et al. datasets. It is very clear that the $W1$ and $W2$ bands have significantly different distributions between the two surveys compared to the other photometry bands. Each plot also shows the KST statistic of the two CDFs for each band.

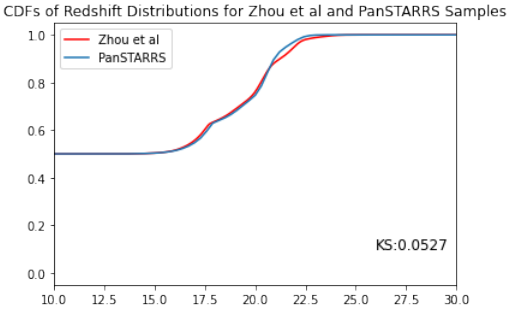


FIGURE 4.3: The cumulative distribution functions of the redshift distributions of the PanSTARRS and Zhou et al. datasets. The plots demonstrate the similarity in the two redshift distributions and the KST statitistic is shown on the plot.

which is significantly larger than the KST values of the g, r and z bands. It is clear that the $W1$ and $W2$ bands are not statistically equivalent between both surveys and have very different ranges due to PanSTARRS being a shallower survey, meaning the $W2$

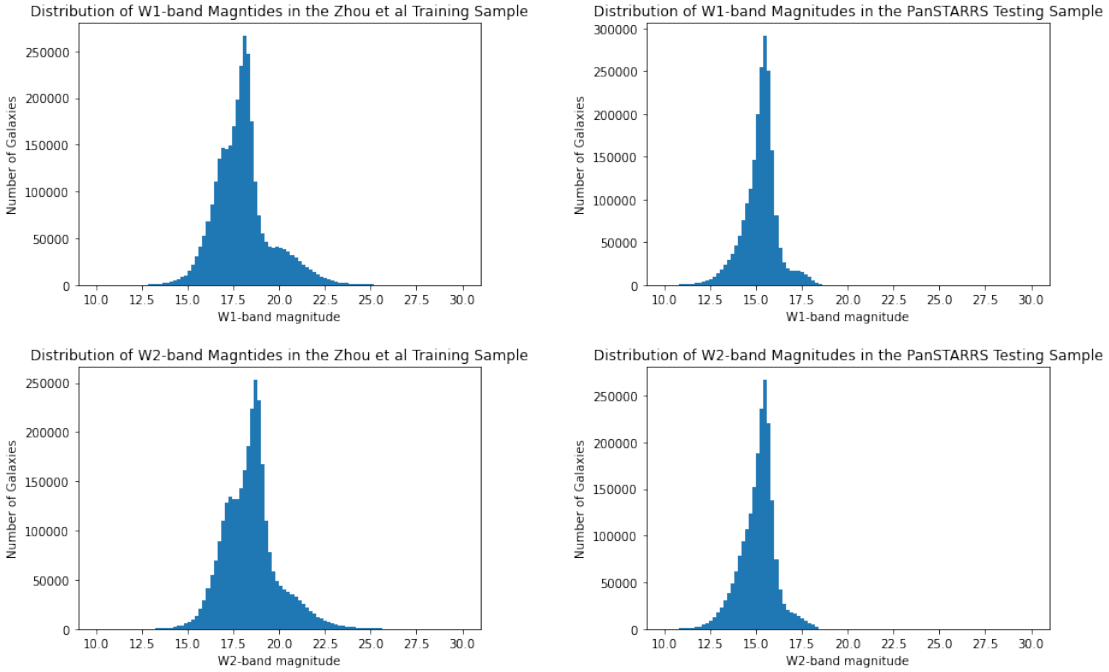


FIGURE 4.4: The W1 and W2 bands of the PanSTARRS and Zhou et al datasets.

band of the PanSTARRS survey is considerably brighter than the Zhou et al. survey. This discrepancy between the $W1$ and $W2$ bands of the two samples made it apparent that the W bands were defined by two different magnitude systems.

The PanSTARRS survey defined the W bands using the *vega* system, whereas the Zhou et al sample. defined these bands using the *ab* system [100]. It is extremely simple to convert between the two magnitudes, yet this conversion could potentially be the difference between a successful or unsuccessful application of the RF therefore it is imperative to ensure both the training and testing samples have compatible magnitude systems. The following equations allow for the conversion of the *vega* system w magnitudes to the *ab* system [103]:

$$m_{ab} = m_{vega} + \Delta m \tag{4.1}$$

with

$$\Delta m = \begin{cases} 2.699 & \text{for the W1 band} \\ 3.339 & \text{for the W2 band} \end{cases} \tag{4.2}$$

Where m_{ab} is the *ab* magnitude, m_{vega} is the *vega* magnitude.

From a glance at the distributions of the W bands from each survey (Figure 4.4) it appears that in each W band, the two surveys have very similar shapes yet are shifted by a value of around three. This difference in the definition of the two magnitudes seems responsible for this discrepancy. Once the PanSTARRS W bands were converted from the *vega* to *ab* system, the CDFs of the W bands can again be compared. These CDFs are shown in Figure 4.5. It is clear from the two CDFs that the W bands are much more statistically equivalent as their CDFs are clearly more similar. The $W1$ bands give a KST statistic of 0.1114 and the $W2$ band gives a result of 0.1105. This conversion to compatible magnitude systems has halved the KST statistic, producing compatible and much more similar W distributions. However, the W bands still do give larger KST statistics than the other g, r and z bands.

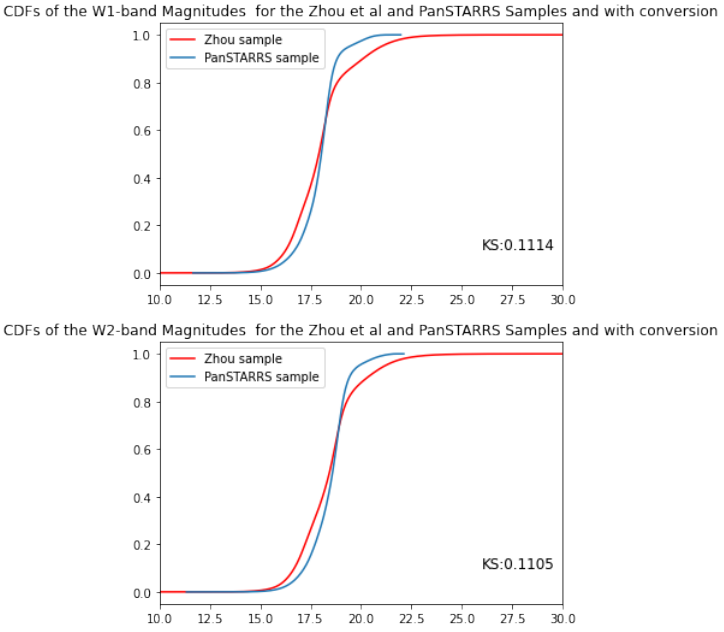
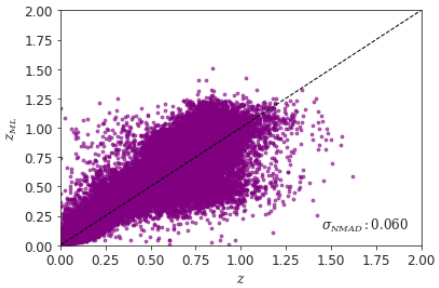


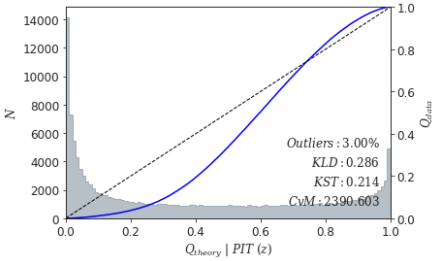
FIGURE 4.5: The CDFs of the W1 and W2 bands from the PanSTARRS dataset with K corrections applied and the magnitudes converted from the *ab* to *vega* system.

With the photometric band and magnitude system corrections applied, the two surveys are now ready to be used by the RF algorithm. GALPRO is to be trained using the full Zhou et al. dataset described in Section 3.2.5.1. A testing set was then generated, comprising of a randomly selected sub-sample of 200,000 PanSTARRS objects for which there were spectroscopic redshifts available. Although the PanSTARRS survey does contain photometric redshift values, they are significantly inaccurate and any comparison between the redshift estimates produced by GALPRO and the photometric

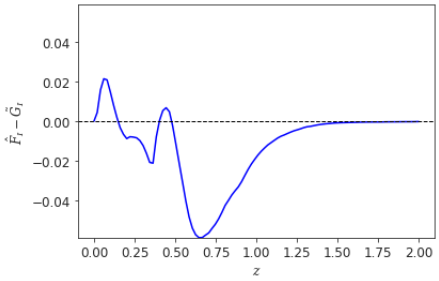
estimates given by the survey would not be useful or appropriate for calibration. The hyperparameters used in this computation are the same as those described in Section 3.



(A) Spectroscopic versus Photometric redshift plot



(B) PIT plot



(C) Marginal Calibration plot

FIGURE 4.6: GALPRO results when trained using the Zhou et al sample and tested using the PanSTARRS dataset with k corrections applied and the W magnitudes converted to a compatible system.

It was hoped that the corrections applied to the PanSTARRS survey eliminated any statistical difference between the two surveys that may be responsible for the failure of the RF algorithm. However, as shown in Figure 4.6, it is clear that there is still some issue present in the training and testing datasets that means the learnt RF algorithm cannot be applied to both datasets. The spectroscopic versus photometric redshift plot does present some rough correlation, however the scatter is large with a σ_{NMAD} value of 0.060. The PIT plot demonstrates that the redshift PDFs are not probabilistically calibrated with a very large CvM score of 2300.603. The PIT plot is extremely not uniform and displayed a hugely concave shape, meaning the redshift PDFs cannot be considered accurate.

This establishes that there must exist some discrepancy between the two samples that means that the mapping between the photometry and redshifts learned by the RF is not transferable between the two. The W bands have been corrected to give compatible magnitude systems, however the KST statistic of the W bands is still much larger than that of the other bands, meaning the two surveys are statistically similar, however a difference in their statistical properties is still present. The statistical equivalence between the two samples may determine how successful the RF application is, which is explored in the following sections. It is key to remember that the PanSTARRS survey is acting here like a completely new survey, for spectroscopic redshifts for testing may not be available.

4.1.3 Comparing Photometric Properties of the DESI and PanSTARRS Surveys

Before exploring how the overlap in the range of the photometry of the DESI and PanSTARRS survey affects the application of GALPRO, the photometric properties of two surveys may be compared. Table 4.1 displays some basic photometric properties of the two surveys, including the number of galaxies in each survey and the mean magnitude value of each band. This table also includes the magnitude completeness limit of each band for both surveys. The magnitude completeness limit involves determining up to what apparent magnitude the data are consistent with all galaxies being observable. The completeness test code ROBUST [104], which was developed in the early 2000s, is applied to both the DESI and PanSTARRS datasets to obtain the limits. This code takes both the spectroscopic redshift and magnitudes of a band for all of the galaxies in the survey and computes the completeness limit of that band using a statistical model [104].

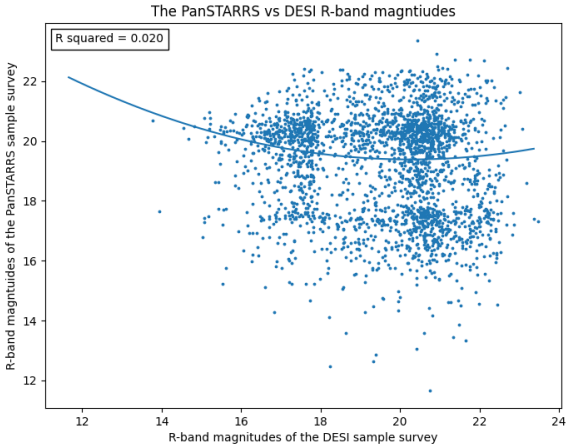
As shown in Table 4.1, the DESI catalogue sample has a slightly greater magnitude completeness limit in each band. This is to be expected, as the PanSTARRS survey is shallower than the DESI survey, and so will be complete out to a lower limit. However, the difference between the completeness limits for each band is relatively small, indicating that there isn't a significant difference in the magnitude completeness of the two surveys.

Properties	DESI Survey	PanSTARRS Survey
Magnitude completeness limit of R band	22.36	22.29
Magnitude completeness limit of G band	23.65	22.27
Magnitude completeness limit of Z band	21.32	21.27
Number of galaxies in survey	3,005,969	2,110,042
Mean R band magnitude	19.5167	19.4975
Mean G band magnitude	20.7776	20.9074
Mean Z band magnitude	18.4699	18.5272
Mean W1 band magnitude	17.9147	17.9224
Mean W2 band magnitude	18.4160	18.4471
Mean spectroscopic redshift	0.4336	0.4139

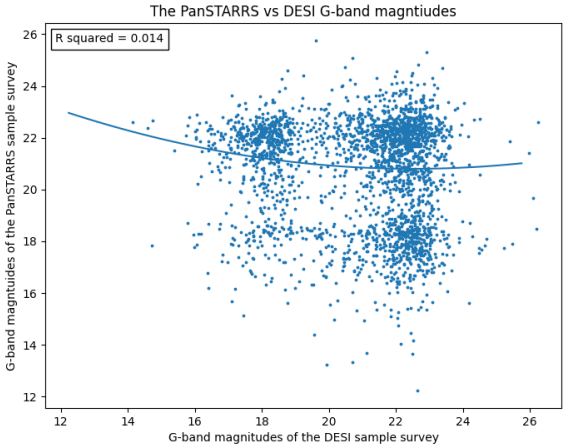
TABLE 4.1: Photometric properties of the DESI and PanSTARRS survey samples.

The photometry of the PanSTARRS and DESI catalogues may also be compared by analysing the common galaxies between the two surveys. The surveys are firstly cross-matched using the right ascension and declination of each galaxy. These two variables are compared for every galaxy in the two surveys, and identified as a common galaxy if both the right ascension and declination match to two decimal places. Due to the cross-matching process taking a very large amount of computational time, the surveys are randomly sampled to form sub-samples, of half a million galaxies each. The random sampling ensures that the sub-samples are completely representative of the complete surveys. Once the cross-matching process is complete, the identified common galaxies have their r, z and g apparent magnitudes in each band plotted, with the DESI magnitude on the x-axis and the PanSTARRS given magnitude on the y-axis. These plots can be seen in Figure 4.7.

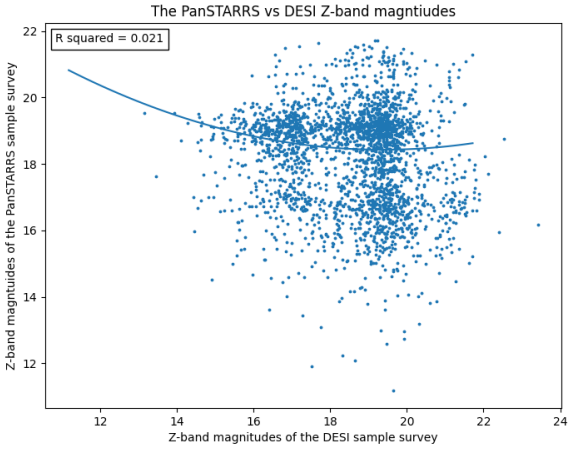
It can be seen in these figures that there appears to be a very large discrepancy between the magnitudes given by the two surveys, although the right ascension and declination are the same. The R-squared fit quantifies how independent the two variables are and is shown on each plot. An R-squared value of 1 would indicate that the two surveys generally agree and quote similar magnitudes for the common galaxies. However, the r, g and z bands all have very small R-squared values, of 0.020, 0.014 and 0.021 respectively. It is very clear from the plots that these surveys contain extremely different photometry values for galaxies with the same sky coordinates. This is very peculiar, as similar sky coordinates should ensure rather similar magnitude values. Tables 4.3 and 4.2 show a very small subsample of the cross-matched data from the DESI and PanSTARRS survey. These tables list the right ascension, declination, r, g and z band magnitudes for



(A) R band plot



(B) G band plot



(C) Z band plot

FIGURE 4.7: PanStarrs versus DESI photometry for galaxies in common.

the cross-matches samples, with each row detailing a single common galaxy. It is clear that the right ascension and declination of the common galaxies are matched correctly, however the two surveys quote very different magnitude values for the 'same' galaxy. Indeed, some of the magnitudes for the common galaxies differ by a rather large amount. The plots above show that even some of these common galaxies have differing quoted band values by an order of magnitude. This is rather unexpected and must be due the way in which the two surveys are compiled.

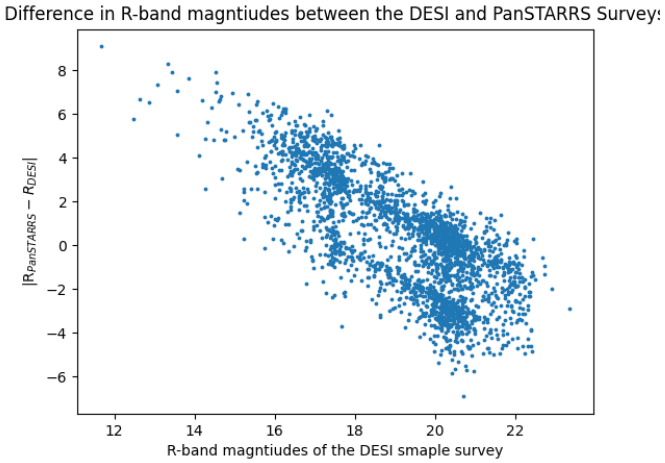
RA	Dec	R mag	G mag	Z mag
209.16	2.92	18.036	19.438	17.303
0.45	-1.16	20.296	22.153	19.328
161.61	9.6	19.441	21.606	18.385
18.55	19.39	18.752	20.528	17.916
175.7	51.64	17.219	18.201	16.642
220.32	25.31	19.554	21.589	18.575

TABLE 4.2: The right ascension, declination and band magnitudes of a subsample of the cross-matched DESI galaxies.

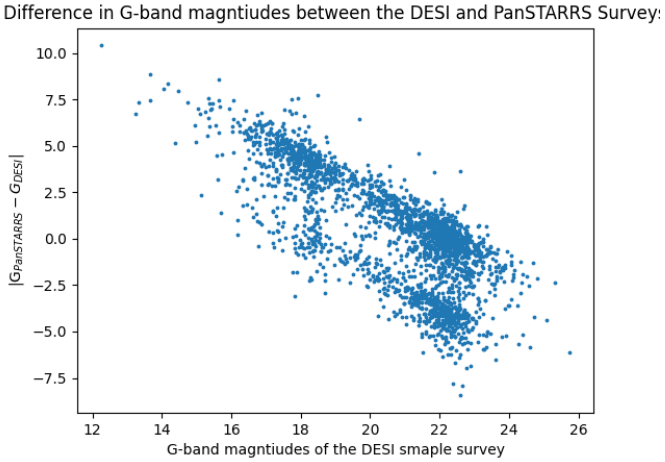
RA	Dec	R mag	G mag	Z mag
209.16	2.92	17.353	18.425	16.605
0.45	-1.16	17.838	18.706	17.187
161.61	9.6	20.57	22.49	19.104
18.55	19.39	20.575	22.468	19.022
175.7	51.64	20.241	22.102	18.949
220.32	25.31	17.38	18.788	16.613

TABLE 4.3: The right ascension, declination and band magnitudes of a subsample of the cross-matched PanSTARRS galaxies.

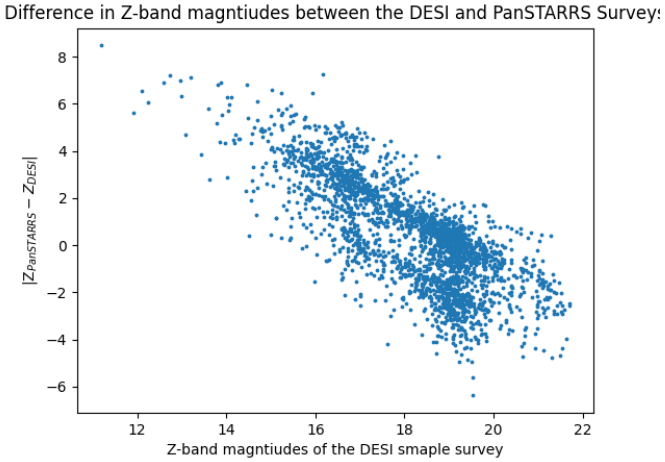
The residuals of these common galaxies are presented in Figure 4.8. Here, the x-axis shows the magnitudes of each band given by the DESI survey and the y-axis gives the difference between the magnitude values for each common galaxy (Magnitude$_{PanSTARRS}$ - Magnitude$_{DESI}$). This serves as a way to assess whether there is some pattern in the residuals. For every band, these plots show a higher positive difference between the two surveys at lower magnitudes, which tends to zero around a DESI magnitude value of 18, and then the residual becomes more negative as it moves towards higher magnitudes. This is further reinforced by Figure 4.9, which shows histogram plots that quantify the residuals. Here, the absolute mean residual over a magnitude range of 1 is calculated and plotted versus the DESI survey magnitudes. It also demonstrates that the residuals are much greater at lower magnitudes and then tend towards zero mid way through the



(A) R band plot

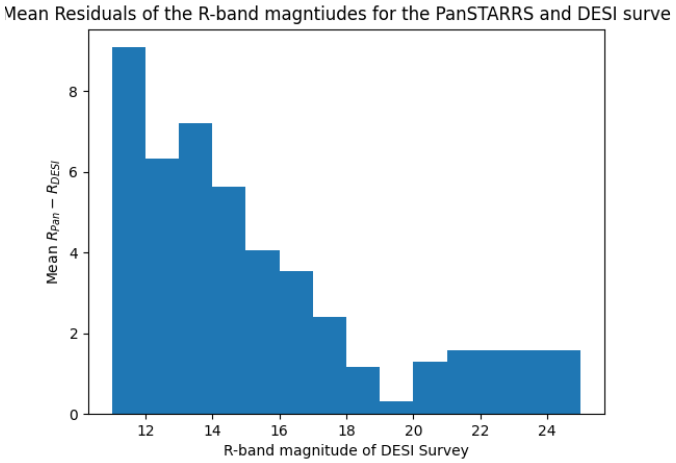


(B) G band plot

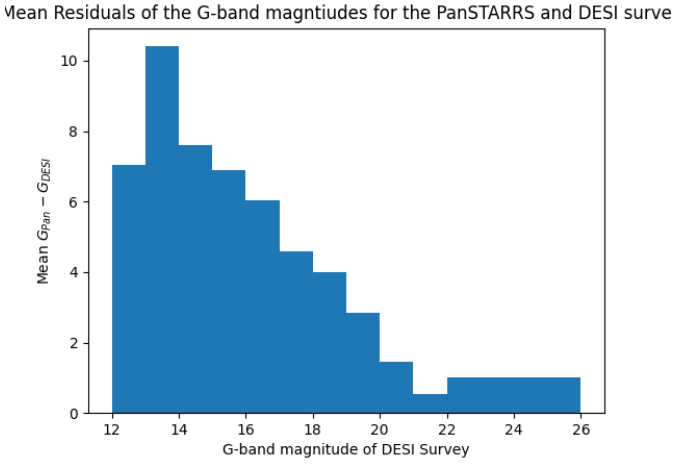


(C) Z band plot

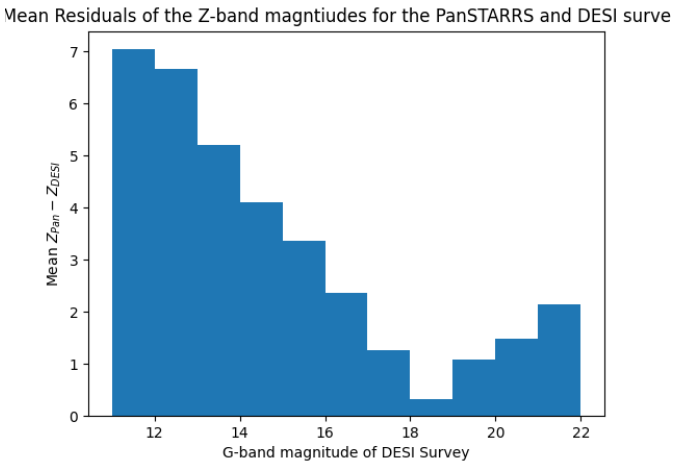
FIGURE 4.8: Scatter of the residuals (DESI - PanSTARRS) of the PanStarrs versus DESI photometry for galaxies in common.



(A) R band plot



(B) G band plot



(C) Z band plot

FIGURE 4.9: Histogram plot of the mean residual of (DESI-PanSTARRS) for common galaxies between the two surveys.

range. The mean residuals then increase in size at larger magnitudes, however their absolute mean value is much less than that of lower magnitudes.

This demonstrates that perhaps there is some systematic difference present in the two photometric surveys. The residual plots show that for smaller values, the PanSTARRS survey quotes much larger magnitudes than the DESI survey in every band. Around mid-way through the magnitude range this difference tends to zero. Then, as we move to larger values, the DESI Survey then provides larger magnitude values for the same galaxy. The reason for this remains unclear, however this systematic pattern present in the differing magnitude values provided for galaxies with the same sky location is concerning. GALPRO requires only photometry and spectroscopic redshifts as input variables, and so the following sections and application of the algorithm may provide insight into whether this difference significantly impacts performance.

4.2 How does the overlap in photometry range affect the performance of the RF?

When plotting histograms of the photometry for the individual bands in the PanSTARRS and Zhou et al. datasets, it becomes apparent that the range and distribution of the data are not consistent between the two surveys. This is expected as the PanSTARRS survey is shallower than the Zhou et al. sample, and these conflicting shapes/ranges may be the reason why the RF cannot be successfully applied to the PanSTARRS sample when trained on the Zhou et al. dataset. The main question explored in this section is how similar the properties of the two survey datasets have to be in order for the RF trained using the first survey to be applicable to the second survey. It is recognised in the previous section that having the same input variables may not be enough, but the range or distribution of those input variables may have to be similar or even identical. It is noted again here that the PanSTARRS and Zhou datasets are just two examples of general surveys which could be used to carry out a similar investigation, i.e. to determine how similar any two surveys may have to be for GALPRO trained on one survey to be successfully applicable to the other survey.

To achieve accurate photometric redshift estimates for the PanSTARRS survey using GALPRO trained on the Zhou et al. sample, some constraints may need to be placed

on the range of the photometric variables for the bands used in the testing and training samples. We might expect, for example, that these constraints will require overlapping ranges in the specified bands, in order to provide a sufficiently representative training sample. The purpose of this next section, therefore, is to understand how similar the properties of the two survey datasets need to be in order to produce reliable results, ie. how much overlap is required in the range and shape of each band in the training and testing samples to give accurate redshift estimates.

One way to quantify this is to take the Zhou et al. sample, which we know gives calibrated results with little error, and artificially split it, as though it is two different surveys. When the Zhou survey is split into two different subsamples, one may be used for training and the other for testing to *simulate* the situation of having two distinctly different photometric surveys. The overlap between the testing and training 'surveys' ranges may then be varied, from the most extreme case with no overlap in range, to a complete match in range, which allows us to assess how statistically equivalent the two general surveys must be to give reliable results. This provides an excellent stress test for when a trained RF can be applied to a new survey with differing distributions of photometry bands.

4.2.1 Case 1: Surveys 1 and 2 are statistically equivalent

To examine how the overlap of luminosity functions affects the redshift results, the Zhou et al. sample is split randomly in half. One half of the randomly sampled Zhou subset will be referred to as 'Survey 1' and the other half as 'Survey 2'. Since these two surveys were randomly selected from the same dataset, they should have identical ranges and shapes in every band. This case represents GALPRO being applied to a new survey which is statistically equivalent to the training survey. Firstly, Survey 1 is used as the training sample and Survey 2 is used as the testing sample. Then, for reassurance, Survey 2 is used as the training sample and the trained RF is applied to Survey 1. The r band magnitude ranges and redshift distributions of both surveys are seen in Figure 4.10 to demonstrate that these two surveys are indeed statistically equivalent. Each survey contains 1502983 galaxies, and initially Survey 1 is used to train GALPRO and Survey 2 is used for testing. Due to the probabilistic calibration process being computationally expensive, a subsample of 150,000 is randomly sampled from Survey 2

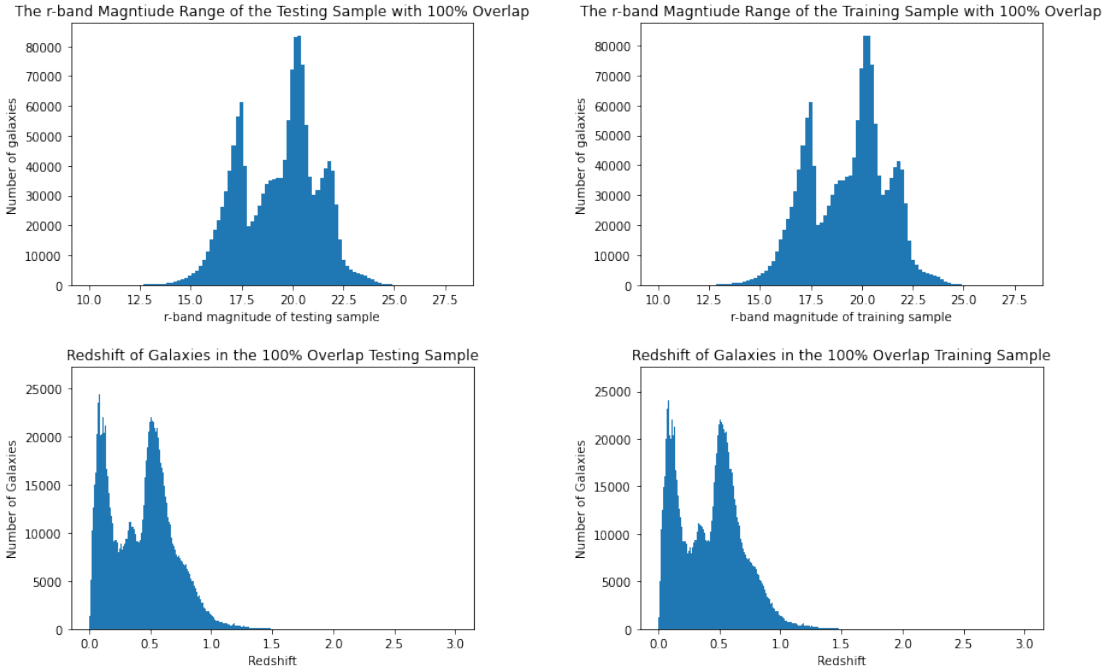
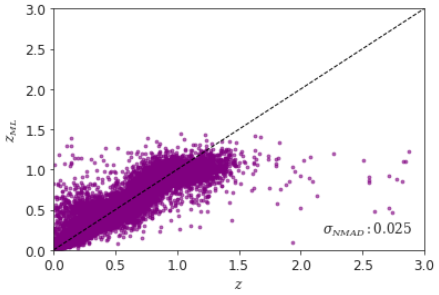


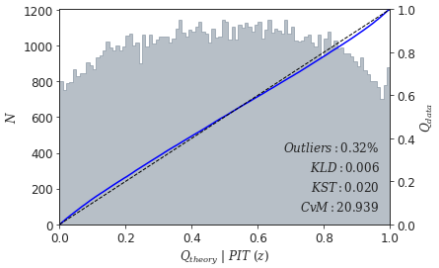
FIGURE 4.10: The r-band magnitude and redshift distributions of the training and testing samples with 100% overlap, meaning they have statistically identical distributions.

for testing purposes. Once the analysis has run, the roles are reversed and Survey 2 is used for training while a randomly selected subsample of 150,000 galaxies from Survey 1 is used for testing. This acts as a sanity check to ensure that both 'Surveys' give the same results to reassure the validity of these tests.

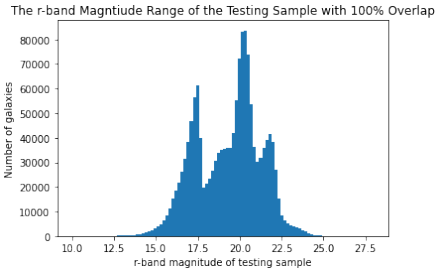
Figures 4.11 and 4.12 show the spectroscopic versus photometric redshift, PIT and marginal calibration plots for when Survey 1 is used for testing and Survey 2 for training and vice versa. As expected, these two tests produce accurate redshift estimates with a uniform PIT as the training and testing samples are statistically equivalent and have no underlying difference in features. Both spectroscopic versus photometric redshift plots give a σ_{NMAD} value of 0.025 and follow a good correlation with a slight increase in scatter as the redshift value increases. This is to be expected, as described in Section 3.2.2.1. Both cases give good marginal calibration plots with little variation around the zero line. The outlier fraction, KLD and CvM tests only vary between the two cases by an insignificantly small percentage which can be attributed to the random sampling of the datasets. Overall, there, the results obtained for Case 1 were as expected, as the randomly sampled 'surveys' with equivalent statistical distributions gave accurate redshift estimates and behaved identically when trained using Survey 1 and tested using



(A) Spectroscopic versus Photometric redshift plot



(B) PIT plot



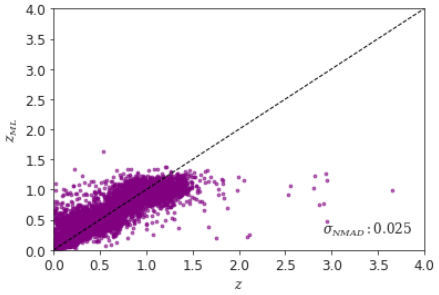
(C) Marginal Calibration plot

FIGURE 4.11: GALPRO results when trained using Survey 1 and tested using Survey 2 with 100% overlap, meaning the testing and training samples have statistically identical distributions.

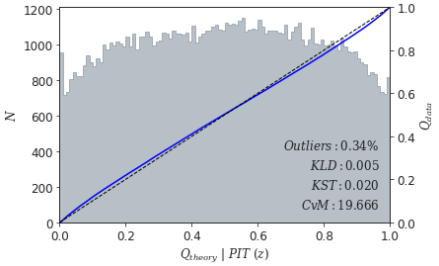
Survey 2 and vice versa. This is reassuring and allows for the following investigation of how the overlap in range and distribution affects the application of the RF.

4.2.2 Case 2: Surveys 1 and 2 have minimal statistical equivalence

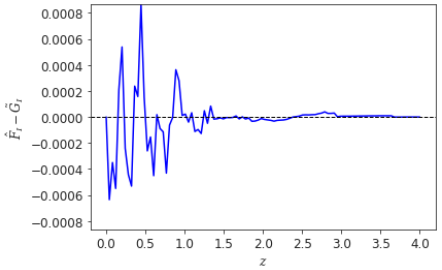
Case 2 explores how the RF algorithm performs in the extreme case where Survey 1 and Survey 2 have no overlapping range and therefore don't have the same statistical distribution. This is done by splitting the Zhou et al. truth dataset in half, but this time finding the median of the r-band magnitude range and taking Survey 1 to include all of the galaxies with $r \leq r_{\text{median}}$ and Survey 2 to include all of the galaxies with $r \geq r_{\text{median}}$. The r-band magnitude is chosen for the splitting process as it has the largest range of all the bands which makes it easier to visualise the plots and also gives



(A) Spectroscopic versus Photometric redshift plot



(B) PIT plot



(C) Marginal Calibration plot

FIGURE 4.12: GALPRO results when trained using Survey 2 and tested using Survey 1 with 100% overlap.

the greatest difference in ranges between the two surveys. Figure 4.13 shows the redshift and r-band magnitude distributions of Survey 1 and Survey 2. Survey 1 was then used as the training sample and a randomly selected subset 150,000 galaxies from Survey 2 was used for testing, as before in Case 1.

The spectroscopic versus photometric redshift, PIT and marginal calibration plots for Case 2 are shown in Figure 4.14. Unsurprisingly, this test gave very inaccurate results with the PIT plot indicating a large amount of bias introduced. The PIT plot is catastrophically non-uniform and has a CvM value of 5157.074. The very steep gradient indicates that the redshift PDFs contain a huge bias. The marginal calibration plot peaks at a value of 0.2, which is much greater than Case 1 and shows that the results are not marginally calibrated. The scatter plot showed no correlation between spectroscopic and photometric redshift, and has a sharp photometric redshift cut off around z

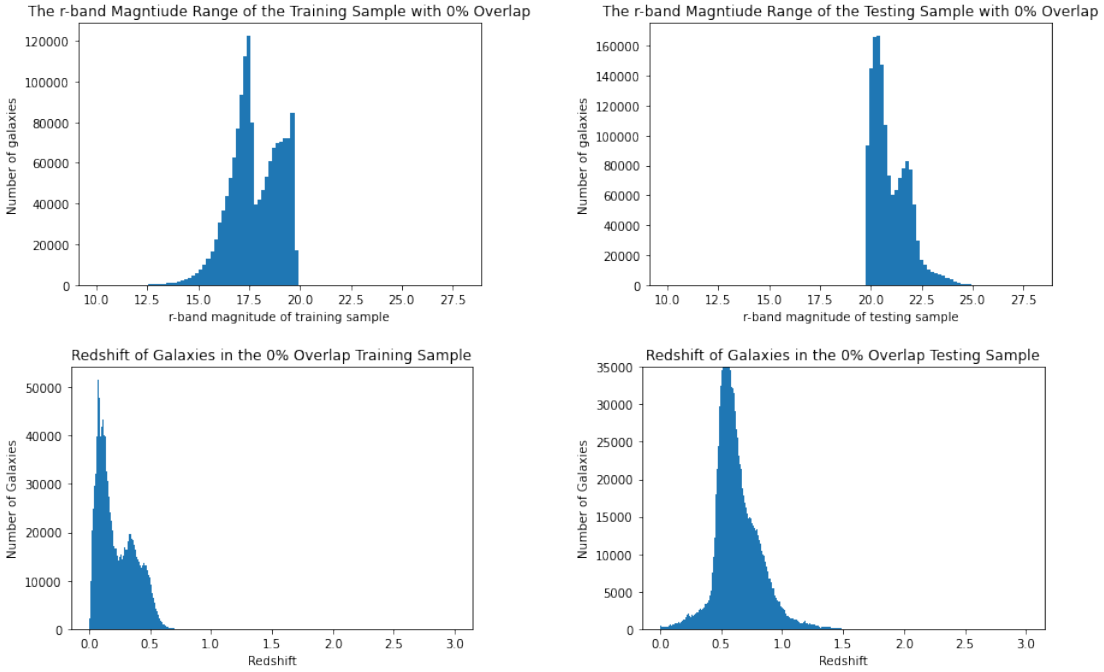
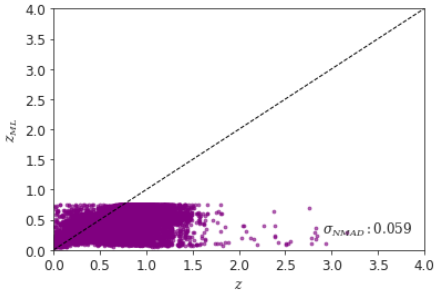


FIGURE 4.13: The r-band magnitude and redshift distributions of the training and testing samples with 0% overlap. The left-hand plots show Survey 1, while the right hand-plots show Survey 2.

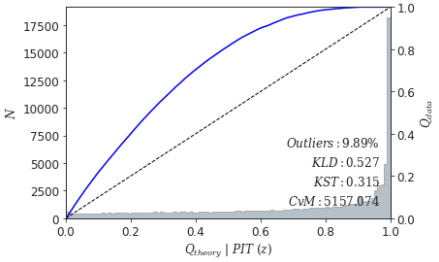
= 0.75. This is interesting, as the redshift distribution plot of Survey 1 cuts off around z = 0.6, while Survey 2 contains objects with mainly z = 0.5-1.5. This demonstrates that the RF learns the mapping between photometry and redshift within the redshift range of the training sample and will not estimate photometric redshifts outside of this redshift range. When a new survey that contains redshifts outside of the training redshift range is used for testing, the RF fails to predict values outside of the training redshift range. This is a very important takeaway from this test, as any future use of GALPRO and potentially other RF algorithms for photometric redshift estimation must have a representative training sample with a large redshift distribution. This is discouraging, as it suggests that it may not be possible to apply GALPRO to a general, new photometric survey that has no spectroscopic redshift information since the redshift range of the testing and training datasets may be insufficiently similar.

4.2.3 Case 3: Varying the statistical equivalence between Surveys 1 and 2

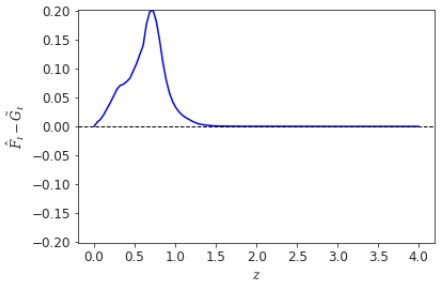
It was shown in the previous subsection that GALPRO cannot be successfully applied to two datasets with significantly differing distributions of redshift and r-band magnitudes.



(A) Spectroscopic versus Photometric redshift plot



(B) PIT plot



(C) Marginal Calibration plot

FIGURE 4.14: GALPRO results when trained using Survey 1 and tested using Survey 2 with 0% overlap.

This is due to the RF not having representative training data to learn the mapping between the photometry and redshifts over these different ranges. These next tests are used to determine how much overlap is required between the training and testing samples for GALPRO to be successfully applied. This is implemented by training using Survey 1, which is restricted to cover a certain range in the r-band magnitude distribution, and testing using Survey 2 which is also restricted to certain r-band magnitude range. These surveys are restricted in such a way that gives 90%, 80% and 70% overlap between the two survey distributions to assess how this degree of overlap affects the application of GALPRO. For example, for the 80% overlap test the galaxies are split such that all of the galaxies included in the 80% range area centered on r_{mean} are randomly sampled so that half are contained in Survey 1 and the other half in Survey 2. This is done by:

- Identifying the 80% overlap region as the range that lies between the 10% (r10) and 90% (r90) percentile in the CDF of the r-band magnitude distribution.
- Construct Survey 1 by sampling X1 galaxies from the 80% overlap region, between r10 and r90, and the remaining Y1 galaxies with r $<$ r10. Construct Survey 2 by sampling X2 galaxies from the 80% overlap region, between r10 and r90, and the remaining Y2 galaxies with r $>$ r90

Then Survey 1 is assigned all of the galaxies below this 80% overlap range and Survey 2 contains all those above the overlap range. Obviously, for the 90% and 70% overlap tests, the above method is used to construct the surveys but with the appropriate Survey 1 and 2 will contain 1502983 galaxies, and 150,000 galaxies from Survey 2 are randomly sampled and used for testing. This allows for the establishment of some baseline as to how similar the two surveys must be for GALPRO to produce accurate photometric redshift estimates.

4.2.3.1 90% Overlap

Firstly, the training and testing samples are split such that Survey 1 and Survey 2 have a 90% overlap region. The two r-band magnitude and redshift distributions can be seen in Figure 4.15. Although similar, the redshift distributions of the two datasets indicate there is still a difference in the depth and shape of the redshifts included in each survey. GALPRO was then trained using Survey 1 and tested using Survey 2 using the same settings as all of the previous tests.

The spectroscopic versus photometric redshift scatter plot, PIT and marginal calibration plots can be seen in Figure 4.16. The PIT produced by this test is generally uniform however it contains a small dip downward as it moves towards the lower values. The Q-Q plots do not show a large difference between the U(0,1) values and results from this test, however the gradient of the PIT indicates some bias has been introduced. The outlier fraction, KLD and KST tests gives reasonable values, however the CvM test gives a value of 50.347 meaning a bias is most certainly present in the results. The marginal calibration plot peaks at an order of magnitude larger than Case 1, and only oscillates above the zero line instead of randomly oscillating about the zero line. This

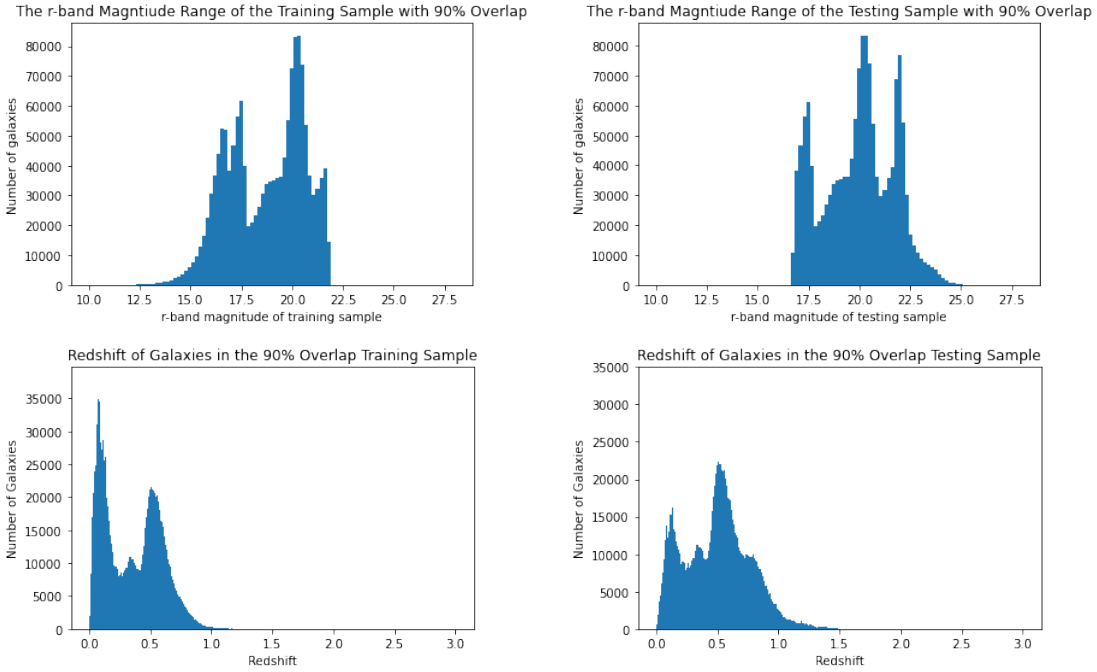
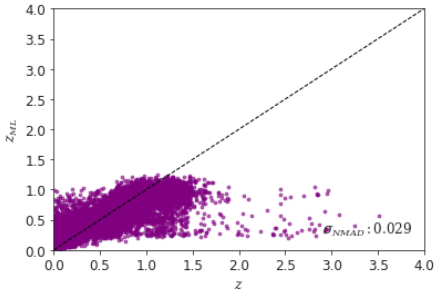


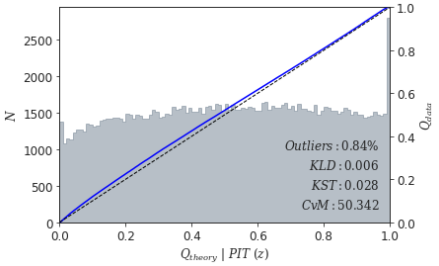
FIGURE 4.15: The r-band magnitude and redshift distributions of the training and testing samples with 90% overlap. The training dsitributions are shown on the left and the testing on the right.

shows that the PDFs are not completely marginally calibrated as the plot peaks around 0.01, whereas a value of less than 0.005 is required for successful marginal calibration.

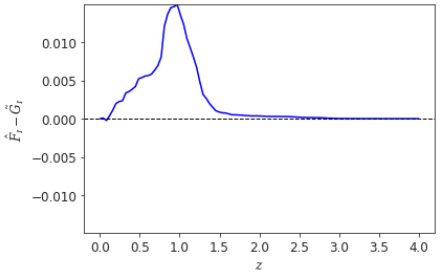
The photometric versus spectroscopic redshift plot gives a value of $\sigma_{NMAD} = 0.029$, which is reasonable; however, the plot itself shows that some bias has obviously been introduced. There is some correlation between the photometric and spectroscopic redshifts in the range $0.5 < z < 1$, however out to larger z values, GALPRO underestimates the photometric redshifts quite significantly. Due to the r-band magnitude cut-off, the training sample contains little to no galaxies with a redshift greater than 0.9 whereas the testing sample reaches out to a redshift of 1.5. The severe underestimation of photometric redshifts out to these higher values demonstrates that GALPRO may only learn the mapping between the fluxes and redshift for the given training sample and does not seem able to extrapolate the mapping out to higher/lower redshift values. Not only the range, but also the shape of the redshift distributions of the testing and training samples seem to affect results. For $z < 0.5$, the redshift values produced by GALPRO are slightly overestimated. The training sample contains many more galaxies with redshift values in the range $0 < z < 0.4$ compared to the testing sample, which explains why there is an overestimate of photometric values around this redshift range. The testing



(A) Spectroscopic versus Photometric redshift plot



(B) PIT plot



(C) Marginal Calibration plot

FIGURE 4.16: GALPRO results when trained using Survey 1 and tested using Survey 2 with 90% overlap.

sample does contain galaxies at lower redshifts; however, the shape of the two redshift distributions for $z < 0.5$ is different, which leads to the underestimation of the photometric redshifts and introduced bias into the results. This further demonstrates that the results produced by GALPRO are affected by the shape of the redshift distribution, as at these low redshifts, the two samples have the same range but a different shape. The overlap in r-band magnitudes between the two samples is 90%, which is very high, giving little reassurance for the 80% and 70% overlap tests, as a decrease in overlap will surely further deteriorate results. We investigate this further in the next sections.

4.2.3.2 80% Overlap

Now, the training and testing samples are split such that Survey 1 and Survey 2 have an 80% overlap region, with Figure 4.17 showing the r-band magnitude and redshift

distributions of the two surveys. The redshift distributions of the two surveys have very differing ranges and shapes, with Survey 1 containing a large amount of galaxies with redshifts in the range $0.25 < z < 0.75$ and reaching out to 0.75 redshift. On the other hand, Survey 2 contains galaxies with a redshift range out to $z < 1.5$ and only has a much smaller number of low redshift galaxies. Following the previous test using the 90% overlap samples, it is expected that the redshift estimates will contain more bias and be less accurate, as the training and testing samples have less overlap in their variables. GALPRO was then trained using Survey 1 and tested using Survey 2, with the same settings as all of the previous tests.

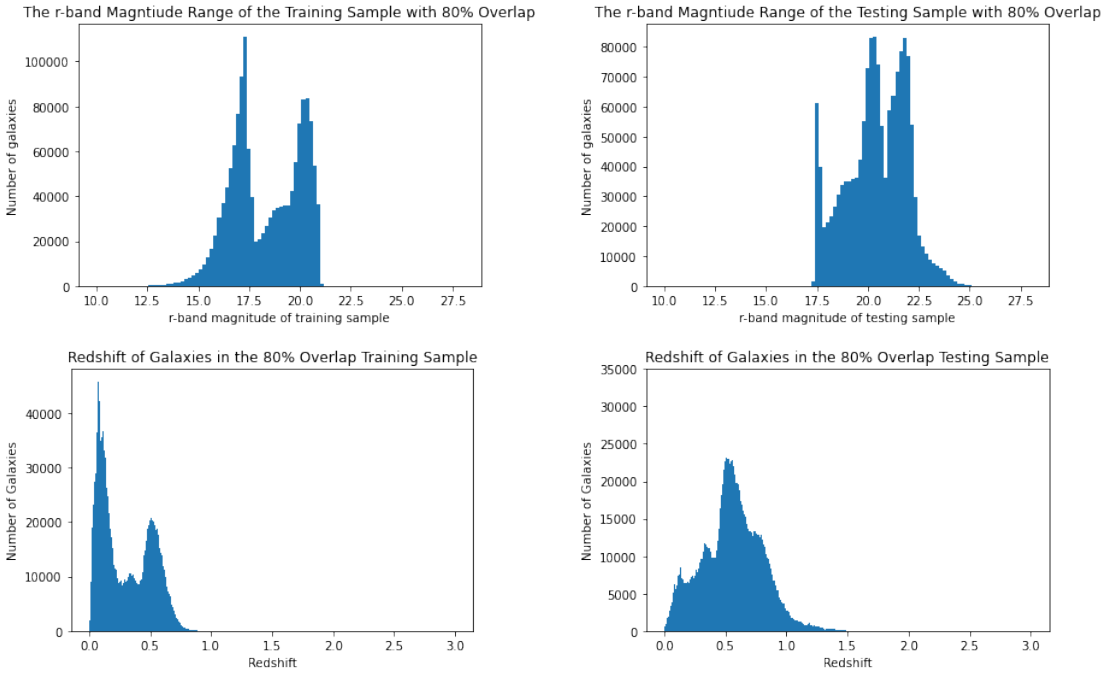
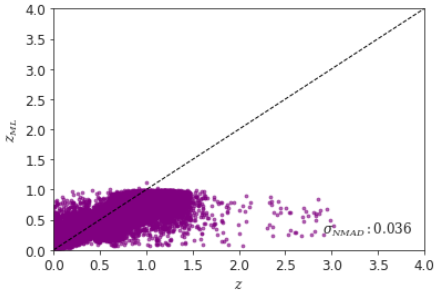
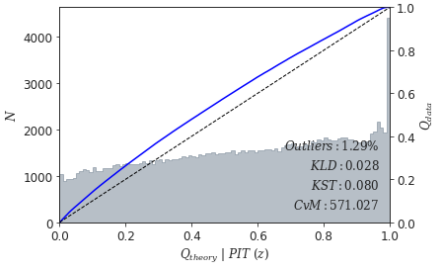


FIGURE 4.17: The r-band magnitude and redshift distributions of the training and testing samples with 80% overlap.

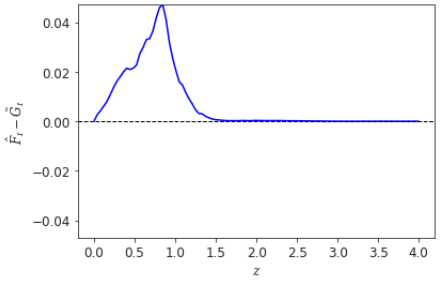
The spectroscopic versus photometric redshift scatter plot, PIT and marginal calibration plot can be seen in Figure 4.18. The PIT plot again shows bias has been introduced to the PDFs as it has a steeper gradient than the 90% overlap test. The outlier fraction has increased as the overlap percentage decreases and the CvM test gives a value of 571.027, indicating that probabilistic calibration has not been successful. The Q-Q plot deviates further from the U(0,1) distribution and the large gradient of the plot shows that the results are far from perfect. The marginal calibration plot again only oscillates in one direction out to 0.04 meaning that the PDFs are not marginally calibrated. This value is four times larger than the 90% overlap test, meaning that both the marginal and



(A) Spectroscopic versus Photometric redshift plot



(B) PIT plot



(C) Marginal Calibration plot

FIGURE 4.18: GALPRO results when trained using Survey 1 and tested using Survey 2 with 80% overlap.

probabilistic calibration have decreased significantly as the overlap percentage decreases by 10%.

The photometric versus spectroscopic redshift plot gives a value of $\sigma_{NMAD} = 0.036$, meaning that the scatter has worsened as the overlap has decreased. There is little correlation between the photometric and spectroscopic redshifts and the plot is starting to behave more like Case 2, where there is 0% overlap. Again, out to large spectroscopic values, the photometric redshift estimates are very inaccurate and GALPRO fails to predict any estimates larger than a redshift of around 1. The training sample doesn't contain galaxies with a redshift greater than around 0.8, yet the testing sample contains galaxies out to a redshift of 3.5 with the majority lying in the $0.5 < z < 1.5$ region. This again confirms that GALPRO is unable to successfully extrapolate the learnt mapping out to redshifts beyond those that are contained in the training sample. Even in the

areas where the training and testing samples have an overlapping redshift range, the results are considerably more inaccurate than the previous 90% overlap test. As the two surveys have differing redshift distribution shapes within the overlapping ranges, the mapping between the fluxes and redshift is not successfully learned for Survey 1 and applied to Survey 2, which is evident in the inaccurate redshift estimates produced for $0.4 < z_{spec} < 0.6$. This is disheartening, as it seems that not only does the range of the two r-band magnitude distributions have to be very similar for each survey, but also the shapes of the redshift distributions must be similar within an overlapping range. The two surveys having an r-band magnitude overlap of 80% is relatively generous, as this is not dissimilar to the sort of overlap two new, real surveys may have, however the RF does not perform to a satisfactory standard.

4.2.3.3 70% Overlap

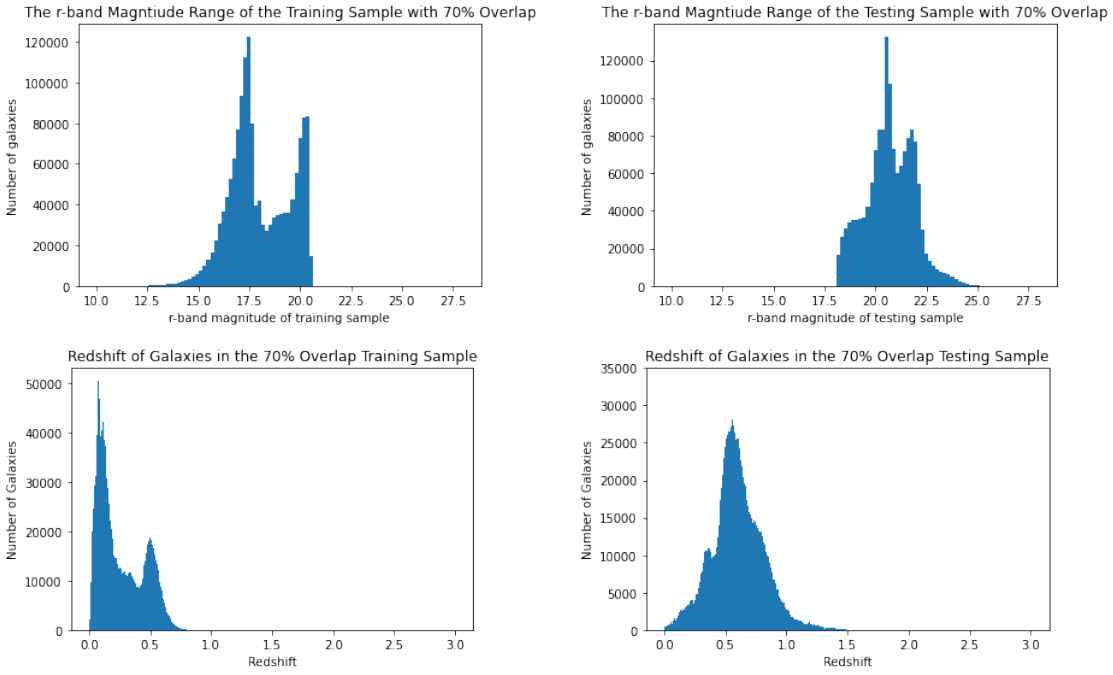
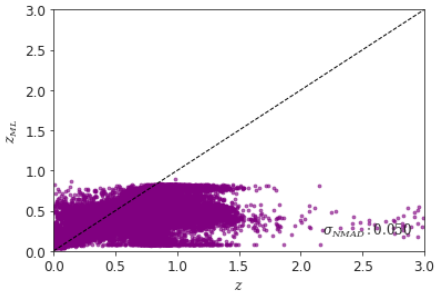


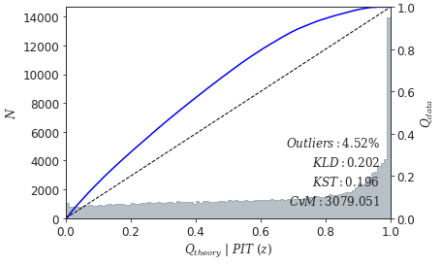
FIGURE 4.19: The r-band magnitude and redshift distributions of the training and testing samples with 70% overlap.

Finally, the training and testing samples are split so that they have a 70% overlap in the r-band magnitude, which can be seen graphically in Figure 4.19, alongside the two survey's redshift distributions. It is expected that this test will produce very inaccurate results containing a large bias, however the interest is in how much the reliability of the results deteriorate with overlap percentage. Survey 1 is again used for training and

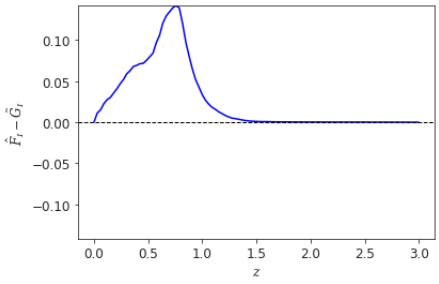
Survey 2 for testing. The r-band magnitudes have a 70% overlap, which produces very dissimilar redshift distributions between the two surveys. These two distributions are opposing in that Survey 1 contains many galaxies with redshifts around 0-0.2 and then dips between 0.3-0.5 and peaks again just after a redshift of 0.5, cutting off at around 0.75. Survey 2 contains few lower redshift galaxies but has a large peak around a redshift of 0.5 and trails out to a redshift of 1.5. Survey 1 was used for training and Survey 2 for testing.



(A) Spectroscopic versus Photometric redshift plot



(B) PIT plot



(C) Marginal Calibration plot

FIGURE 4.20: GALPRO results when trained using Survey 1 and tested using Survey 2 with 70% overlap.

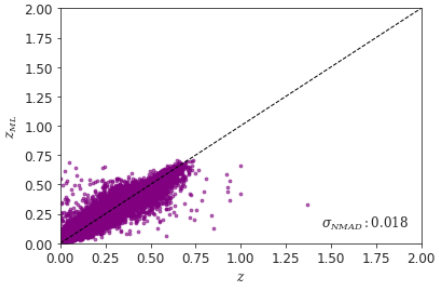
As expected, this test produced highly inaccurate redshift estimates, which are seen in Figure 4.20. The PIT shows catastrophic bias was introduced with a very steep gradient and a CvM value of 3079.051. It appears that with each 10% overlap increment decreased, the CvM value increases by a power of ten. The Q-Q plot deviated greatly from the uniform distribution and this plot shows that probabilistic calibration was

not successful. The marginal calibration peaks around 0.15, which is much greater than the 80% overlap test, and only oscillates above the zero line, again showing that marginal calibration was not successful. The spectroscopic versus photometric redshift plot shows highly inaccurate results with little to no correlation and a value of σ_{NMAD} of 0.05. GALPRO fails to predict any photometric redshift value greater than around 0.75, which correlates with the cut-off in redshift of the training sample. This again reinforces the fact that GALPRO may only predict redshift values within the limits of the training sample. The decrease in overlap by 10% has deteriorated the accuracy in redshift estimates and calibration by a significant amount and demonstrates that even a small overlap decrease can greatly affect the performance of the RF. Although a 70% overlap in r-band magnitude range doesn't seem like a large difference in the flux distributions, it is clear that this difference is enough to cause a complete failure in the RF algorithm and lead to inaccurate and unreliable results.

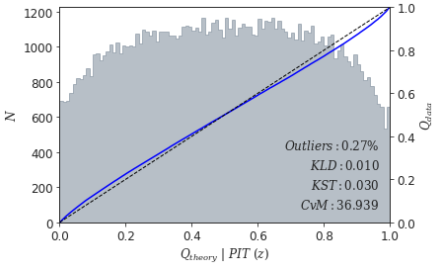
4.2.4 Case 4: Survey 1 has a larger statistical range than Survey 2

Case 3 explored how the overlap of the r-band magnitudes of the training and testing samples affected the application of the RF algorithm. This test is similar, and splits the Zhou et al. dataset into two new 'surveys', Survey 1 and Survey 2, each containing 1502983 galaxies. However, now Survey 1 contains randomly sampled galaxies over the full range of the r-band magnitude while Survey 2 is restricted. Firstly Survey 2 is restricted to only contain galaxies below the r_{mean} of the Zhou et al dataset. The r-band magnitudes and redshift distributions of Survey 1 and Survey 2 can be seen in Figure 4.22. Survey 1 is used for testing and Survey 2 for training and GALPRO is run using the same settings as all previous tests. This test is similar to those in Case 3, however now the entirety of the range and shape of the Survey 2 overlaps with Survey 1.

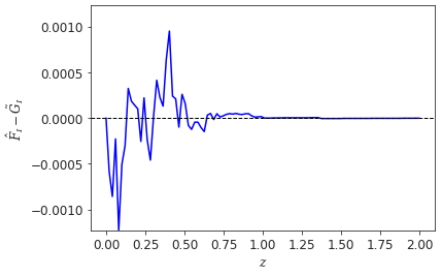
The results of this test are shown in Figure 4.21. The marginal calibration plot oscillates about the zero line out to a maximum of around 0.001 which shows that marginal calibration was successful. The PIT plot has an outlier fraction of 0.27% and the KLD and KST tests give values close to zero. The CvM test has a value of 36.939 which shows some deviation from the uniform distribution and the PIT is slightly convex showing that the PDFs are overly narrow, however these results are much better than the 80% and 70% overlap tests. The photometric versus spectroscopic scatter plot shows a good



(A) Spectroscopic versus Photometric redshift plot



(B) PIT plot



(C) Marginal Calibration plot

FIGURE 4.21: GALPRO results when trained using the entire r-band magnitude range and testing is restricted to below r_{mean}.

correlation between the true and estimated redshift values with no outstanding tendency to under or overestimate at a particular redshift value.

This test is repeated, with Survey 1 still spanning the entirety of the r-band magnitude range but now Survey 2 is restricted to only contain galaxies with an r-band magnitude above r_{mean}. The r-band magnitude distributions of the two surveys are shown in Figure 4.23. GALPRO is applied again using Survey 1 for training and Survey 2 for testing.

The results of this test are shown in Figure 4.24. The marginal calibration plot oscillates about the zero line with a maximum of 0.001, again showing that marginal calibration was successful. The PIT is very uniform, with an outlier fraction of 0.39% and KLD and KST tests close to zero. The CvM value is 8.674 and the Q-Q plot shows a very close match between the results and the uniform PIT graph. In comparison to the previous result where the testing sample is restricted to galaxies below the r_{mean} value, this PIT

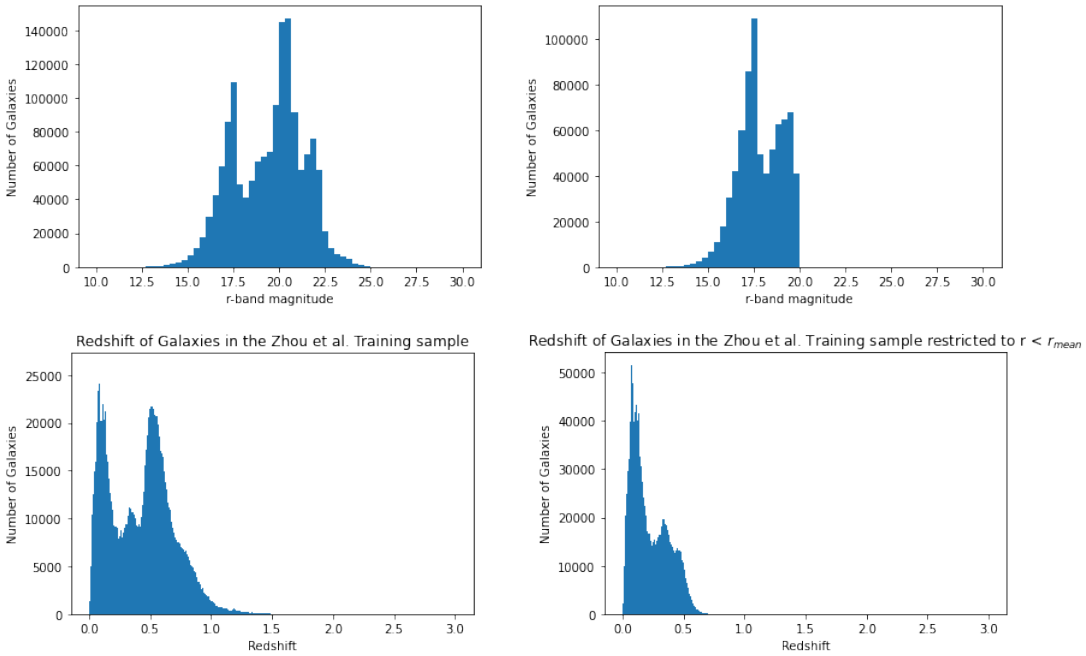


FIGURE 4.22: The r-band magnitude and redshift distributions of the training and testing samples, where the training sample covers the whole range and the testing sample is restricted to below r_{mean}.

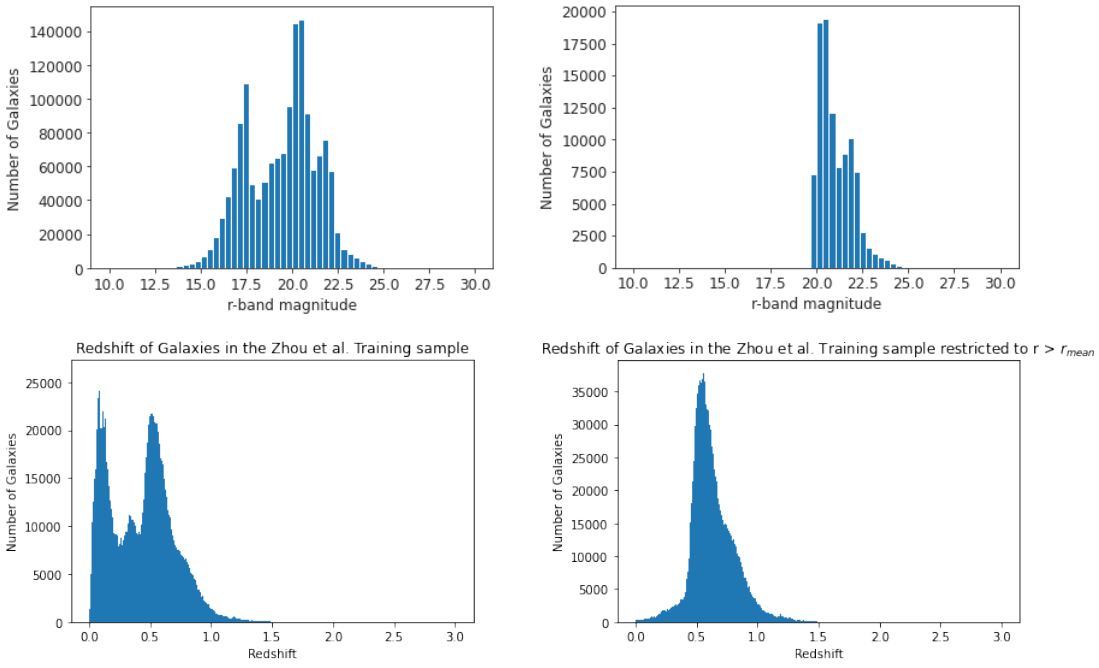
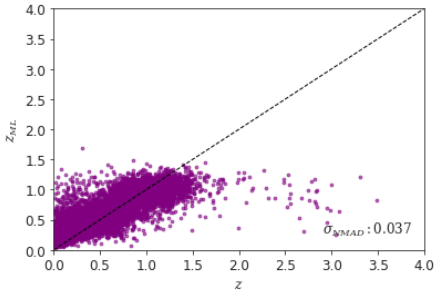
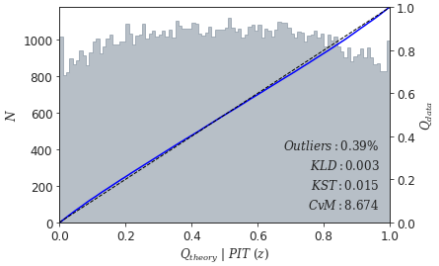


FIGURE 4.23: The r-band magnitude and redshift distributions of the training and testing samples, where the training sample covers the whole range and the testing sample is restricted to above r_{mean}.

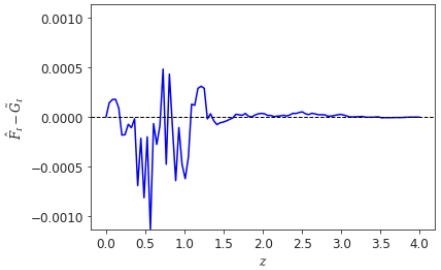
is only slightly more uniform and also doesn't show any issues with the PDFs. This may be due to the r-band magnitude distribution, as Survey 1 contains more galaxies above r_{mean}, and so the training sample is more representative at the higher r-band magnitude



(A) Spectroscopic versus Photometric redshift plot



(B) PIT plot



(C) Marginal Calibration plot

FIGURE 4.24: GALPRO results when trained using the entire r-band magnitude range and testing is restricted to above r_{mean}.

range, giving more accurate results. The scatter plot shows a good correlation between spectroscopic and photometric redshift, which underestimates redshifts as it moves out to higher redshift values. This is to be expected as it follows the same trend as Section 3.2.2.1.

Overall, it seems that if the range of the photometry of the training sample is as large, or larger, than the range of the testing sample, the RF produces probabilistically and marginally calibrated results to a satisfactory degree of accuracy. This means that the testing sample may only have to overlap with the training sample by 50% to generate accurate results, provided that the training sample range completely covers the range of the testing sample. Case 3 demonstrated that a decrease in overlap between the training and testing sample ranges decreases the quality of the results. However Case 4 indicates that it is the fraction of the testing sample that isn't covered by the range of the training

sample that causes the degradation of the results and not just a difference in the range values of the data. It also appears that whether the testing sample range corresponds to the lower or higher end of the training sample range doesn't affect the outcome and still provides satisfactory results. Interestingly, this case performs much better than the previous cases with 90/80/70% overlap. This may be due to the range and more importantly, the shape of the redshift distributions. In the previous cases, the redshift distributions have mostly overlapped in range with the testing sample reaching out to slightly deeper redshifts than the training sample. This still caused a large introduction of bias and inaccurate redshift estimates, even in areas of overlapping redshift. However, it is key to note that in these areas of overlap, the redshift distributions had different shapes due to the way in which the sampling was performed. Since the previous results showed a large amount of inaccuracy in these areas where the redshift range was the same but the shape was different, it can be deduced that not only the range but also the shape of the redshift distributions must be the same to achieve accuracy. In this case, the redshift distribution of the testing sample is identical to the training sample over a restricted range, giving much more accurate results. This indicates that the shape of the redshift distribution must be similar or identical for the RF to perform well.

4.2.5 Overlap Tests Conclusion

The aim of Section 4.2 was to establish how a difference in the range and shape of the variables of two surveys affects the application of GALPRO. It is clear from the above results that as the percentage overlap of the r-band magnitude decreases between the testing and training surveys, the results become increasingly inaccurate. A 90% overlap in r-band magnitudes causes the RF to contain a small amount of bias, as shown in the PIT plot. This bias only increases as overlap decreases, which indicates that GALPRO may not be suitable for the estimation of photometric redshifts when applied to a 'new' survey of unknown depth and range. It is important to note that it is not the general difference in overlap percentage that introduces bias to the results, but specifically the percentage of the testing sample range that isn't covered by the training sample. It seems that the depth and properties of the training and testing surveys must be very similar for the RF to be applied and give trustworthy results, which is not usually possible when dealing with new catalogues.

As the overlap percentage decreases, the maximum redshift of the training sample decreases and the resulting predicted photometric estimates are limited to this maximum redshift. This demonstrates that GALPRO cannot extrapolate the mapping it has learnt between fluxes and redshift outside of the redshift range of the training sample. This means that any future use of GALPRO to predict photometric redshifts must be cautious to that fact that the RF can only predict within its learnt range. If GALPRO is trained using a certain training sample, any new, unknown survey must have redshift and colour distributions within the same range. This is obviously an issue, as the aim of this work is to apply GALPRO to a new survey which doesn't have associated redshift values, yet if this new survey had redshift values outside of the training dataset range, it would be unsuccessful. Due to this, GALPRO may not be suitable for generating redshift PDFs and point estimates for new photometry surveys, as we would be unsure as to whether the training sample is representative of the testing sample with regards to its flux and redshift distributions.

As the overlap percentage increases to 70%, even the galaxies with overlapping r-band magnitudes have inaccurate photometric predictions indicating that not only the range but also the shape of the variables must be similar or even identical to produce accurate results. As the overlap percentage decreases, the redshift distributions become increasingly dissimilar in shape, meaning that the mapping between the fluxes and redshift learnt by the RF are not applicable to both surveys. Case 4 demonstrates that GALPRO can be reliably applied when the testing sample has r-band magnitude ranges that are contained within the range of the testing sample, and reinforced that the RF may only produce accurate predictions when the redshift distributions of the testing and training samples are similar.

When applying GALPRO to an unknown survey, one can not always be certain that the new survey has spectroscopic redshifts contained within the ranges of the training sample due to the nature of the survey being unknown. However, it is possible to test the compatibility of the photometry of the new survey with the training data, meaning the application of GALPRO to a new survey is certainly feasible. Nevertheless, caution must be taken when applying the RF to new surveys, as training and testing samples must be statistically equivalent to produce accurate, reliable results.

4.3 Applying GALPRO to the PanSTARRS survey with statistically equivalent samples

Section 4.2 established that the training and testing datasets to which GALPRO is applied must be statistically equivalent in order to produce accurate photometric redshift estimates and PDFs that are marginally and probabilistically calibrated. Now attention turns back to the computation of redshift PDFs for the previously described PanSTARRS catalogue, to investigate further the question of whether GALPRO is applicable to this survey when trained using the full Zhou et al. dataset. As previously shown in Section 4.1, the RF, trained using the Zhou et al. dataset, cannot be straightforwardly applied to the PanSTARRS sample and accurate results produced. The aim of this section is to explore whether, taking into account the conclusions drawn from Section 4.2, any restrictions can be applied to the PanSTARRS data that can improve the redshift PDF results obtained for PanSTARRS.

The previous section made it clear that the two surveys must overlap by 90% to produce accurate and calibrated redshift estimates. The CDFs of the training and testing samples using the in the 90% overlap test gave KST statistics of around 0.2 for each of the bands. It is expected that if any two surveys have photometry such that all of the bands have at least the required 90% overlap in range, then a satisfactory results will be produced. When comparing the $W1$ and $W2$ bands from the PanSTARRS and Zhou et al. datasets, they KST statistics give values of around 1, and the other photometry bands produce even smaller results. It is clear that the PanSTARRS and Zhou et al. surveys overlap by over the required 90% in all of the photometric bands, meaning that the RF should be successful when trained using the Zhou et al. dataset and tested using the PanSTARRS survey. This indicates that there may be some underlying difference between the two surveys that cause the learnt mapping between the photometry and redshifts of one survey to be inapplicable to the other.

Firstly, it is useful to examine the spectroscopic redshift distributions of the two samples, shown in Figure 4.25. From these plots it is clear that the two datasets have similar redshift distributions, both with peaks around $z = 0.2$ and $z = 0.7$. They also both contain a smaller peak around $z = 0.5$, however the PanSTARRS sample has a deeper minimum at $z = 0.25$ and a steeper tail off at around $z = 1$. The Zhou sample reaches

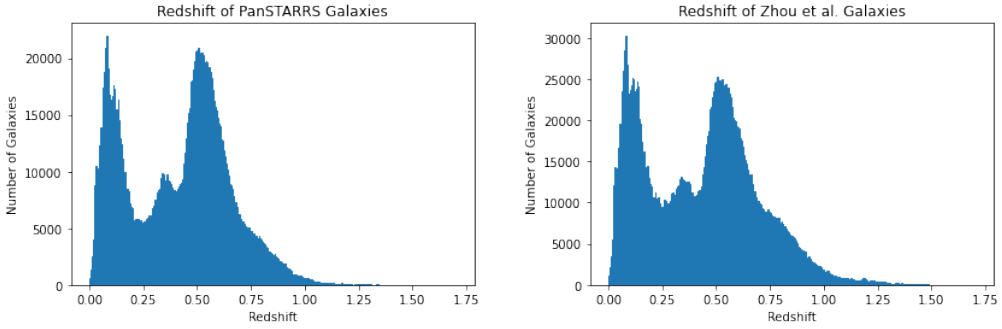


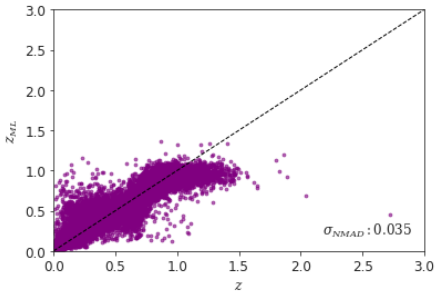
FIGURE 4.25: The redshift distributions of the PanSTARRS and Zhou et al. datasets. The plots demonstrate the similarity in the two redshift distributions.

higher redshifts of around $z = 1.5$ with a more gradual decline in population at these high redshifts. Despite these minor differences, the redshift distributions of the two samples are generally very similar, with the CDFs of both distributions being shown in Figure 4.3. It is clear from this plot that the two redshift distributions are very similar amd have a KST statistic of 0.0527. This means that the Zhou et al. and PanSTARRS datasets have more similar redshift distributions than the 90% overlap case considered in Section 4.2.3.1, which had a KST statistic of 0.198. Hence it seems unlikely that the shape of the redshift distribution in the PanSTARRS is contributing to the unsatisfactory results obtained for the photometric redshifts derived with GALPRO.

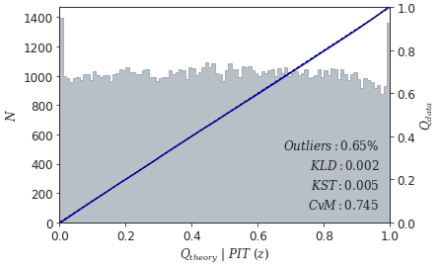
Despite the PanSTARRS and Zhou et al. datasets overlapping by over the required 90% in each photometry band, some discrepancy between the two surveys is causing the RF algorithm to fail. In theory, the two surveys should have enough statistical equivalence for the RF algorithm to be successful. Although all of the bands show over 90% overlap, the W bands still present less statistical similarities than the other bands, as shown by their larger KST statistic values. One final test to explore how this difference in photometry affects the application of the RF is to remove the W bands altogether from the analysis leaving only the g, r, z photometry, as these bands are almost completely equivalent between the two surveys.

As described in Section 3.2.5.1, GALPRO requires 4 inputted arrays, two of which contain the spectroscopic redshifts of the galaxies for the training and testing datasets. The other two arrays contain the r, $g-r$, $r-z$, $z-W1$, $W1-W2$ and their associated error columns for the testing and training datasets. These two arrays therefore have the shape [10,N] with N being the number of galaxies in each dataset. To eliminate any statistical difference between the two datasets, the columns containing any $W1$ and $W2$

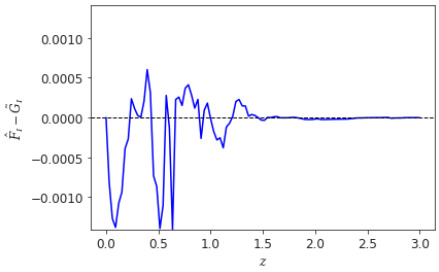
values were removed, with the goal of hopefully leading to more accurate and calibrated results. This means that the $z-W1$, $W1-W2$ and associated errors columns should be removed from both the training and testing input arrays, leaving them with the shape [6,N].



(A) Spectroscopic versus Photometric redshift plot



(B) PIT plot



(C) Marginal Calibration plot

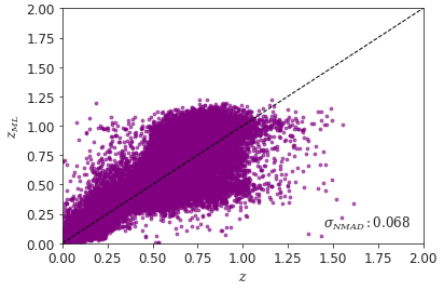
FIGURE 4.26: GALPRO results when trained and tested using the Zhou et al. dataset with the $W1$ and $W2$ columns omitted.

Obviously, the removal of these columns from the input arrays may affect the performance of the RF algorithm. How this removal affects the RF performance can be quantified using the Zhou et al dataset, which was done before applying this method to the PanSTARRS dataset to avoid the introduction of any other bias to the RF. Firstly, the Zhou et al. dataset is randomly sampled so that 150,000 galaxies are selected for testing and the rest for training, in the exact same manner as Section 3.2.1. The GALPRO settings/sampling/sample sizing is identical to those used in Section 3.2.5.1, which is already known to be successful and produce reliable redshift results. However, now the $z-W1$, $W1-W2$ and their associated error columns are removed from both the

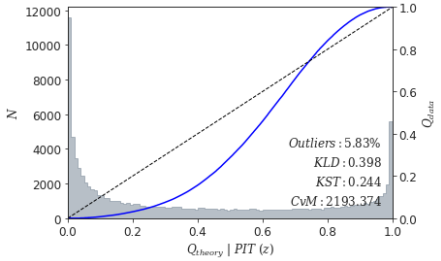
testing and training input arrays. The PIT, marginal calibration and spectroscopic versus photometric redshift estimate plots produced in this case can be seen in Figure 4.26. The PIT is very uniform, with excellent KST, KLD, CvM and outlier fraction results. The Q-Q plots show an almost identical match between the results and the U(0,1) distribution. The marginal calibration plot oscillates about the zero line with a maximum of around 0.0012. These plots both show that the PDFs are marginally and probabilistically calibrated and don't show any bias introduced. The scatter plot also shows a strong correlation between the spectroscopic and photometric values and has a value of $\sigma_{NMAD} = 0.035$, and is very similar to the scatter plot produced when the W columns are included (Section 3.2.2.1). This is encouraging, as the removal of the W columns does not appear to affect the performance of the RF.

With this reassurance, it is now time to apply GALPRO, trained using the Zhou et al dataset, to the PanSTARRS sample with the W columns removed from both the testing and training arrays. It is noted here that the PanSTARRS sample used has the K corrections applied, as described in Section 4.1.1. As before, the $z - W1$, $W1 - W2$ and their associated error columns are removed from the Zhou et al. training array and the PanSTARRS testing array. The hope is that the removal of the W columns will eliminate any significant statistical difference between the two samples, leading to accurate photometric redshift estimates and PDFs being produced for the galaxies in the PanSTARRS sample. GALPRO was run with the above inputs and the same hyperparameters described in Section 3.2.1.

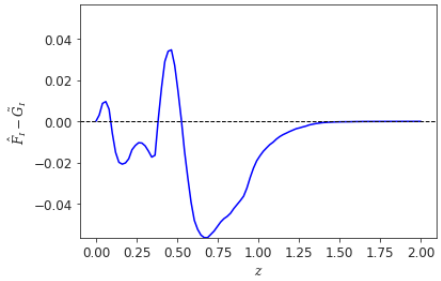
The PIT, marginal calibration and spectroscopic versus photometric redshift estimate plots produced by this test can be seen in Figure 4.27. Unfortunately, the results show that the RF was not successfully applied to the PanSTARRS testing sample. The PIT is extremely concavity, indicating that the PDFs are overly narrow by a significant amount. The CvM test gives a value of 2193.374 which reflects this concave, and the Q-Q plot deviates greatly from the uniform distribution. The marginal calibration does oscillate about the zero line, but in a very smooth manner and peaks at around 0.06, meaning marginal calibration has not been successful. The spectroscopic versus photometric redshift plot does follow the diagonal, unlike previous unsuccessful tests where there is no correlation between the spectroscopic and photometric values. However, the scatter either side of the diagonal is significant and gives a σ_{NMAD} value of 0.068. There is no significant cut off in the scatter plot as the redshift range of the training sample



(A) Spectroscopic versus Photometric redshift plot



(B) PIT plot



(C) Marginal Calibration plot

FIGURE 4.27: GALPRO results when trained using the Zhou et al. dataset and tested with the PanSTARRS sample with the $W1$ and $W2$ columns omitted.

encapsulates the redshift range of the PanSTARRS sample. This scatter does not have a strong enough correlation for the photometric estimates to be defined as accurate, and the PIT and marginal calibration graphs show that the PDFs are not probabilistically or marginally calibrated.

These results are discouraging, as it seems that GALPRO may not be trained on the Zhou et al. dataset and applied to the PanSTARRS sample, no matter how statistically equivalent the two are. The overlap tests, described in Section 4.2, show that a discrepancy in the range of the photometry bands introduced a bias to the PDFs, as shown by the gradient present in the PIT plot. Here, the PIT does not have a uniform gradient but is a concave shape, meaning that the PDFs are overly narrow by a significant amount. Since the PIT here is concave and not a gradient, this shows that the inaccuracy is not related to the overlap (or lack thereof) of the bands introducing a bias.

Both surveys have very similar redshift distributions and this final test assured that they also had consistent photometric bands, which should have been enough to produce reliable results, as demonstrated in Section 4.2. GALPRO only requires the photometry and spectroscopic redshifts of the two surveys as input variables, and all of these variables have been assessed for consistency, yet the RF algorithm still doesn't give accurate results. Any obvious statistical difference between the two surveys, such as the differing K-corrections or magnitude definitions, has been dealt with accordingly which points towards only one conclusion: There is some underlying statistical difference between the Zhou et al. and PanSTARRS datasets, causing the mapping between the fluxes and redshifts to differ between the two surveys. The catastrophic deviation of the PIT from the uniform indicates that there is not some small issue regarding miss-calibration but an underlying inherent difference between the surveys. This difference may be attributed to the fact that the common galaxies provided by both surveys with the same sky location give different magnitude values.

It is important to remember here that the PanSTARRS and Zhou surveys act as two general surveys, one with known spectroscopic data and one only containing photometry. This serves as a cautionary tale, as it demonstrates that no matter how statistically equivalent the two surveys may be, there can exist some underlying difference that means that an RF is not applicable for the estimation of photometric redshift PDFs. One must proceed with great caution when hoping to apply an RF to two different surveys, as no matter how similar they may be, the results can be far from perfect. The range in band magnitudes and redshift appear very similar, however the two surveys differ as the galaxies with the same sky location do not give the same magnitude values. This may be the reason behind the catastrophic failure of the algorithm. When applying GALPRO to a new, unseen survey, one cannot be certain that the surveys do not contain systematic differences, and so a more malleable software may be required to perform this task. GALPRO may not be a suitable choice for the generation of redshift PDFs to be included in the inference of H_0 using gravitational wave data, as even after thorough evaluation the PDFs cannot be considered as reliable. This inference of H_0 may be pivotal in the world of cosmology and any contributing factors in its measurement must have an assured accuracy and reliability before being used in the measurement process. Unfortunately, GALPRO cannot provide this reliability and therefore should not be considered as choice for the generation of redshift PDFs for this purpose.

Chapter 5

Conclusion

The goal of this work has been to investigate an RF-based approach, GALPRO, to estimate photometric redshifts from galaxy survey data, and to evaluate whether GALPRO can be reliably applied to new surveys containing only photometry, i.e. without spectroscopic redshifts that can be used directly for testing. Photometric redshifts derived in this way may then be used to populate the galaxy catalogues utilised by the software gwcosmo in the inference of the Hubble constant. The intention was that GALPRO may be trained using a trusted dataset and applied to an unknown survey to compute redshift PDFs that may then be used alongside GW data to make a constraining measurement of H_0.

However, it was found GALPRO could not be successfully applied to new catalogues when trained on a trusted dataset as the mapping learnt between the colours and redshift of that trusted dataset could not be applied to a new catalogue, despite having verified the statistical similarity of the two, because the mapping learnt between the colours and redshifts for the trusted dataset was found not to be applicable to the new catalogue. Thus, even when the statistical equivalence of the GALPRO input data has been demonstrated, there may still exist some inherent difference between the two surveys that causes the application of GALPRO to the new survey to fail. From this, it can be concluded that GALPRO may not be reliable when applied to a new survey, and should not be implemented in the `gwcosmo` analysis pipeline unless spectroscopic training data is available for each new survey under consideration.

GALPRO was initially calibrated using a known, trustworthy sample containing around 3 million objects compiled by Zhou et al. to produce marginally and probabilistically calibrated redshift PDFs between $0 < z < 1.5$. The sample was split randomly such that 90% is used for training and the other 10% for testing. The spectroscopic versus photometric redshift plot demonstrated a strong correlation between photometric and true values, indicating that the estimates were indeed reliable. Above a spectroscopic redshift value of around $z = 1.5$, GALPRO tended to underestimate the photometric values as the training sample did not contain as many objects at higher redshifts. This highlights the need for representative training samples when using an RF algorithm for predictions. The calibration and performance of the redshifts is assessed using the probability integral transforms (PIT) of the redshift PDFs and it was found that the outlier percentage and uniformity of the PIT improved as the photometry of the training sample is scattered. This scattered sample produced accurate and reliable redshift PDFs which could then be used as a trustworthy training sample to explore if GALPRO could be applied to new, unknown photometry surveys. The software GALPRO is advantageous in that it can produce photometric redshift posterior distributions alongside redshift point estimates, which many ML softwares are incapable of doing. The computational expense of training the algorithm and calibrating the PDFs is quite high, however once this has been executed, the expense of generating the redshift PDFs is much more reasonable.

The sample described above generated reliable redshift PDFs over a redshift range which is representative of the type of galaxies used to infer H_0 from GW dark sirens. This allowed for the assessment of whether the current assumption implemented in the `gwcosmo` code, that photometric redshift estimate errors can be assumed as Gaussian, is valid. The redshift posteriors computed using the above testing sample were each evaluated using the D'Agostino's K-squared test to determine how non-Gaussian the PDFs are. It was found that out of 101945 galaxies, 78241 galaxies had a non-Gaussian redshift PDF and 23704 were found to follow the Gaussian distribution. This demonstrates that the majority of the galaxies in the testing sample have redshift PDFs that do not follow a normal distribution. This highlights the need for accurate redshift PDF generation to populate galaxy surveys, as the current assumption that the errors are Gaussian is obviously not satisfied. A more accurate representation of the redshift posteriors in future galaxy surveys used for GW cosmology analyses could lead to less biased and more accurate results in the overall inference of H_0.

The above results make it clear that there is a need for methodology to accurately generate photometric redshift PDFs and calls for these methods to be applicable to new, unknown surveys that only contain photometry data in order for it to be useful. To assess how well the random forest works for new surveys for which we only have photometry available, the PanSTARRS galaxy catalogue was chosen as the testing sample and acted as a 'new' survey. The PanSTARRS catalogue does have a subsample of objects with cross-matched spectroscopic redshifts, which are selected as the testing sample for validation purposes. The PanSTARRS survey acts as a sanity check and is used to represent other examples of redshift surveys for which we only have photometry available.

It was carefully checked that the PanSTARRS and Zhou et al samples had consistent ranges and distributions of data, including compatible K corrections and magnitude definitions. Despite these checks, however, when the Zhou et al. sample is used for training and the PanSTARRS for testing, the results showed general inaccuracy and unsuccessful calibration. This indicates that there is some residual difference between the statistical properties of the two surveys, meaning that the mapping learnt between the colours and redshifts of the Zhou et al. sample is not applicable to the PanSTARRS sample. The two surveys were cross-matched by sky location and it was found that common galaxies have largely differing band magnitudes between the two samples. This is alarming, however it demonstrates that generally, two photometric surveys may contain systematic differences. GALPRO is perhaps not malleable enough to cope with the difference between the two surveys, which should be taken in account with it's use.

To explore how statistically equivalent the training and testing samples must be in order for the RF to be applicable, the Zhou et al. sample was used as a means to gauge how much overlap of statistical properties is needed for GALPRO to give reliable results. The sample was split in half to form two artificial 'new' surveys with varying degrees of overlap in the r-band magnitude range. Analysis was run to determine how the overlap in colour bands affects the performance of the algorithm.

It was found that even if the two surveys overlapped in the r-band magnitude range by 90%, the results contained a small amount of bias in comparison to the initial calibration tests. As the overlap percentage decreased to 80% and 70%, the accuracy of results deteriorated dramatically. An overlap of 70% in the r-band magnitude range showed

little to no correlation between photometric and spectroscopic values and a catastrophic failure of the probabilistic calibration of the PDFs. The tests also indicated that not only the range of colour bands, but also the shape of the redshift distributions must be very similar for the RF to be applicable between the two surveys. This demonstrates that GALPRO may not be suitable for the estimation of photometric redshifts when applied to a 'new' survey of unknown (but likely greater) depth and range. The depth and properties of the training and testing surveys must be almost identical for the RF trained on one survey to be applied to the other survey and give robust and accurate results, which is not usually possible when dealing with unknown catalogues. The tests also show that GALPRO is unable to predict redshift estimates greater than those of the maximum redshift of the training sample. This demonstrates that GALPRO cannot extrapolate the mapping it has learnt between fluxes and redshift outside of the redshift range of the training sample. This means that any future use of GALPRO to predict photometric redshifts must take account of the fact that the RF can only predict within its learnt range.

It was demonstrated that the PanSTARRS and Zhou et al. samples overlapped in each photometry band and in redshift distribution by over the required 90%, meaning there should be no reason why the mapping between the photometry and redshift would differ between the two samples. The inputs from the two surveys were now almost completely statistically equivalent, with the ranges in all colour bands and redshift distribution having over 90% overlap. The above described tests demonstrated that this overlap should in principle be sufficient to generate accurate redshift estimates. However, when GALPRO was trained using the reliable Zhou sample and tested using the PanSTARRS survey, the results still showed large amounts of inaccuracy in the derived photometric redshifts. The photometric estimate versus spectroscopic redshift plot showed a relatively large scatter indicating that the estimates were not accurate. The PIT was catastrophically non-uniform, showing that the redshift PDFs were much too narrow to be considered accurate.

Although the statistical equivalence of the relevant GALPRO data for each survey was verified, GALPRO was still unable to generate accurate and reliable results when applied to a new survey. This work, therefore, serves as a cautionary tale about the dangers of applying RF algorithms to new, unknown galaxy surveys. Even if the two surveys

appear to be statistically almost equivalent, in terms of the variables with which GALPRO constructs the mapping from photometry to redshifts, there may still exist some underlying, fundamental difference between the two surveys that is not apparent in the photometry or spectroscopic redshift distributions of the surveys. Although a thorough evaluation of the photometric and spectroscopic distributions has been performed, there exists some unknown, elementary difference between these two samples that hinders the performance of the RF.

Hence, in conclusion, GALPRO is shown to be potentially unsuitable for generating estimates of photometric redshifts and their PDFs from new surveys. It may be useful in the case where a catalogue is nearly complete, yet is missing some spectroscopic values, as it has been shown that GALPRO can be reliable when trained and tested using the same survey. However, it is clear here that the algorithm is unable to extrapolate the learnt mapping between the colours and redshift for new surveys or even redshift values beyond which it has been trained on. This acts as a warning, not only for GALPRO but also the future use of any RF algorithm to generate photometric redshifts to be used in astronomical analysis.

In the broader context of the inference of H_0 using gravitational data, it is clear that there is still a need for an accurate method to estimate redshift posteriors which can be used in the measurement of H_0. This work has made it apparent that the current assumption that the redshift errors are Gaussian is inaccurate and to obtain a better constraint on H_0 will require moving beyond the Gaussian assumption to a more precise description of the redshift PDF for each galaxy. However, it seems that GALPRO is not ideally suited for this purpose as it cannot produce reliable photometric PDFs when applied to a new survey that differs from the training dataset, even when restrictions are applied to the input data to ensure statistical equivalence. The PanSTARRS survey, which is to be used in the upcoming fourth LVK observational period, does contain some spectroscopic redshift data and so there is not an urgent need for photometric redshift generation techniques. However, future observational runs and new photometry surveys will call for accurate, calibrated photometric techniques to populate galaxy catalogues and further constrain H_0.

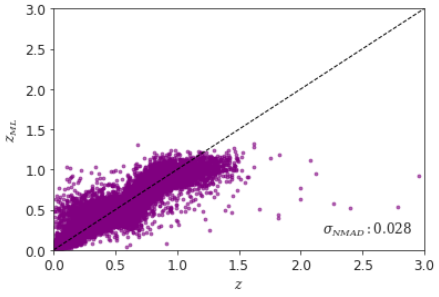
For now, the LVK collaboration will continue to use both spectroscopic data and search for other ways in which we can access information about non-Gaussian PDFs, as this

work has made it clear than in general, photometric redshifts are not Gaussian. A true assessment of the photometric error would of course lead to a more accurate inference of H_0, however for the time being, this assumption will do. Future work may include exploring the application of other ML algorithms, such as neural networks, for the purpose of generating redshift PDFs that can quantify the non-Gaussian photometric errors to lead to a further constraint of H_0. It could also be useful for the fundamental, underlying difference between the two surveys that made GALPRO inapplicable to be identified. As previously stated, the issue was not in fact the photometry or redshift values/range. Some rigorous test that identifies this difference may not only be useful to the application of RF algorithms to galaxy surveys but could give deeper insight into the use of galaxy catalogues as a whole.

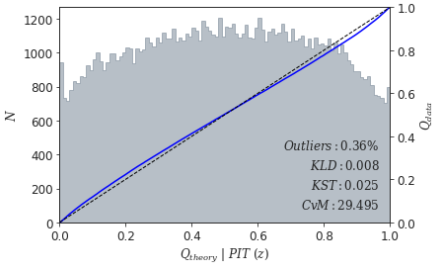
Although the application of GALPRO was unsuccessful for this purpose, there is a bright future ahead for new and exciting photometric methods and an ever brighter future for cosmology as a whole. Further constraints of the Hubble constant may bring about new physics entirely or at least shed light on the systematics involved in the measurement process of our universe around us. Either way, the constraint of H_0 will bring us one step closer to understanding our universe. In the great words of Plato, "Astronomy compels the soul to look upward, and leads us from this world to another".

Appendix A

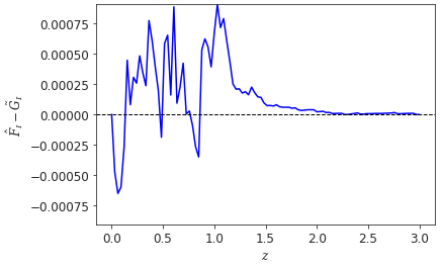
Appendix



(A) Spectroscopic versus Photometric redshift plot



(B) PIT plot



(C) Marginal Calibration plot

FIGURE A.1: The PIT, marginal calibration and spectroscopic versus photometric redshift plots produced by GALPRO when trained and tested using the Zhou et al. truth dataset with min_leaf_sample = 5.

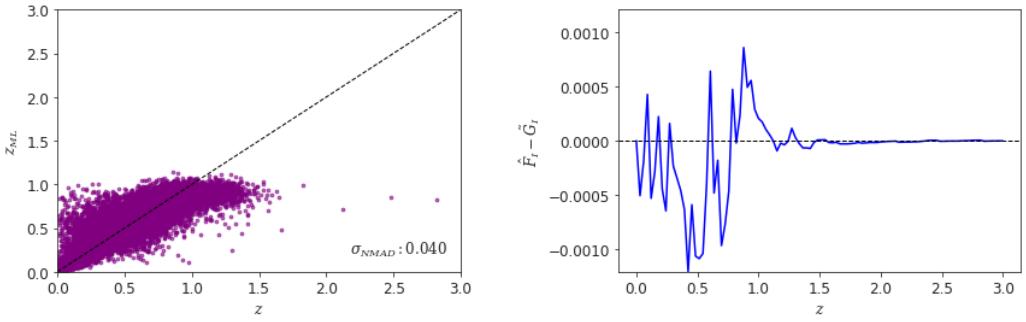


FIGURE A.2: The marginal calibration and spectroscopic versus photometric redshift plots produced by GALPRO when trained and tested using the Zhou et al. truth dataset with the photometry scattered.

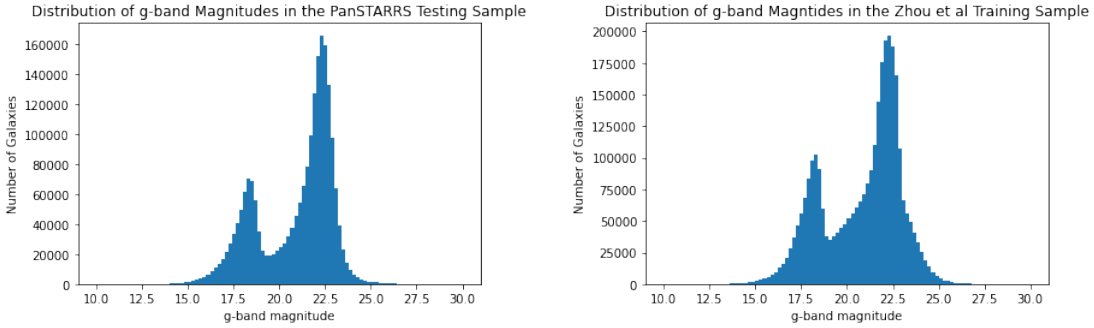


FIGURE A.3: The g-band distributions of the PanSTARRS and Zhou et al. surveys which are used to compute the cumulative distributions functions.

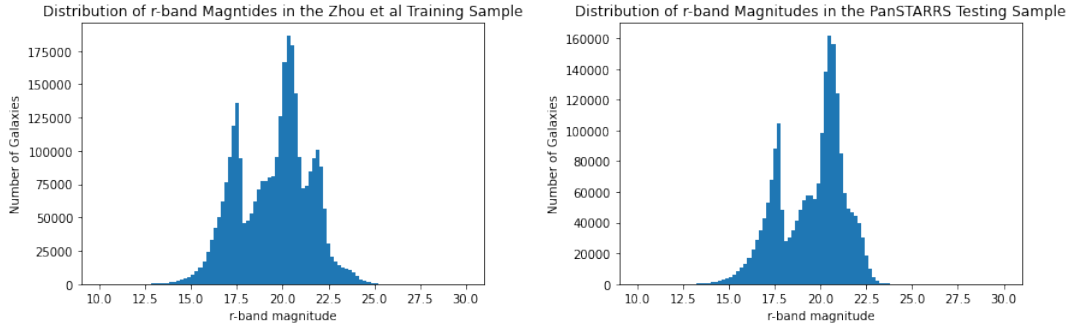


FIGURE A.4: The r-band distributions of the PanSTARRS and Zhou et al. surveys which are used to compute the cumulative distributions functions.

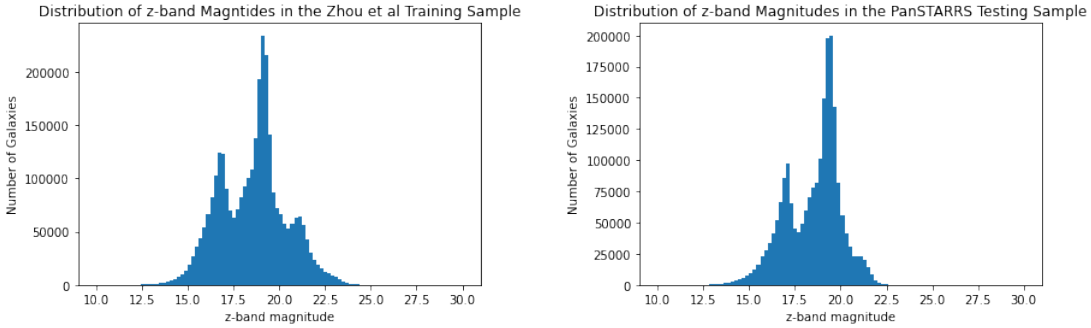


FIGURE A.5: The z-band distributions of the PanSTARRS and Zhou et al. surveys which are used to compute the cumulative distributions functions.

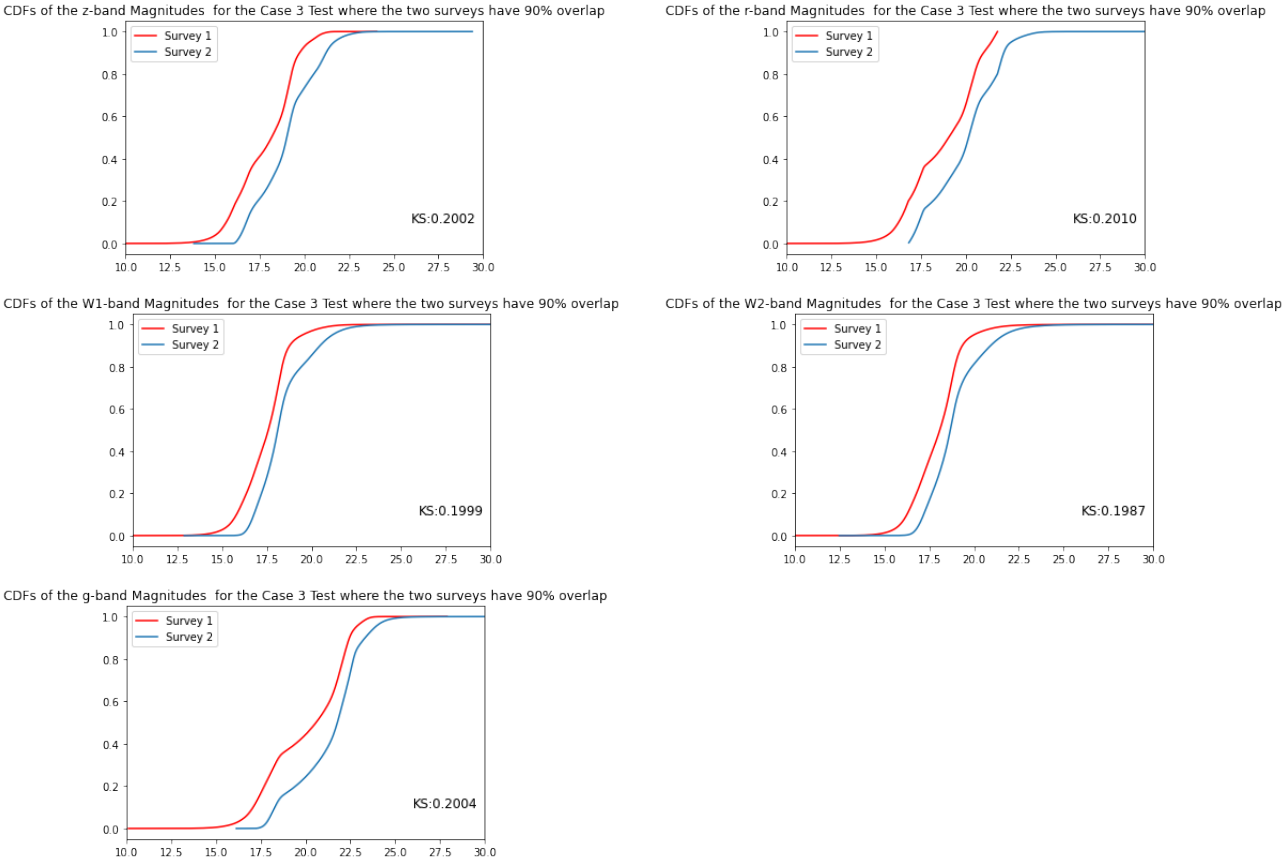


FIGURE A.6: The CDFs of the $g, r, z, W1$ and $W2$ bands for the training and testing surveys used in the 90% overlap tests.

Bibliography

[1] D. Darling. The cosmic distance ladder. URL ttps://www.daviddarling.info/encyclopedia/C/cosmic_distance_ladder.html.

[2] M Bovin, V. Millon. URL https://github.com/vbonvin/H0_tension.

[3] The 2df galaxy redshift survey: spectra and redshifts. (328(4)). doi: 10.1046/j.1365-8711.2001.04902.

[4] ogrisel.github.io. (n.d.). Regression: Photometric redshifts of galaxies — scikit-learn 0.11-git documentation. URL https://ogrisel.github.io/scikit-learn.org/sklearn-tutorial/tutorial/astronomy/regression.html.

[5] C. Bakshi. Random forest regression. URL https://levelup.gitconnected.com/random-forest-regression-209c0f354c84.

[6] R. Zhou. The clustering of desi-like luminous red galaxies using photometric redshifts. *Monthly Notices of the Royal Astronomical Society*, 501(3):3309–3331, March 2021. doi: 10.1093/mnras/staa3764.

[7] V. M. Slipher. The radial velocity of the andromeda nebula. *Lowell Observatory Bulletin*, (1):56–57, 1913. doi: 10.1103/PhysRevD.53.2878.

[8] E. F. Bunn. The kinematic origin of the cosmological redshift. *American Journal of Physics*, (77(8)):688–694, 2009. doi: 10.1119/1.3129103.

[9] A. Friedmann. Über die krümmung des raumes. *Zeitschrift fur Physik*, (10):377–386, January 1922. doi: 10.1007/BF01332580.

[10] G. Lemaître. Un univers homogène de masse constante et de rayon croissant rendant compte de la vitesse radiale des nébuleuses extra-galactiques. *Annales de la Société Scientifique de Bruxelles*, (47):49–59, January 1927.

[11] A relation between distance and radial velocity among extra-galactic nebulae. (15(3)). doi: 10.1073/pnas.15.3.168.

[12] Planck Collaboration. et al. Planck 2018 results. vi. cosmological parameters. *Astron. Astrophys*, (641:A6), September 2020. doi: 10.1051/0004-6361/201833910.

[13] A. G. Riess et al. Cosmic distances calibrated to 1% precision with gaia edr3 parallaxes and hubble space telescope photometry of 75 milky way cepheids confirm tension with cdm. *Astrophys. J*, (908(1)), February 2021. doi: 10.3847/2041-8213/abdbaf.

[14] T. Verde, L. Treu and A. G. Riess. Tensions between the early and late universe. *Nature Astronomy*, (3):891–895, September 2019. doi: 10.1038/s41550-019-0902-0.

[15] E. D. Valentino. letter of interest cosmology intertwined ii: The hubble constant tension. *Astroparticle Physics*, (131:102605), 2021. doi: https://doi.org/10.1016/j.astropartphys.2021.102605.

[16] R Gray. *Gravitational wave cosmology: measuring the Hubble constant with dark standard sirens.* PhD thesis, University of Glasgow, 2021.

[17] A. Friedmann. Determining the hubble constant from gravitational wave observations. *Nature*, (323:310), September 1986. doi: 10.1038/323310a0.

[18] B. P. Abbott et al. Gw170817: Observation of gravitational waves from a binary neutron star inspiral. *Phys. Rev. Lett.*, (119:161101), October 2017. doi: 10.1103/PhysRevLett.119.161101.

[19] B. P. Abbott et al. A gravitational-wave standard siren measurement of the hubble constant. *Nature*, (551):85–88, November 2017. doi: 10.1038/nature24471.

[20] A Liddle. *An Introduction to Modern Cosmology.* John Wiley Sons, 2015. ISBN 9780198520115.

[21] J.-M. Goff and V. Ruhlmann-Kleider. Cosmological constraints from baryonic acoustic oscillation measurements. *Scholarpedia*, (10(9)):32149, 2015. doi: 10.4249/scholarpedia.32149.

[22] A. Friedmann. Über die möglichkeit einer welt mit konstanter negativer krümmung des raumes. *Zeitschrift fur Physik*, (21(1)):326–332, December 1924. doi: 10.1007/BF01328280.

[23] G. Lemaître. L'univers en expansion. *Annales de la Société Scientifique de Bruxelles*, (53:51), January 1933.

[24] H. P. Robertson. Kinematics and world-structure. *Astrophys. J*, (82:284), November 1935. doi: 10.1086/143681.

[25] H. P. Robertson. Kinematics and world-structure ii. *Astrophys. J*, (83:187), April 1936. doi: 10.1086/143716.

[26] H. P. Robertson. Kinematics and world-structure iii. *Astrophys. J*, (83:257), May 1936. doi: 10.1086/143726.

[27] A. G. Walker. On milne's theory of world-structure. proceedings of the london mathematical society. (42):90–127, January 1937. doi: 10.1112/plms/s2-42.1.90.

[28] M. Bolte D.N. Spergel and W. Freedman. The age of the universe. 94, 1997.

[29] S. Aiola et al. The atacama cosmology telescope: Dr4 maps and cosmological parameters. *Journal of Cosmology and Astroparticle Physics*, (2020(12)):047–047, December 2020. doi: 10.1088/1475-7516/2020/12/047.

[30] G. E. Addison. Elucidating cdm: Impact of baryon acoustic oscillation measurements on the hubble constant discrepancy. *Astrophys. Jk*, (853(2):119), January 2018. doi: 10.3847/1538-4357/aaa1ed.

[31] J. Lemos P. Cuceu, A. Farr and A. Font-Ribera. Baryon acoustic oscillations and the hubble constant: past, present and future. *Zeitschrift fur Physik*, (2019(10)): 44, October 2019. doi: 10.1088/1475-7516/2019/10/044.

[32] A. G. Riess et al. Large magellanic cloud cepheid standards provide a 1% foundation for the determination of the hubble constant and stronger evidence for physics beyond cdm. *Astrophys. J*, (876(1):85), 2019. doi: 10.3847/1538-4357/ab1422.

[33] D. W. Pesce et al. The megamaser cosmology project. xiii. combined hubble constant constraints. *Astrophys. J*, (891(1):L1), February 2020. doi: 10.3847/2041-8213/ab75f0.

[34] W. L. Freedman et al. The carnegie-chicago hubble program. viii. an independent determination of the hubble constant based on the tip of the red giant branch. *Astrophys. J*, (882(1):34), August 2019. doi: 10.3847/1538-4357/ab2f73.

[35] W. Yuan et al. Consistent calibration of the tip of the red giant branch in the large magellanic cloud on the hubble space telescope photometric system and a redetermination of the hubble constant. *Astrophys. J*, (886(1):61), November 2019. doi: 10.3847/1538-4357/ab4bc9.

[36] Planck Collaboration et al. Planck 2015 results. xiv. dark energy and modified gravity. *Astron. Astrophys.*, (594:A14), September 2016. doi: 10.1051/0004-6361/201525814.

[37] A. Di Valentino, E. Melchiorri and J. Silk. Cosmological hints of modified gravity? *Phys. Rev. D*, (93(2):023513), January 2016. doi: 10.1103/PhysRevD.93.023513.

[38] A. Mena O. Di Valentino, E. Melchiorri and S. Vagnozzi. Interacting dark energy in the early 2020s: A promising solution to the h0 and cosmic shear tensions. *Physics of the Dark Universe*, (30:100666), December 2020. doi: 10.1016/j.dark.2020.100666.

[39] F. Kreisch, C. D. Cyr-Racine and O. Doré. Neutrino puzzle:anomalies, interactions, and cosmological tensions. *Phys. Rev. D*, (101:123505), June 2020. doi: 10.1103/PhysRevD.101.123505.

[40] A. Di Valentino, E. Melchiorri and J. Silk. Planck evidence for a closed universe and a possible crisis for cosmology. *Nature Astronomy*, (4):196–203, February 2020. doi: 10.1038/s41550-019-0906-9.

[41] W. Handley. Curvature tension: Evidence for a closed universe. *Phys. Rev. D*, (10.1103/PhysRevD.103.L041301), February 2021. doi: 10.1007/BF01332580.

[42] A. Einstein. Näherungsweise integration der feldgleichungen der gravitation. *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften*, pages 688–697, January 1916.

[43] A. Einstein. Die feldgleichungen der gravitation. *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften*, pages 844–847, January 1915.

[44] F. A. Pirani. Invariant formulation of gravitational radiation theory. *Phys. Rev.*, (105(3)):1089–1099, February 1957. doi: 10.1103/PhysRev.105.1089.

[45] R. A. Hulse and J. H. Taylor. Discovery of a pulsar in a binary system. , 195: L51–L53, January 1975. doi: 10.1086/181708.

[46] B. P. Abbott et al. Observation of gravitational waves from a binary black hole merger. *Phys. Rev. Lett.*, (116):6–12, February 2016.

[47] L. S. Finn and D. F. Chernoff. Observing binary inspiral in gravitational radiation: One interferometer. *Phys. Rev.*, (D47):2198–2219, 1993. doi: 10.1103/PhysRevD.47.2198.

[48] P. R. Creighton J. D. Anderson, W. G. Brady and E. E. Flanagan. Excess power statistic for detection of burst sources of gravitational radiation. *Phys. Rev. D*, (63(4):042003), February 2001. doi: 10.1103/PhysRevD.63.042003.

[49] J. C. Usman, S. A. Mills and S. Fairhurst. Constraining the inclinations of binary mergers from gravitational-wave observations. *Astrophys. J.*, (877(2)):82, June 2019. doi: 10.3847/1538-4357/ab0b3e.

[50] S. Fairhurst. Triangulation of gravitational wave sources with a network of detectors. *New Journal of Physics*, (11(12):123006), December 2009. doi: 10.1088/1367-2630/11/12/123006.

[51] B Schutz. Determining the hubble constant from gravitational wave observations. *Nature*, (323), 1986. doi: 10.1038/323310a0.

[52] W. Del Pozzo. Inference of cosmological parameters from gravitational waves: Applications to second generation interferometers. *Phys. Rev. D*, (86(4):043011), August 2012. doi: 10.1103/PhysRevD.86.043011.

[53] Li L. X. and B. Paczynski. Transient events from neutron star mergers. *Astrophys. J. Lett.*, (507(1)):L59–L62, November 1998. doi: 10.1086/311680.

[54] B. D. Metzger et al. Electromagnetic counterparts of compact object mergers powered by the radioactive decay of r-process nuclei. *Mon. Not. R. Astron. Soc*, (406(4)):2650–2662, August 2010. doi: 10.1111/j.1365-2966.2010.16864.x.

[55] W. Fong and E. Berger. The locations of short gamma-ray bursts as evidence for compact object binary progenitorss. *Astrophys. J*, (776(1):18), October 2013. doi: 10.1088/0004-637X/776/1/18.

[56] D. E. Hughes S. A. Dalal, N. Holz and B. Jain. Short grb and binary black hole standard sirens as a probe of dark energy. *Phys. Rev.*, (D74:063006), 2006. doi: 10.1103/PhysRevD.74.063006.

[57] B. F. Sathyaprakash, B. S. Schutz and C. Van Den Broeck. Cosmography with the einstein telescope. *Class. Quant. Grav*, (27:215006), 2010. doi: 10.1088/0264-9381/27/21/215006.

[58] S. Nissanke et al. Exploring short gamma-ray bursts as gravitational-wave standard sirens. *Astrophys. J.*, (725):496–514, 2010. doi: 10.1088/0004-637X/725/1/496.

[59] S. Nissanke et al. Determining the hubble constant from gravitational wave observations of merging compact binaries. 2013.

[60] B. P. Abbott et al. Multi-messenger observations of a binary neutron star merger. *Astrophys. J.*, (848:L12), October 2017. doi: 10.3847/2041-8213/aa91c9.

[61] M. Soares-Santos et al. The electromagnetic counterpart of the binary neutron star merger ligo/virgo gw170817. i. discovery of the optical counterpart using the dark energy camera. *Astrophys. J. Lett*, (848:L16), October 2017. doi: 10.3847/2041-8213/aa9059.

[62] S. Mukherjee et al. Velocity correction for hubble constant measurements from standard sirens. *Astron, Astrophys.*, (646:A65), 2021. doi: 10.1051/0004-6361/201936724.

[63] C. Nicolaou et al. The impact of peculiar velocities on the estimation of the hubble constant from gravitational wave standard sirens. *Mon. Not. R. Astron. Soc*, (495(1)):90–97, 2020. doi: 10.1093/mnras/staa1120.

[64] C. Guidorzi et al. Improved constraints on h0 from a combined analysis of gravitational-wave and electromagnetic emission from gw170817. *Astrophys. J.*, (851(2):L36), 2017. doi: doi:10.3847/2041-8213/aaa009.

[65] M. Chen, H. Y. Fishbach and D. E. Holz. A two per cent hubble constant measurement from standard sirens within five years. *Nature*, (562(7728)):545–547, October 2018. doi: 10.1038/s41586-018-0606-0.

[66] S. M. Feeney et al. Prospects for resolving the hubble constant tension with standard sirens. *Phys. Rev. Lett.*, (122(6):061105), 2019. doi: 10.1103/PhysRevLett.122.061105.

[67] R. Gray et al. Cosmological inference using gravitational wave standard sirens: A mock data analysis. *Phys. Rev. D*, (101(12):122001), 2020. doi: 10.1103/PhysRevD.101.122001.

[68] B. P. Abbott et al. Gwtc-1: A gravitational-wave transient catalog of compact binary mergers observed by ligo and virgo during the first and second observing runs. *Phys. Rev*, (9(3):031040), 2019. doi: 10.1103/PhysRevX.9.031040.

[69] B. P. Abbott et al. A gravitational-wave measurement of the hubble constant following the second observing run of advanced ligo and virgo. *Astrophys. J.*, (909(2):218), March 2021. doi: 10.3847/1538-4357/abdcb7.

[70] R. Abbott et al. Gwtc-2: Compact binary coalescences observed by ligo and virgo during the first half of the third observing run. *Phys. Rev. X*, (11:021053), June 2021. doi: 10.1103/PhysRevX.11.021053.

[71] C. Messenger and J. Read. Measuring a cosmological distance-redshift relationship using only gravitational wave observations of binary neutron star coalescences. *Phys. Rev. Lett.*, (108:091101), 2012. doi: 10.1103/PhysRevLett.108.091101.

[72] L. S. Finn. Observational constraints on the neutron star mass distribution. *Phys. Rev. Lett.*, (73):1878–1881, 1994. doi: 10.1103/PhysRevLett.73.1878.

[73] J. R. Taylor, S. R. Gair and I. Mandel. Hubble without the hubble: Cosmology using advanced gravitational-wave detectors alone. *Phys. Rev.*, (D85:023535), 2012. doi: 10.1103/PhysRevD.85.023535.

[74] S. R. Taylor and J. R. Gair. Cosmology with the lights off: standard sirens in the einstein telescope era. *Phys. Rev.*, (D86:023502), 2012. doi: 10.1103/PhysRevD.86.023502.

[75] B. P. Abbott et al. Gw190425: Observation of a compact binary coalescence with total mass 3.4 $m_{\odot}$. *Astrophys. J. Lett*, (892(1):L3), March 2020. doi: 10.3847/2041-8213/ab75f5.

[76] M. Ye J. Farr, W. M. Fishbach and D. Holz. A future percent-level measurement of the hubble expansion at redshift 0.8 with advanced ligo. *Astrophys. J. Lett*, (883:L42), October 2019. doi: 10.3847/2041-8213/ab4284.

[77] S.S. Vasylyev and A.V. Filippenko. A measurement of the hubble constant using gravitational waves from the binary merger gw190814t. *The Astrophysical Journal*, (902(2)):149, 2010. doi: 10.3847/1538-4357/abb5f9.

[78] Jurgen Freund. *Special Relativity for Beginners*. World Scientific, 2008. ISBN 120. ISBN 978-981-277-160-5.

[79] Lanzetta K.M. Chen H. Pascarelle S.M. Fernandez-Soto, A. and N. Yahata. On the compared accuracy and reliability of spectroscopic and photometric redshift measurements. *The Astrophysical Journal Supplement Series*, (135(1)):41–61, 2001. doi: 10.1086/321777.

[80] N. Benitez. Bayesian photometric redshift estimation. *Astrophys. J.*, (536(2)): 571–583, 2000. doi: 10.1086/308947.

[81] A. Bergström, L. Ariel Goobar. *Cosmology and Particle Astrophysics (2nd ed.)*. Springer, 2006. ISBN 978-3-540-32924-4.

[82] M. Carrasco Kind and R.J. Brunner. Tpz: photometric redshift pdfs and ancillary information by using prediction trees and random forests. *Monthly Notices of the Royal Astronomical Society*, 432(2):1483–1501, 2013. doi: 10.1093/mnras/stt574.

[83] O. Salvato, M. Ilbert and B. Hoyle. The many flavours of photometric redshifts. *Nature Astronomy*, 3(3):212–222, 2018. doi: 10.1038/s41550-018-0478-0.

[84] W. A. Baum. Problems of extra-galactic research. *IAU Symposium*, (15):390, 1962.

[85] M. Fernandez-Soto, A. Kenneth Lanzetta and A. Yahil. A new catalog of photometric redshifts in the hubble deep field1. *Astrophys. J.*, (513):34–50, 1999. doi: 10.1007/s11749-016-0481-7.

[86] Cooper M. Davis M. Faber S. Coil A. Guhathakurta P. Koo-D. Phillips A. Conroy C. Dutton A. Finkbeiner D. Gerke B. Rosario D. Weiner B. Willmer C. Yan R. Harker J. Kassin S. Konidaris N. Newman, J. and K. Lai. The deep2 galaxy redshift survey: Design, observations, data reduction, and redshifts. 2012. URL https://arxiv.org/pdf/1203.3192.pdf.

[87] T. Heinis S. Priebe C. Carliles, S. Budavári and A. S. Szalay. Random forests for photometric redshifts. *Astrophys. J.T*, (712(2)), 2010. doi: 10.1088/0004-637X/712/1/511.

[88] G. Biau and E. Scornet. A random forest guided tour. *TEST*, (25(2)):197–227, 2016. doi: 10.1007/s11749-016-0481-7.

[89] F. Marocco et al. The catwise2020 catalog. *Astrophy. J. Supplementary Series*, 235(1), March 2021. doi: 10.3847/1538-4365/abd805.

[90] Malz A.I. Soo J.Y.H. Almosallam I.A. Brescia M. Cavuoti S. Cohen-Tanugi J. Connolly A.J. DeRose J. Freeman P.E. Graham M.L. Iyer K.G. Jarvis M.J. Kalmbach J.B. Kovacs E. Lee A.B. Longo G. Morrison C.B. Newman J.A. Schmidt, S.J. and E. Nourbakhsh. Evaluation of probabilistic photometric redshift estimation approaches for the rubin observatory legacy survey of space and time (lsst). *Monthly Notices of the Royal Astronomical Society*, 499(2):1587–1606, 2020. doi: 10.1093/mnras/staa2799.

[91] Breiman L. Random forests. *Machine Learning*, (45(1)):5–32, 2001. doi: 10.1023/A:1010933404324.

[92] Liaw A. Documentation for r package randomforest. (PDF), October 2012.

[93] S Mucesh. A machine learning approach to galaxy properties: joint redshift-stellar mass probability distributions with random forest. *Monthly Notices of the Royal Astronomical Society*, 502(2):2770–2786, April 2021. doi: 10.48550/arXiv.2012.05928.

[94] T Dahlen et al. A critical assessment of photometric redshift methods: A candels investigation. *American Astrophys. J.*, (775(2)), 2013. doi: 10.1088/0004-637X/775/2/93.

[95] Bom C.R. Mucesh S. Palmese, A. and W.G. Hartley. A standard siren measurement of the hubble constant using gravitational wave events from the first three ligo/virgo observing runs and the desi legacy survey. 2021. doi: arXiv:2111.06445[astro-ph]. URL https://arxiv.org/abs/2111.06445.

[96] T.M. Hamill. Evaluation of probabilistic photometric redshift estimation approaches for the rubin observatory legacy survey of space and time (lsst). *Monthly Weather Review*, 129(3):550–560, 2001. doi: 10.1175/1520-0493(2001)129⟨0550:IORHFV.

[97] M.A. Stephens. Introduction to kolmogorov (1933) on the empirical determination of a distribution. *Springer Series in Statistics*, pages 93–105, 1992. doi: 10.1007/978-1-4612-4380-9_9.

[98] R.A. Kullback, S. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, (22(1)):79–86, 1951. doi: 10.1214/aoms/1177729694.

[99] H. Cramér. On the composition of elementary errors. *Scandinavian Actuarial Journal*, pages 13–74, 1928. doi: 10.1080/03461238.1928.10416862.

[100] Dr9 data release description. URL https://www.legacysurvey.org/dr9/description/.

[101] Balabdaoui F. Gneiting, T. and A.E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society*, (69(2)):243–268, 2007. doi: 10.1111/j.1467-9868.2007.00587.x.

[102] A. Dey et al. Overview of the desi legacy imaging surveys. *American Astron. Soc.*, (157(5)), April 2019. doi: 10.3847/1538-3881/ab089d.

[103] Wise data processing. URL https://wise2.ipac.caltech.edu/docs/release/allsky/expsup/sec4_4h.html#conv2ab.

[104] S Rauzy. A simple tool for assessing the completeness in apparent magnitude of magnitude-redshift samples. *Monthly Notices of the Royal Astronomical Society*, 324(1):51–56, June 2001. doi: 10.1046/j.1365-8711.2001.04078.x.