UNIVERSIDAD DE LAS AMÉRICAS PUEBLA

Escuela de Ciencias

Departamento de Actuaría, Física y Matemáticas

Redshift Estimation via Artificial Intelligence Methods

Trabajo de investigación que presenta el estudiante

Pablo Huerta Ocaña

167495

Director

Milagros Zeballos Rebaza

San Andrés Cholula, Puebla. noveno, 2023

TRABAJO DE INVESTIGACIÓN QUE PRESENTA EL ESTUDIANTE

PABLO HUERTA OCAÑA, 167495

ASESOR DEL PROYECTO

Milagros Zeballos Rebaza

PRESIDENTE DE TESIS

Miguel Angel Reyes Cortes

SECRETARIO DE TESIS

Ruben Blancas Rivera

Contents

1 Introduction **3**

2 Theoretical Framework **6**

2.1 Cosmological redshift 6

2.1.1 Spectral Redshifts 8

2.2 Galaxies 9

2.3 Photometry 11

2.3.1 Flux density and magnitude 11

2.3.2 Photometric redshifts 12

2.4 Artificial neural networks 16

3 Metodology **20**

3.1 ANNz2 20

3.2 SDSS method 23

3.3 Keras Neural network 24

3.4 Metrics 26

4 Input data **28**

4.1 Sloan Digital Sky Survey data 28

4.2 Wide-field Infrared Survey Explorer data 30

5 Results **37**

5.1 SDSS experiments 37

5.2 WISE experiments 41

6 Conclusions **51**

Abstract

This research focuses on the phenomenon of cosmological redshift, a crucial aspect in contemporary cosmology and our understanding of the universe. The study uses photometric redshift determination techniques based on artificial intelligence, utilizing the $ugriz$ photometric bands extracted from the Sloan Digital Sky Survey (SDSS) and the infrared bands from Wide-field Infrared Survey Explorer (WISE) as input. The primary objective is to demonstrate the effectiveness of neural networks in solving regression problems, as compared to empirical techniques.

The work introduces ANNz2, a code developed by Sadeh et al. (2016), and compares it with a neural network implemented in Keras, as well as the empirical method for determining redshifts from the SDSS. The results exhibit superior performance of neural networks in the range of $0.0 < z < 0.8$ when compared to the empirical method of the SDSS. Additionally, it is identified that the $ugriz$+WISE bands are not sufficient for predicting redshifts in the interval of $0.8 < z < 1.5$, although they still outperform other empirical techniques.

Key Words:Galaxies: distances and redshifts – Catalogs: SDSS – Large-scale structure of Universe – Methods: data analysis – Methods: numerical

Chapter 1

Introduction

This research work will focus on explaining the techniques for determining photometric redshifts and applying them in the inference of a vast number of galaxies. Models for determining redshifts using photometric information have been the subject of study for many years. The original idea, proposed by Baum in 1962, is centered on estimating the redshifts of galaxies and other celestial objects solely based on the relationships between the available photometric information and the redshifts, contrasting with the more conventional method that relies on utilizing spectroscopic information of the object in question.

These types of models have been extensively tested, successfully demonstrating the estimation of redshift values quite close to reality and showing that Photometric redshifts will be one of the key components for enhancing our understanding of the Universe in the next decade (Abdalla et al., 2011). These photometric determination techniques range from the most conventional physical methods, motivated by seeking theoretical relationships that explain redshifts based on physical and mathematical foundations, to photometric models driven by empirical experimentation, aiming to derive redshifts through statistical models that best fit the observed data.

Thanks to computational advances that have facilitated the handling and processing time of large databases, the interest of the scientific community has focused on exploring this virtues. This is also the case in cosmology, considered a theoretical science twenty or thirty years ago when there was not enough information to determine the validity of a model. Nowadays, it is impossible to think about cosmology without dealing with large accumulated observational data on the terabyte scale (Csabai et al., 2007).

In this context of technological innovation, artificial intelligence models begin to emerge, with the advantage of handling large databases and discovering complex relationships often surpassing physically motivated methods. Nowadays, artificial intelligence is a thriving field with many practical applications and active research topics (Goodfellow et al., 2016). Among these, one of the most crucial areas for astronomy is the study of redshifts.

The study of redshifts has played a fundamental role in shaping our understanding of the universe. Thanks to this, we have different perspectives on the evolution and expansion of the cosmos, as well as cosmological probes of dark matter and dark energy. All these measurements rely on accurately and precisely measuring galaxy redshifts of hundreds of millions of galaxies (Jones et al., 2023). However, the precision in measuring these redshifts, especially in the context of large astronomical datasets, has been a challenge. Conventional techniques are often constrained by the costs of spectroscopic information or the complexity in photometric data. It is in this context that the application of artificial intelligence emerges as a promising tool capable of overcoming these barriers. This evolution in methodology has the potential to revolutionize our understanding of the universe.

The objective of this research is to compare the performance of photometric redshift determination methods using artificial intelligence. This will be achieved by employing a neural network code named ANNz2 (Sadeh et al., 2016) and training it on a photometric database. The model's outcomes will be contrasted with another neural network implemented in Keras, as well as a photometric method independent of artificial intelligence techniques.

This research work will be divided as follows: In the theoretical framework (2), key concepts necessary to understand this investigation will be explored. The cosmological redshift will be addressed, along with its interpretation and the various methods to determine its value, including both spectroscopic and photometric techniques. Additionally, a detailed examination of galaxies, their properties, and the common methods for determining their redshifts will be conducted. Furthermore, pivotal concepts in photometry will be analyzed, with a focus on the measurement of fluxes and magnitudes. Finally, artificial neural networks will be explored as fundamental tools for the precise prediction and estimation of redshifts from photometric data.

In the methodology section (3), the procedures and techniques underlying the three photometric redshift determination methods used in this thesis will be detailed. ANNz2, based on various machine learning techniques, the photometric redshift determination method used by the Sloan Digital Sky Survey database, and a neural network developed specifically for this research. The implementation of each of these methods will be explained. Additionally, metrics and notation will be introduced to measure and compare the results obtained from each of the methods.

In the input data section (4), the methodology used to obtain the databases, which served for both training our models and testing them, will be detailed. In the results section (5), the outcomes of evaluating the models in different redshift intervals will be presented, followed by a brief discussion. Finally, in the conclusions section (6) the results will be summarized, and suggestions for future research in this field will be provided.

Chapter 2

Theoretical Framework

2.1 Cosmological redshift

In the field of astrophysics, there are various challenges that the scientific community must face, and among them, one of the most important and recurring problems is the determination and characterization of astronomical objects. Properties such as luminosity, temperature, mass, metallicity, and position are necessary for the proper study of the structure and evolution of the universe. Among all these, knowing the position of galaxies is one of the most important quantities that we are capable of measuring, allowing the study of several phenomena such as the identification of galaxy clusters, understanding the nature of dark energy, e.g. through surveys like the Dark Energy Survey (DES) and telescopes like the Large Synoptic Survey Telescope (Sadeh et al., 2016). The position of an astronomical object is determined via the right ascension (R.A.), declination (Dec.) and distance to us. Although obtaining the values of right ascension and declination can be achieved relatively easily, estimating the distance between the observer and the galaxy constitutes a more complicated problem that has required the use of various techniques throughout the history of physics.

Thanks to the current cosmological model, we are aware that the universe is expanding. This phenomenon causes a redshift (z) in the spectral lines of any astronomical object observed from Earth, i.e. we detect spectral lines at longer wavelengths compared to their original values. This change in wavelength, caused by the expansion of the universe can be compared to wavelength shift due to the Doppler effect:

$$z = \frac{\lambda - \lambda_0}{\lambda_0}, \tag{2.1.1}$$

where λ_0 refers to the original wavelength, and λ represents the wavelength that reaches our measuring instruments. Because this effect is caused by the expansion of the universe, we refere to it as cosmological redshift.

This redshift was first observed by Hubble in the late 1920s. He noticed that more distant galaxies exhibited more pronounced redshifts, suggesting that they were receding at higher speeds. This discovery not only confirmed the expansion of the universe but also established a relationship between cosmological redshift (z) and distance (r), known as the Hubble's Law.

$$z = \frac{H}{c} r \tag{2.1.2}$$

In this equation, c represents the speed of light, and H corresponds to the Hubble constant. It is important to note that this equation becomes more complex in current cosmological models as the rate of cosmic expansion changes over time, and other factors such as the density of matter and dark energy are taken into account. Nevertheless, the correlation between distance and recession velocity remains. Therefore, with a robust cosmological model, it is possible to determine distances using redshifts. In general, the most reliable method for estimating the distance to a source is through its redshift, except for limited and mostly local samples of redshift-independent ”distance indicators” (Bilicki et al., 2018). It is for all these reasons that the concept of redshift is so important in the field of astrophysics.

2.1.1 Spectral Redshifts

Like almost all physical properties of an astronomical object, redshifts are determined through the study of their spectra. Spectroscopy, which is the study of the interaction between matter and electromagnetic radiation, provides a way to obtain information by using instruments that separate light into its different wavelengths, such as dispersion spectrographs, slit spectrographs, integral field spectrographs, etc.

The basic equation of spectroscopy establishes a relationship between the energy difference of two quantum states (ΔE) of an atom and the wavelength (λ) of the light emitted or absorbed when the atom moves between those quantum states. It involves Planck's constant (h) and the speed of light in a vacuum (c).

$$\triangle E = h\frac{c}{\lambda} \tag{2.1.3}$$

Because energy differences are characteristic of each atomic element, this equation reveals that each atom emits or absorbs radiation only at specific wavelengths. Thanks to laboratory experiments, the wavelengths of some of the most important absorption or emission lines are known, such as the hydrogen Balmer lines, the lines of neutral helium, the iron lines, and the H and K doublet of ionized calcium (Karttunen et al., 2017). In this way, if we can accurately identify the elements interacting with the light emitted by an astronomical object, such as the elements in the atmosphere of a star or the elements in the intergalactic medium (IGM) for a galaxy, we can then determine the actual value of λ_0.

This is achieved by comparing at least two absorption or emission lines from the source with reference lines measured in a laboratory, which would be observed if there were no redshift. Once these values are known, the spectroscopic redshift (spec_z) can be calculated. (see equation 2.1.1).

2.2 Galaxies

In the night sky, there are many celestial bodies that can be observed thanks to various astronomical instruments. Throughout the history of humanity, different objects have been observed, from stars of various sizes and colors, to nebulae, asteroids, and more recently, objects such as black holes or neutron stars. Today, it is known that all these objects do not exist in isolation, but are components of much larger systems called galaxies.

These systems form when a large number of stars, gas, and dust interact gravitationally. More recently, this concept has been redefined to include the presence of dark matter; therefore, a star cluster would be classified as a galaxy only if it contains dark matter in addition to stars (Karttunen et al., 2017). Given the immense variety of shapes that galaxies can exhibit, the most common way to classify them is through the system proposed by Edwin Hubble, which categorizes galaxies into four primary groups based on their shapes: A) The elliptical galaxies, which have no distinctive features apart from their smooth, elliptical shape, can vary greatly in size and are typically found in dense galaxy clusters. B) The lenticular galaxies, which are a hybrid between ellipticals and spirals, characterized by a flat, disk-like shape with a bright central bulge. C) The spiral galaxies, which are characterized by large arms around a luminous core, and whose most interesting member is our own Milky Way galaxy. Lastly, D) the irregular galaxies, wich have no distinct shape such as the Milky Way´s satellite galaxies, the large and small Magellanic clouds. We can think of galaxies as the fundamental blocks of the universe, even though they are not uniformly distributed, since they form clusters of various sizes. Their study as individual pieces has been of great importance in the history of physics.

As the information we obtain from galaxies takes time to reach us while light travels across the cosmos, the farther a galaxy is from Earth, the further back in time we observe it. Since this relationship depends on the distance of the observer from the source, we can immediately relate cosmological redshift to the study of galaxies over time. Time is commonly discussed using z as a unit of measurement; for example, it is expected for the first stars to have been born at redshifts close to 20. When z=6, almost all the hydrogen in the universe was ionized, and the most massive stars were already born. Finally, at z=2, we observe the abundance peak of galaxies whit Active Galactic Nuclei galaxies (AGN) (Karttunen et al., 2017).

In order to study galaxy properties, the spectral energy distributions (SEDs) are our primary source of information (Walcher et al., 2011). We can think of the SED of a galaxy as a graphical representation of the energy emitted as a function of wavelength, ranging from ultraviolet (UV) wavelengths, $\lambda < 3500\,\text{Å}$ ($10\,\text{Å} = 1$nm), through the visible region, and extending into the far-infrared (FIR), $25 < \lambda < 250\ \mu$m.

Excluding AGN-type galaxies, the electromagnetic radiation from all galaxies comes mainly from the light emitted by the different stars that compose them, each one distinct in luminosity, mass, size, color, metallicity, etc. The method for creating the SED of a galaxy, called stellar population synthesis (Tinsley, 1972; Searle et al., 1973), is based on assuming that the radiation from a galaxy can be represented as the sum of the spectra of simple stellar populations (SSP). Each of these SSPs is an idealized set of stars that share one or more attributes, depending on the employed model. This way, the flux for each frequency ν of an SSP with mass M, age t and metallicity z is given by the contribution of its individual stars

$$L_v(t, Z) = \int_M \phi(M)_{t,Z} L_v(M, t, Z). \tag{2.2.1}$$

In this equation $\phi(M)_{t,Z}$ represents the stellar mass function and describes when a star will end its life cycle. Constructing these types of models is a challenge itself, and there is a significant amount of research in this area (Pietrinferni et al., 2009; Weiss and Schlattl, 2008). Among many other properties, spectroscopic redshifts can be derived from the SED of a galaxy using spectroscopic techniques as the analysis of absorption patterns of light at different wavelengths.

2.3 Photometry

Throughout history, humanity has endeavored to measure the brightness of celestial objects. From this intrigue and curiosity arises photometry, which is the branch of astronomy that measures the brightness of radiating objects in the sky (Budding and Demircan, 2007). This discipline plays a vital role in the field of astrophysics as it strives to quantify the light emitted by galaxies and stars. In the following discussion, we will explore the key concepts within this discipline and the reasons why its study can help us with the determination of redshifts.

2.3.1 Flux density and magnitude

One of the most important concepts when discussing photometry is the flux density, which represents the average power per unit area and per unit frecuency (W/m²/Hz). Flux density is a useful measure for determining the total amount of radiant energy reaching a measurement instrument, such as a radiometer. In radio astronomy, janskys are commonly used as the flux density unit, where one jansky (Jy) is equal to 10^{-26} W/m²/Hz.

It is important to emphasize that the flux density is a measurement that varies depending on the distance to the radiating object, whether we are talking about flux density at a specific frequency (F_v) or referring to the total flux density (F). In astronomy, the total power emitted by a star is called luminosity (L), and it is also defined in relation to the frequency $L_v = (WHz^{-1})$. Unlike flux,density luminosity does not depend on the distance (r), and a relationship can be established between them as

$$F_v = \frac{L_v}{4\pi r^2}. \tag{2.3.1}$$

Despite having techniques nowadays to measure the flux density and therefore the apparent brightness of astronomical objects, the study and characterization of celestial brightness has been doing since around 200 B.C. During that time, magnitude (m) was the value used to determine brightness and was assigned only through visual observation. In 1856, Norman R. Pogson realized that this scale followed a logarithmic increment derived from the nature of the human eye, and established a relationship between these magnitudes (m) and the flux density (F) (Budding and Demircan, 2007)

$$m = -2.5\log_{10}(\frac{F}{F_0}). \tag{2.3.2}$$

Where F_0 is the total flux density of an object such that its magnitude is zero. It is important to emphasize that in this scale, brighter objects have smaller magnitudes, while the faintest objects we have been able to observe exhibit magnitudes of 30. The color is another very useful concept in photometry and it is defined as the difference in magnitude of any object observed at two specified wavelength regions of observation (Budding and Demircan, 2007), It is expressed for example as $B - V$ where B and V are magnitudes in the blue and green regions of the electromagnetic spectrum.

Depending on the observation method and the value of F_0 , many systems of magnitudes can be defined, such as the multicolor or UBVRI system. This work will focus deeply on the use of the AB magnitude system. which is characterized by a value of $F_0 = 3631$ Jy, and is the same system used by the Sloan Digital Sky Survey (see section 4.1).

2.3.2 Photometric redshifts

As previously mentioned, the most efficient way to determine important properties of a galaxy, is through its spectral information. However, due to the nature of the method, it is time-consuming and also limited by the spectral range of multi-object spectrographs (de Diego et al., 2021), making it impractical for the dimmest galaxies. On the other hand, obtaining photometric values is an easier task to perform and due to limited telescope time, photometry becomes a better option for studying faint galaxies that may not be accessible for spectroscopic analysis (Li et al., 2007). It is in this context that the idea of determining redshifts using photometric data arises.

Photometric redshifts (photo$_z$) are a technique originally proposed by Baum (1962) and were used in the 1980s on low-redshift samples. Their interest has recently increased with the development of large field and deep field surveys (Bolzonella et al., 2000). This method relies on collected observations in different and specific wavelength bands, contrasting with the detailed spectral analysis provided by spectroscopy. A more practical definition can be given by Koo (1999), who mentions that photometric redshifts are those derived from images or photometry with spectral resolution $\lambda/\triangle\lambda \leq 20$, In this way, they are not taking into account images derived from slit and slitless spectra, narrow band images, ramped-filter imager, Fabry-Perot images and Fourier transform spectrometers.

Due to the limited information available, photometric redshifts have lower precision compared to spectroscopic values, which are generally assumed to be the true redshift. However, they remain highly useful for determining the properties of numerous galaxies. Consequently, determining the photo$_z$ of a galaxy is quite tolerable and sometimes even more effective than spec$_z$ (Li et al., 2007).

The estimation of photometric redshifts has been a broad field of research for the scientific community, as it enables measuring the distance of a much larger number of galaxies whose spectral information is not known (Beck et al., 2016), The different approaches to obtain photometric redshifts rely on having some kind of information about the galaxy in question, whether it is its magnitudes, one, two or more colors, its surface brightness, light profile, etc. Due to the multitude of parameters and the wide range of ways in which they can be related to estimate photometric redshifts, different types of techniques have been developed over the years. These various methods, whose compilation can be found in Hildebrandt et al. (2010) or Abdalla et al. (2011), can be easily divided into two categories: empirical and template fitting methods.

Template fitting methods

The template fitting methods were the first technique developed to obtain photometric redshifts and became the most popular tool among the scientific community in the past decade. This method was adopted by Puschell et al. (1982), who introduced the maximum likelihood approach (χ^2) fitting by modeling SEDs to broad-band (RIJHK) photometry. This method has the advantage of predicting redshifts using SED templates without having any spectroscopical redshifts. A more precise definition can be provided by Benítez (2000), who mentions that any template fitting method is based on storing a library of template spectra, either generated empirically or through population synthesis techniques. These templates, after being redshifted and corrected for intergalactic extinction, are then compared with the galaxy's colors to determine the photo$_z$ that best fits the observations. In general the χ^2 method it is based on minimizing the following function:

$$\chi^2(z) = \sum_{i=1}^{Nfilters} (\frac{F_{obs,i} - b * F_{temp,i}(z)}{\sigma^2})^2 \tag{2.3.3}$$

where F_{obs} is the observed flux density, F_{temp} are the template flux densities and σ the uncertanty for the filter i, In this equation b is a normalization constant.

Templates spectra are intended to represent a great variety of the galaxies, e.g. diferent types of galaxy morphology and luminosities. The most popular template libraries can be found in Coleman et al. (1980) or Bruzual A. and Charlot (1993). these templates also incorporate astrophysical effects, such as dust extinction in the Milky Way or in the observed galaxy (Sadeh et al., 2016).

One of the main problems we may encounter when using a template fitting method is mismatches between the templates and the galaxies in the sample. As we can see, the entire method relies on assuming that the SED templates provide a good representation of the actual SEDs. Therefore, they depend on proper calibration, which is commonly performed using spectroscopic data (Sadeh et al., 2016). This method can encounter challenges due to factors such as the influence of emission lines, templates that may not adequately represent the majority of galaxies in the sample, sensitivity to various other measurements (e.g., bandpass profiles and photometric calibrations), the impact of dust-induced reddening, and the presence of AGN, which necessitate the use of significantly different templates (Walcher et al., 2011).

In summary, template fitting methods offer a way to estimate the photometric redshift for a galaxy primarily using SED fitting techniques. Despite the potential inherent errors in the method, they are a good choice when exploring new regimes in a survey or when investigating a region of space for which we lack spectroscopic information. This method has been widely used since 1982, as it allows for the determination of other physical properties of galaxies in addition to their redshifts.

Empirical methods

Unlike template fitting methods, which rely on physically motivated models, the concept behind empirical methods is based on discovering relationships between photometric variables and spectroscopic redshifts. Once these relationships have been derived in a training dataset, the method can be applied to datasets for which spectroscopic information is not available.

To understand the bases of this method, we can think of $C = c_1, c_2, ...c_n$ as the set of colors (or magnitudes) that are sufficient for estimating photometric redshifts. The method aims to fit a surface $z = z(C)$ by training the data with its corresponding set of spectroscopic data. For this to make sense, the surface $z = z(C)$ must be a well-defined function in the color space, where for each value of C there is only one vaule of z, In reality, the relationship between redshifts and colors does not behave in this way, as a galaxy with photometric values (C) may have slightly diferent redshifts (Benítez, 2000), However, it has been observed that for $z < 1$ it is possible to assume that the surface $z = z(C)$ does not curve back on itself, making it a good approximation to the real picture (Brunner et al., 1997).

There are many ways to derive a relationship $z = z(C)$ over the training data, ranging from a simple polynomial fit (Connolly et al., 1995) to the implementation of advanced machine learning techniques like the use of neural networks (see Section 2.4). Since its inception, this method has proven to be superior to template fitting techniques. Its success stems from the fact that its training set consists of real galaxies, which helps avoid the problem of inaccurate templates. Additionally, since the training set is a subsample of the survey, it also encompasses the effects of filter bandpasses and flux calibrations (Walcher et al., 2011).

While it is true that empirical methods tend to outperform template fitting (Bilicki et al., 2018; Hildebrandt et al., 2010), it is important to emphasize that they also come with considerations. This method becomes more uncertain when applied to objects further from the training data. By establishing relationships solely with this data, an empirical method may struggle to associate a redshift with a distance greater than what was covered in the training set. Additionally, the training dataset must be large enough to span the entire color space. This is why the selection of training sets is crucial when calibrating a model of this nature.

One of the most significant empirical methods, which will be discussed in this research work, involves the use of neural networks. Therefore, a brief explanation about the subject and its applications in photometric redshifts will be provided below.

2.4 Artificial neural networks

Neural Networks (NN) are an information processing architecture inspired by the functioning of the human brain. This concept was first introduced in 1943 by the neurologist Warren McCulloch and the mathematician Walter Pitts. Since then, various neural networks architectures have been proposed and interest in NN has evolved over time. In the 1960s, psychologist Frank Rosenblatt built the first computer capable of learning through trial and error using a neural network. However, it was in the 1980s when architectures with more robust training methods were proposed, proving to be a useful tool for tasks involving a large amount of data.

The interest in neural networks waned in the 1990s, as other machine learning techniques showed better results for classification and regression problems. However, recently, the research and use of neural networks have experienced a resurgence in popularity. This resurgence is mainly due to the availability of a huge quantity of data for training, a significant increase in computing power since the 1990s, improvements in training algorithms, and the demonstration that theoretical issues, such as algorithms getting stuck in local optima, do not pose a real problem (Géron, 2017).

Neural networks are part of machine learning techniques, which, as the name suggests, are based on algorithms that allow the computer to learn from a set of data. Tom Mitchell (1997) give us a more formal definition : 'a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.'

It is expected that a neural network can improve its performance on a task with more experience, similar to how neurons in a brain function. While a single neuron does not exhibit any form of intelligence, a large ensemble of interconnected neurons can exhibit intelligent behavior (Géron, 2017).

The basic unit of information processing in an NN is the neuron. Each neuron receives a set of numerical inputs, either from other neurons or from the input data, and produces a single output. These neurons are grouped together to form layers. The first layer (i=0) is known as the input layer, where the number of neurons is equal to the number of input variables. The last layer (i=N) is called the output layer, and in a regression problem, this layer would have only one neuron representing the predicted value. The intermediate layers are known as hidden layers, and the number of layers and neurons per layer is arbitrary depending on the chosen architecture. The value of each neuron j in layer n is calculated using the following formula:

$$a_j^{(n)} = f(\sum_{i=0}^{M}(w_{ji}a_i^{(n-1)}) + b_j), \tag{2.4.1}$$

where M is the set indexing the number of neurons in the previous layer (n-1), w_{ji} represents the weights connecting neuron $a_i^{(n-1)}$ in the previous layer to neuron $a_j^{(n)}$ in the current layer. b_j is known as the bias of the neuron, and it is a value that allows adjusting the starting point of the activation function f.

The activation function introduces nonlinearities in the neural network, enabling it to learn and represent more complex relationships in the data. The choice of the number of layers in the network and the number of neurons is arbitrary, meaning that there are no fixed rules to determine how many layers and neurons should be used. On the other hand, the choice of the activation function depends on the nature of the problem, although the most common ones in the neural network literature are:

Function Sigmoid: $f(x) = \frac{1}{1+e^{-x}}$

Function ReLU (Rectified Linear Unit): $f(x)$ =MAX$(0, x)$

Function Tanh: $f(x) = \tanh(x)$

The objective of an NN is to find the values for w_{ij} that minimize the cost function C during the training process:

$$C(w, b) = \frac{1}{n}\sum_{i=1}^{n}(o_i - t_i)^2. \tag{2.4.2}$$

Here, w and b are the values of all the weights and biases in the network, n is the number of training inputs, o_i refers to the values predicted by the NN for the ith sample and t_i is the actual output value that these samples have. In this way, when $C(w, b) \approx 0$ we can see that o_i approximates t_i $\forall i \in n$. Therefore, we can say that the algorithm is doing a good job finding the weights and biases.

The way an NN minimizes the cost function is through a process called backpropagation. This process initializes random values for each weight (w) and through minimization algorithms such as gradient descent, the value of the function is iteratively reduced until reaching a local minimum. Once this training process is completed, the neural network is ready to make predictions on new data. By adjusting the weights and biases, the network has learned to detect patterns in the training data, enabling it to generalize these patterns and make inferences on unknown data.

In the field of astronomy, neural networks are widely employed instruments. Their capability to establish relationships between input and output variables allows for versatile applications across various problem types. They have found extensive use in classification tasks, such as galaxy classification (Ball et al., 2004), spectral classification (Weaver, 2000), and the classification of astronomical objects (Zhang and Zhao, 2007), among others.

One of their most popular applications in regression problems lies in the determination of photometric redshift estimations. In this context, each value of C (colors or magnitudes) is inputted into the neural network. The predicted value by the network corresponds to the redshift (Firth et al., 2003; Collister and Lahav, 2004; Singal et al., 2011), or a probability distribution of the redshift (Pasquet et al., 2019).

Chapter 3

Metodology

This work will compare the performance of three methods for determining photometric redshifts: ANNz2, NNk, and the SDSS method. This section will encompass a comprehensive description of the methods, including how they derive redshifts and the required inputs for their operation. Additionally, the metrics used to evaluate the performance of each method will be defined in the results section (see Section 5).

3.1 ANNz2

The first method employed for calculating photometric redshifts relies on empirical approaches utilizing various machine learning techniques. This method is known as ANNz2.

ANNz2 is a code developed by Sadeh et al. (2016)[1]. It is a program that employs various machine learning techniques to address both classification and regression problems. ANNz2 is a new implementation of the public software for computing photometric redshifts, originally created by Collister and Lahav (2004). The innovation of this code lies in how uncertainties for the predicted redshift values are estimated. Unlike its predecessor, ANNz2 incorporates the generation of probability distribution

[1] https://github.com/IftachSadeh/ANNZ

functions.

ANNz2 is based on the Toolkit for Multivariate Data Analysis (TMVA) package (Hoecker et al., 2007). Many machine learning techniques are included in this package, such as cuts, likelihood, k-nearest neighbors (KNN), boosted decision and regression trees (BDTs), linear discriminant analysis, Fisher discriminants, support vector machine, among others. However, for this research only neural networks and BDTs will be used.

The available architecture in ANNz2 for neural networks consists of a multi-layer perceptron (MLP). The number of hidden layers as well as the number of neurons can be defined by the user and vary depending on the selected operating mode. The activation function for each neuron is `TANH`. Unlike a classic MLP, which employs the backpropagation algorithm (see Section 2.4), ANNz2 uses a minimization algorithm called Broyden-Fletcher-Goldfarb-Shannon (BFGS). The main difference between the two methods is that in backpropagation, minimization is done with gradient descent, which is a first-order approach, while the BFGS method minimizes by estimating the inverse of the Hessian matrix, which is a second-order approach. This difference can offer advantages in terms of convergence and overcoming local minima, but the algorithm becomes computationally more expensive.

Among the operating modes of ANNz2, the first one is the ”single regression” mode. In this mode, a simple regression is performed using a single neural network. The architecture of this network defaults to three hidden layers with $N + 1$, $N + 9$, $N + 4$ neurons, where N is the number of input variables. This is the most basic mode of ANNz2 and it also allows for running a classification algorithm with the same architecture.

The second available mode in ANNz2, and the one to be used for this research work, is the ”random regression” mode. Unlike single regression, this mode employs an ensemble of different machine learning techniques, which can be either BDTs or NN. These techniques differ from each other in terms of architecture, initial random seed, and the input parameters used for training.

The random regression mode in ANNz2 offers several advantages over a simple NN. Once a large number of NNs are trained—100 in this research—ANNz2 initiates an optimization process. This process derives a distribution of photo-z solutions for each galaxy and selects a subset of NNs that achieve optimal performance. In this stage, uncertainty estimates are also generated using a KNN method (Oyaizu et al., 2008).

Another significant advantage of this code is the high level of automation achieved through Python scripts. This means that the user does not need to individually define the properties of each NN or manually set the training criteria. This automation is particularly valuable when training a large number of neural networks, in such cases, the number of hidden layers and neurons are randomly generated for each NN.

The final output of random regression includes the best solution from all the NNs, referred to as ”best Z”. Additionally, the code provides the full binned probability distribution function. The number of bins is a parameter that can be defined by the user, and for this research work, it was set to 80. Finally, the code outputs the uncertainty value for each ”best Z”. This uncertainty should not be interpreted as an error of the method with respect to the original redshift (defined in Section 3.4), but rather as a quantification of the uncertainties inherent to the photo-z derivation method.

Sadeh et al. (2016) explains that the uncertainties of the method primarily arise from three distinct phenomena. First, there are uncertainties associated with the inputs used in training. This is because the magnitudes utilized may not always be sufficient to accurately determine the redshift. The second type of uncertainty stems from the inherent variability of a neural network, which can be influenced by factors such as different random seeds, parameters, or even the choice between using NNs and BDTs. The third type of error can result from the training database not being entirely representative, leading to an incomplete color space. This effect can be exacerbated if the evaluation data does not reside in the same color space defined by the training database.

3.2 SDSS method

The second methodology for obtaining photometric redshifts corresponds to the empirical method underlying the database of the Sloan Digital Sky Survey Data Release 16 (Ahumada et al., 2020). Unlike the other methods presented in this research, this one was not tested on a new database nor replicated using any code. Instead, the predicted photo$_z$ values by the SDSS method are available for download along with the magnitude values and spec$_z$ values that were necessary to train the other methods (see SDSS in Section 4.1).

Similar to earlier SDSS releases (Csabai et al., 2007), this method relies on adopting a local linear model to describe the dependence between broad-band colors/magnitudes and redshifts. A detailed explanation of the method can be found in Beck et al. (2016). In this section, only the relevant parts for this research work will be described.

If we index the galaxies in a set of objects(Q) for which we would like to estimate their redshifts, we can define z_i as the redshift and d_i as the magnitude vector for the i-th galaxy. Similarly, let j be the index of the galaxies in the training set (T), for which we know both the values of z_j and d_j. The local linear model is defined as follows.

$$z_i \approx c_i + a_i d_i = z_{phot,i}. \tag{3.2.1}$$

In this ecuation $z_{phot,i}$ refers to the redshift value estimated by the method, while c_i is a constant and a_i is a vector with linear coefficients. To determine these values, an empirical relationship must be found using the training set T. In this method, this relationship is found by identifying the k-nearest neighbors of the galaxy i within T. This means finding the k galaxies whose d_j magnitudes are close to d_i in Euclidean distance. Therefore, these parameters can be determined by minimizing the following expression.

$$\chi_i^2 = \sum_{j \in O} \frac{(z_j - c_i - a_i d_i)^2}{w_j} \tag{3.2.2}$$

In this ecuation the set O comprises the k nearest galaxies to i, w_j is a weight that serves to represent uncertainties in z_j and the minimization must be performed over every galaxy j in O.

For the determination of redshifts in the SDSS DR16, the components of vector d consist of the r-band magnitude and the $u-g, g-r, r-i, i-z$ colors (As defined in section 4.1).The nearest neighbours are weighted equally, $w_j = 1 \; \forall j \in T$. Additionally, it is assumed that the reported error of spec$_z$ in SDSS is zero, therefore $Z_{spec,j} = z_j$.

3.3 Keras Neural network

The final method for obtaining photometric redshifts was carried out under an empirical approach, where the relationships between magnitudes and redshifts were derived using a neural network implemented in Keras and executed in TensorFlow (Abadi et al., 2016). This method will be referred to as Keras Neural Network (NNk).

The architecture of the neural network consists of an input layer with N neurons, one for each magnitude value. It includes nine hidden layers, each with two hundred neurons and a ReLU activation function. The output layer consists of a single neuron without an activation function.

The loss function of the neural network is defined as `mean_squared_error`. This means that the network will seek to minimize the squared difference between the network's predictions and the actual training value.

An important parameter when creating a neural network in Keras is the optimizer, which is responsible for adjusting the weights and biases of the neurons. The optimizer used was `Adam`, which combines the stochastic gradient descent method with learning rate adaptation techniques. The learning rate is a parameter that controls the size of the steps taken during optimization. A small learning rate allows for more precise adjustments at the cost of being computationally more expensive, while a large learning rate can make the process unstable but significantly reduce training time. The commonly used learning rate for deep learning is 0.001, so we use this value as the default parameter.

The training parameters were kept constant for all experiments reported in the results (see 5). These are `epochs=200`, `batch_size=30`, and `validation_split=2.5`.

The parameter `epochs` refers to the number of iterations through the entire training dataset. During an epoch, the neural network adjusts the weights using each example from the training database. This parameter needs to be tuned to strike a balance between precision with the trained data and generalization for making inferences on unknown data.

The `batch_size` is a parameter in Keras that enables the division of data during the training process. This way, the model's weights will be adjusted after processing each batch, meaning that 30 examples will be processed before recalculating the weights. It is recommended to keep this parameter small to prevent overfitting.

The `validation_split` parameter allows for regulating what percentage of the training data will be exclusively used as a validation set. The purpose of this division is to test the model's performance on a dataset that was not used for training. This allows for simulating the behavior of the network in terms of generalization and adjusting the weights to prevent overfitting.

While the first two parameters in Keras were chosen arbitrarily, attempting to strike a balance between the model's learning capacity and computational cost, the third parameter was selected specifically to replicate the default `validation_split` parameter of ANNz2. This ensures that both methods use 25% of the training data for validation.

Metric	Equation
Bias	$\delta_z = \text{photo}_z - \text{spec}_z$ (1)
Normalised bias	$n\delta_z = \frac{\text{photo}_z - \text{spec}_z}{1+\text{spec}_z}$ (2)
Outlier	$O : \frac{\lvert\text{photo}_z - \text{spec}_z\rvert}{1+\text{spec}_z} > 0.15$ (3)
Catastrophic Outlier	$O_c : \lvert\text{photo}_z - \text{spec}_z\rvert > 1.0$ (4)
Loss function	$L = 1 - \frac{1}{1+(\frac{n\delta}{\gamma})^2}$ (5)
Standard deviation of normalised error	$\sigma_{n\delta} = \sqrt{\frac{1}{N-1}\sum(n\delta - \langle n\delta\rangle)^2}$ (6)
Scaled Median Absolute Deviation (SMAD)	$\Omega \cdot \text{median}(\lvert n\delta - \text{median}(n\delta)\rvert)$ (7)
Scaled Root Mean Square (RMS)	$\text{RMS}_{n\delta} = \sqrt{\frac{1}{N}\sum(n\delta)^2}$ (8)

Table 3.1: Metrics employed for evaluating the photometric redshifts methods.

3.4 Metrics

In order to effectively quantify the performance of the various methods previously presented, it is essential to establish a set of specific metrics that will allow us to evaluate their performance. These metrics will provide us with a method to compare the performance of the three photometric redshift techniques presented in this research work.

These metrics will be meaningful after testing each of the methods with an evaluation dataset containing N galaxies, for which we have both the spectroscopic redshift (spec_z) and the photometric redshift derived by the methods (photo_z). For each galaxy, the bias (δ) defined in the Eq. 1 of the table 3.1 represents the difference between the predicted value and the true redshift value, which will always be considered as the spectroscopic redshift. This measure is highly useful as it enables us to observe whether the method is overestimating or underestimating the values. Additionally, it serves as a clear indicator of the method's performance, with smaller bias values indicating better performance.

While the bias is an absolute measure of error, the Normalized Bias ($n\delta$) in the Eq. 2 of the table 3.1 serves as a relative measure of error. It quantifies the error in relation to the actual value. This metric helps us understand how significant the error is relative to the magnitude of the redshift. It proves especially valuable when analyzing redshift bins with varying magnitudes.

For this work, it was decided to categorize two different types of outliers in the photometric redshifts.

In Eqs. 4 and 5, an outlier (O) is defined if $|n\delta_z| > 0.15$, and a catastrophic outlier (O_c) if $|\delta| > 1.0$. The choice of 0.15 to categorize an outlier was made to be consistent with other analyses of photometric methods (Hildebrandt et al., 2010; de Jong et al., 2017), while the catastrophic outlier is presented as provided by Jones et al. (2023). Similarly, the Loss function (L) in Eq. 5, which was evaluated using $\gamma = 0.15$, is useful in regression problems as the metric produces a normalized value between 0 and 1, Proximities to zero in values imply a high degree of proficiency concerning γ, whereas closeness to 1 indicates a deterioration in performance. The outliers will be presented as a percentage of the number of galaxies N, while the rest of the metrics as an average over the data, where the symbol $\langle A \rangle$ refers to the average of A.

The standard deviation of normalized error ($\sigma_{n\delta}$) in Eq. 6 is a metric that indicates how much dispersion exists in the data with respect to $\langle n\delta \rangle$. The scaled median absolute deviation (SMAD) in Eq. 7 is also a measure of variability, but unlike $\sigma_{n\delta}$, it is less sensitive to outlier values. This is why visualizing both can help understanding the complete nature of the training. We chose $\Omega = 1.4826$ as used in Bilicki et al. (2018). The last included metric is the Scaled Root Mean Square of normalized error (RMS) as shown in Eq. 8.

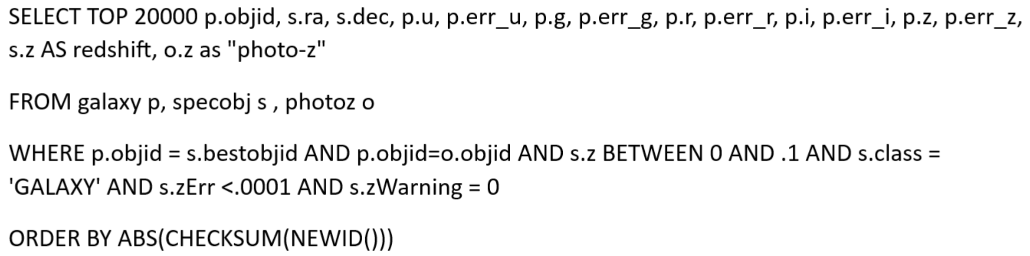
Chapter 4

Input data

In this section, we present the samples that will be used to test the three photometric redshift methods, along with a description of the surveys from which they were obtained. For this research work, data from SDSS DR16 were used to create two datasets. The first sample covers redshifts $0 < z < 0.8$. This interval was chosen in order to compare the methods in a space where they have been experimentally shown to yield good results (Sadeh et al., 2016). The second sample, with $0 < z < 1.5$ aims to include the WISE bands to test if the empirical methods can extend their accuracy to regions of higher redshift.

4.1 Sloan Digital Sky Survey data

The Sloan Digital Sky Survey (SDSS) is a space research project that provides both photometric and spectroscopic data over a large region of the sky, specifically covering approximately π steradians with a Galactic latitude of 30°, in five broad optical bands (York et al., 2000). Since 1998, the data published by the SDSS has been collected from the Apache Point Observatory using the 2.5 m Sloan Foundation Telescope (Gunn et al., 2006). Since 2017, observations have also been made using the 2.5 m du Pont telescope located at the Las Campanas Observatory (LCO). The SDSS observes in five filters, with



```
SELECT TOP 20000 p.objid, s.ra, s.dec, p.u, p.err_u, p.g, p.err_g, p.r, p.err_r, p.i, p.err_i, p.z, p.err_z,
s.z AS redshift, o.z as "photo-z"

FROM galaxy p, specobj s , photoz o

WHERE p.objid = s.bestobjid AND p.objid=o.objid AND s.z BETWEEN 0 AND .1 AND s.class =
'GALAXY' AND s.zErr <.0001 AND s.zWarning = 0

ORDER BY ABS(CHECKSUM(NEWID()))
```

Figure 4.1: Query used in the DR16 SDSS

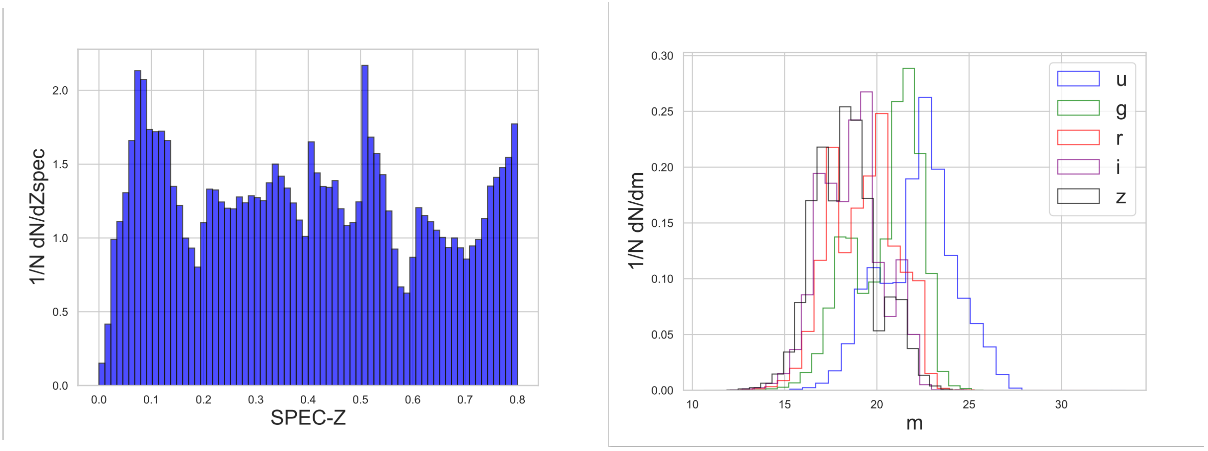
effective design wavelengths of 3550, 4770, 6230, 7620, and 9130 $\AA$ for the u, g, r, i, and z bands, respectively (Fukugita et al., 1996).

This research project retrieved data from the 16th data release (DR16) of the SDSS [1]. For each of the two databases, celestial coordinates (Right Ascension and Declination), Pogson galaxy magnitudes in five bands ($ugriz$), the value of the spectroscopic redshift, and the value of the photometric redshift derived using the SDSS method were obtained.

fig 4.1 is an example of the query used in the SDSS DR16 database for a redshift range from 0 to 0.1. It is requested that `s.class='GALAXY'`, as the color space for quasars is fundamentally different from other galaxies and they are avoided to achieve better performance in empirical methods. Since one of the most important premises in the methodology is that spectroscopic redshifts are interpreted as the actual redshift, the value of the spectroscopic error was limited to `s.zErr<0.0001`. If a magnitude is not associated with an error in the survey, a warning is assigned to it, as this is often an indication that the measurement is not correct, only values with warnings `a.zWarning=0` were admitted.

The sample 1 consists of 156,573 galaxies with an average redshift of $\langle z\rangle = .39$ distributed in the range of $0 < z < .8$. We aimed for the data to be uniform within the interval in order to achieve a complete color space. The general properties of the data set are shown in fig. 4.2 and fig. 4.3, figure 4.2 shows the distribution of spectroscopic redshifts in the dataset, as well as the distribution of the five magnitudes. On the other hand, Figure 4.3 displays the correlations between the colors of the magnitudes.

[1] https://skyserver.sdss.org/dr16/en/tools/toolshome.aspx



(a)

(b)

Figure 4.2: Properties of galaxies in the first sample: (a) differential distribution of the spectroscopic redshift. (b) differential distribution of the magnitudes in the five bands $ugriz$.

The sample 2 consists of the same galaxies as the first database plus 47,633 galaxies corresponding to all the galaxies available in the SDSS DR16 for a redshift interval of $0.8 < z < 1.5$. This leaves us with a database of 204,206 galaxies with $\langle z \rangle = .5$ in the interval of $0 < z < 1.5$. the distribution of spectroscopic redshifts are shown in fig. 4.4 while fig. 4.5 displays the correlations between the colors of the magnitudes. We can observe that there are few values for the farther redshifts.

4.2 Wide-field Infrared Survey Explorer data

The Wide-field Infrared Survey Explorer (WISE) is an astronomical space telescope operating in the infrared wavelength range, launched on December 14, 2009. It is a medium-class Explorer mission funded by NASA. WISE is mapping the sky using four infrared bands, W1, W2, W3, and W4, which are centered at 3.4, 4.6, 12, and 22 micrometers, respectively. This photometric information is obtained using a 40 cm telescope that feeds arrays with a total of 4 million pixels (Wright et al., 2010).

The data processing and analysis of the band fluxes are carried out by the Infrared Processing and Analysis Center (IPAC) [2]. For this research work, the magnitude values in the W1, W2, W3, and W4

[2] https://irsa.ipac.caltech.edu/cgi-bin/Gator/nph-scan?submit=Select&projshort=WISE

filters for the galaxies were obtained from the second sample acquired from the SDSS.

The AllWISE Source Catalog enables users to perform a multi-object search using spatial coordinates (Right Ascension and Declination) to retrieve information pertaining to specific objects. Since the values of Right Ascension and Declination for the same object may vary slightly from those reported in other surveys, an exact match is not the primary objective. Rather, WISE endeavors to identify all objects whose spatial coordinates fall within a defined search cone. The output of this query provides, for each source, all the objects found within the search radius, along with the value of the distance between the search coordinates and the coordinates of the output.

In order to align the WISE magnitudes with the galaxies obtained from the SDSS, objects were sought with spatial coordinates within a cone search radius of 10 arcseconds. The instrumental profile-fit photometry magnitude and photometry flux uncertainty values were requested for each of the four bands. For each point source, only the object with the smallest associated distance was selected.

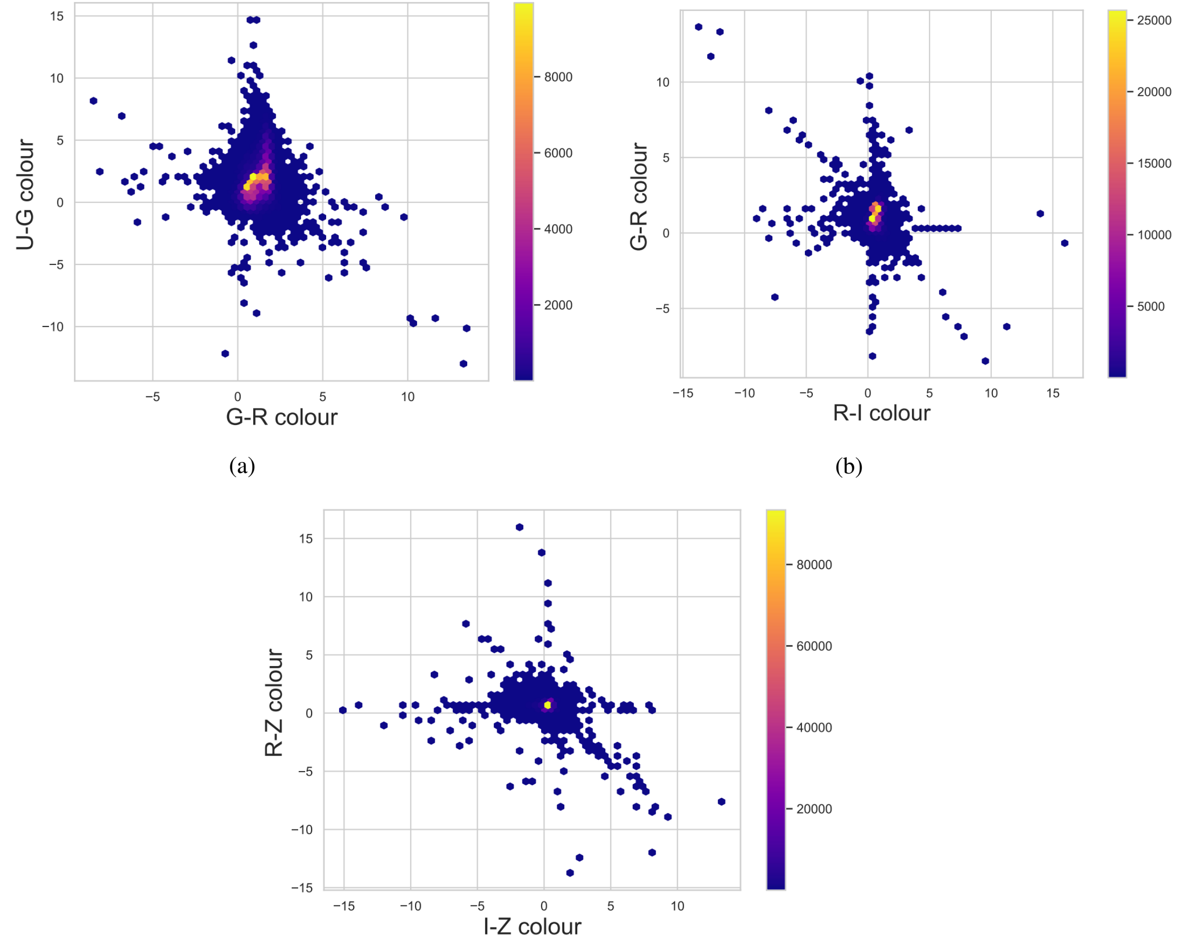
Out of the 204,206 galaxies subjected to the search, only 181,601 were found with corresponding magnitudes in the WISE database. One notable observation about these galaxies was that the majority of their magnitudes did not have numerical values in the band's errors. This indicated that the magnitude values corresponded to a 95% confidence upper limit, or the source was not measurable. As a result, all values lacking a numerical error were excluded.

After removing these values, Sample 2 was divided into two databases. The first one (fig 4.6) represents the redshift distribution of the galaxies after removing the values with uncertainties in all WISE bands, and the second one (fig 4.7) represents the redshift distribution of the galaxies after removing only values with uncertainties in the W1 and W2 bands.

This cleaning process not only resulted in having few values beyond $0.8z$, but also drastically reduced the number of galaxies. For the database where all four WISE filters were retained, the number of galaxies was reduced to 21,302, with an average redshift of $\langle z \rangle = 0.28$.

The aim of expanding the predictor variables to bands closer to the infrared spectrum is to enhance the model's characterization capability by incorporating emissions from this region of the electromagnetic spectrum. However, the quantity of data in the training set is also pivotal for the model's performance. This is why we retained the database where only the W1 and W2 bands were included, while excluding the W3 and W4 bands due to their higher measurement uncertainties.

The figure 4.7 displays the redshift distribution of galaxies for which information is available only in the W1 and W2 bands. This refinement yielded 166,999 galaxies with an average redshift of $\langle z \rangle = 0.44$. While this approach results in a loss of information about a portion of the electromagnetic spectrum, it achieves a greater uniformity in the color-magnitude space.

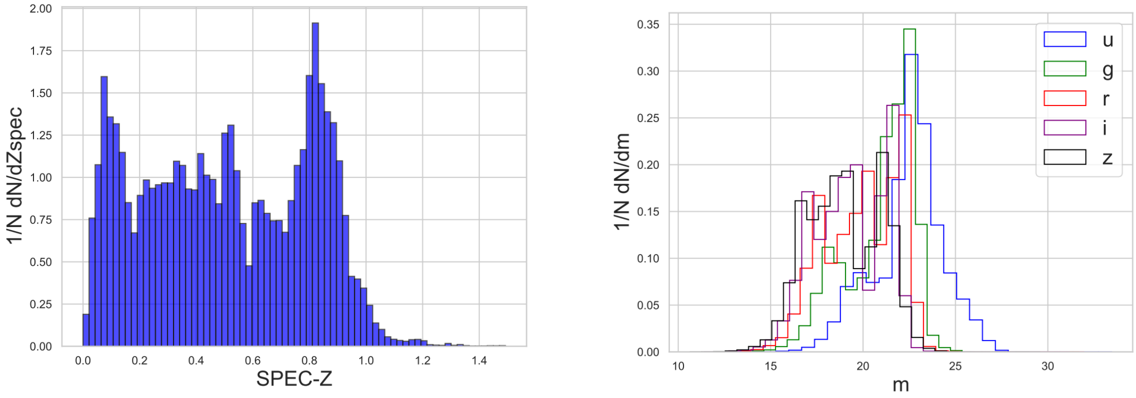


(a)

(b)

(c)

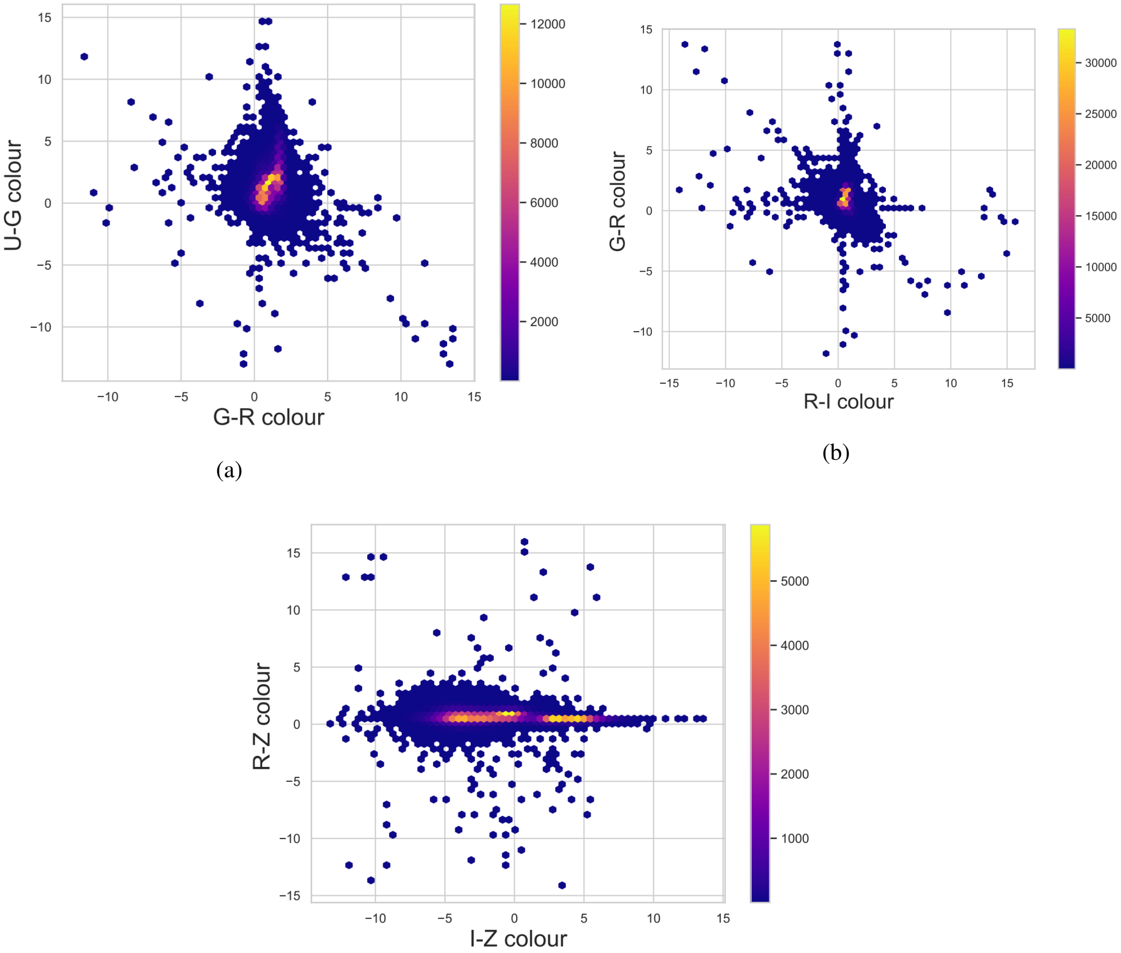
Figure 4.3: Correlation between different color combinations of the first sample, where the color of each hexagon indicates the point density in that area of the plot.



(a)

(b)

Figure 4.4: Properties of galaxies in the second sample: (a) differential distribution of the spectroscopic redshift. (b) differential distribution of the magnitudes in the five bands $ugriz$.

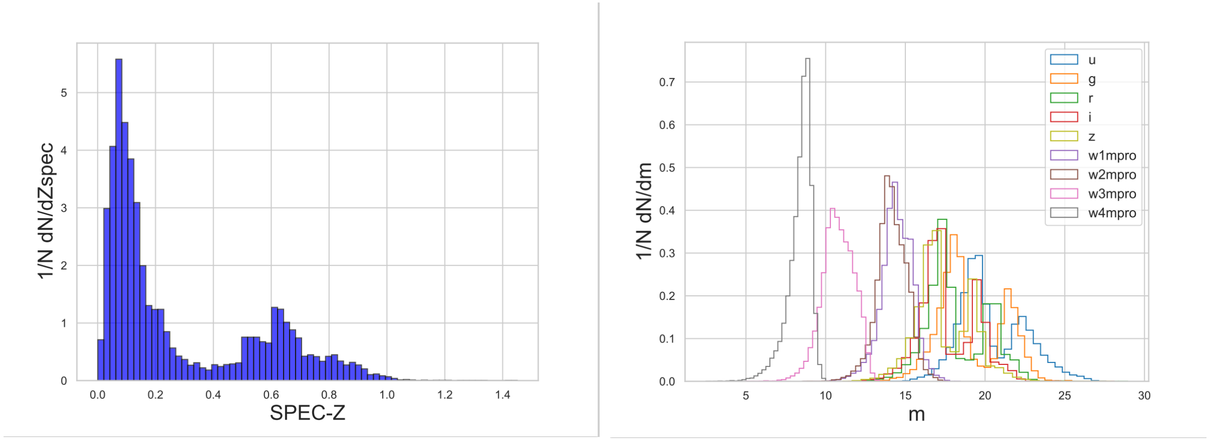


(a)

(b)

(c)

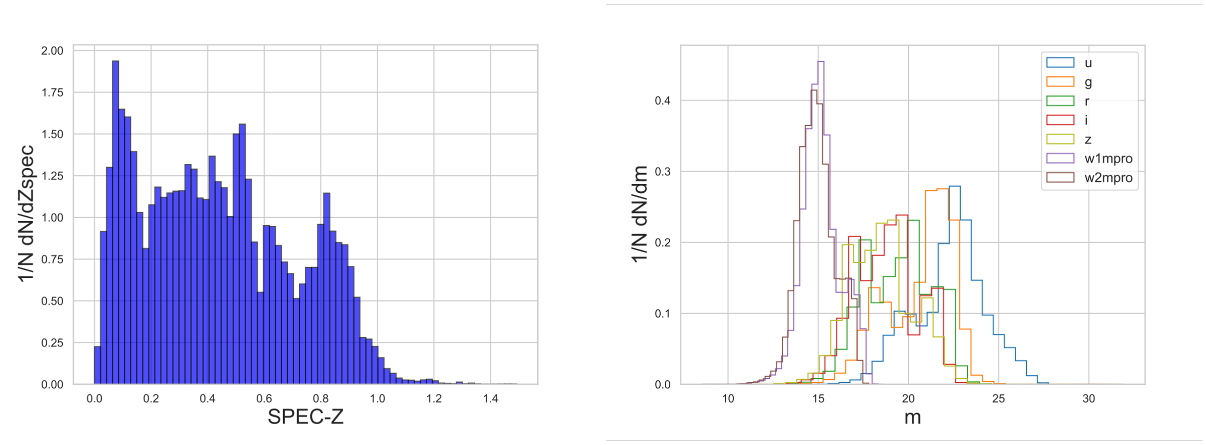
Figure 4.5: Correlation between different color combinations of the second sample, where the color of each hexagon indicates the point density in that area of the plot.



(a)

(b)

Figure 4.6: Properties of galaxies with $ugriz$ and full WISE bands: (a) differential distribution of the spectroscopic redshift. (b) differential distribution of the magnitudes



(a)

(b)

Figure 4.7: Properties of galaxies with $ugriz$, W1 and W2 bands: (a) differential distribution of the spectroscopic redshift. (b) differential distribution of the magnitudes

Chapter 5

Results

This section shows the performance of two photometric redshift determination techniques. The results displays the different metrics obtained from training the neural networks multiple times, always comparing these methods to the photometric predictions provided by the SDSS.

It is important to remember that these results should be understood as quantifiers of empirical methods on a subset of the actual photometric set. Their accuracy or impact on a complete survey of galaxies will be constrained by the spectroscopic information available in the survey. Extrapolations of these algorithms to other regions of the sky may yield entirely different results, especially if incomplete spec$_z$ samples are used as calibrators, as is the case for most modern photometric surveys (Hildebrandt et al., 2010).

5.1 SDSS experiments

The initial phase of this study utilizes data sourced from SDSS DR16 to compare various methods of determining redshifts within the range of $0 < z < 0.8$. To conduct these experiments, we divided the original database into two sets: a training set and a testing or evaluation set. We aimed for the training set to be as comprehensive as possible, encompassing the entire color spectrum. Consequently, the test set size was set at 0.003% of the total data. This allowed us to train and validate ANNz2 and NNk using a dataset of 156,103 galaxies. Subsequently, the performance of the method will be assessed using 470 galaxies. The division between these two sets was carried out randomly, ensuring that both samples (train and test) are statistically comparable. Additionally, we opted for an evaluation with 470 galaxies to obtain sufficiently significant information in each redshift range.

The first method to be presented involves the training and evaluation of ANNz2 using the dataset decribed in section 4.1. We employed the random regression mode of ANNz2 and trained it using the colors $U, G, (G-R), (R-I)$, and $(I-z)$ as inputs. The primary objective of this experimental phase was to identify the best combination of parameters that would yield satisfactory results.

ANNz2 offers the user the ability to modify various aspects, including the number of machine learning methods (MLM) to be executed. Given that, ANNz2 reports the best value obtained among the n methods, it is reasonable to assume that increasing the number of MLM during training may contribute to improved results on the same testing set.

We utilized 1, 5, 20, 50, 100, and 150 MLM for these experiments, all employing neural networks. figure 5.1 illustrates the execution time in relation to the number of MLM. While the time increases linearly with the rise in MLM, the performance of ANNz2 regarding the normalized bias and the standard deviation of the normalized bias does not increase in the same manner. This can be observed in figure 5.2. Given this behavior in the number of MLM, it was decided to train the remaining experiments associated with ANNz2 with a maximum of 100 MLM, as the computational cost was not proportionate to the improvement in redshift predictions.

Another training parameter that the user can modify in ANNz2 is the type of MLM with which the dataset will be trained, with neural networks and boosted decision trees (BDTs) being the options to choose from in this phase. Table 5.1 presents the results of evaluating ANNz2 using these two different learning techniques, both with 100 MLM. Both techniques report values of bias and normalized bias

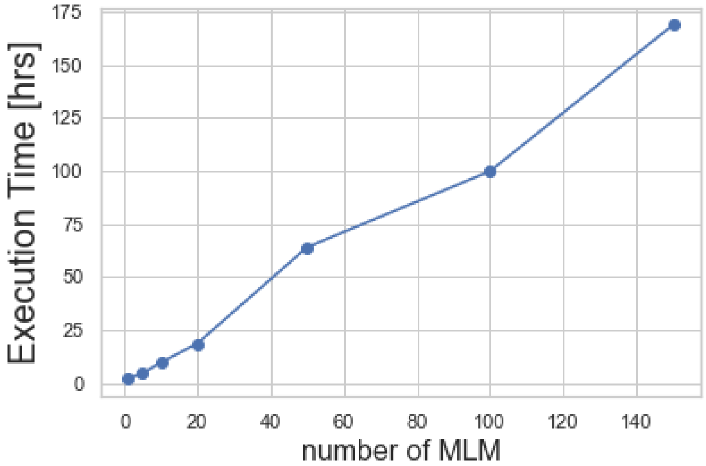


Figure 5.1: Execution Time of ANNz2 as a function of the number of machine learning methods.

ANNz2 **MLM**	**st.dev. of** $n\delta_z$	**SMAD of** $n\delta_z$	**loss of** $n\delta_z$	**RMS of** $n\delta_z$	**% of** O
NN	0.053	0.025	0.047	0.053	2.12
BDT	0.056	0.026	0.053	0.056	2.34

Table 5.1: Metrics of photometric redshift performance of ANNz2 for Neural Networks vs Boosted Decition Trees.

that are practically identical, and they do not have any catastrophic outliers. It can be observed in Table 5.1 that neural networks outperform BDTs in all the reported metrics, as noted in other research papers on the subject (Bilicki et al., 2018). Although the difference in metrics is not significantly large in comparison with their magnitude, it has been demonstrated that neural networks exhibit superior performance. Therefore, the rest of the experiments related to ANNz2 will be trained using neural networks as MLM. The results of these experiments can be seen in figures 5.3 and 5.4, which display the performance of neural networks and decision trees against the spectroscopic data, respectively.

The selection of colors to incorporate into a neural network is also an important discussion. Even though we only have the five bands from SDSS ($ugriz$), it has been demonstrated that a proper choice of colors often leads to better performance than just using the five magnitudes (Li et al., 2007). In the case of ANNz2, the choice of colors was left as provided by Sadeh et al. (2016). However, we wanted to

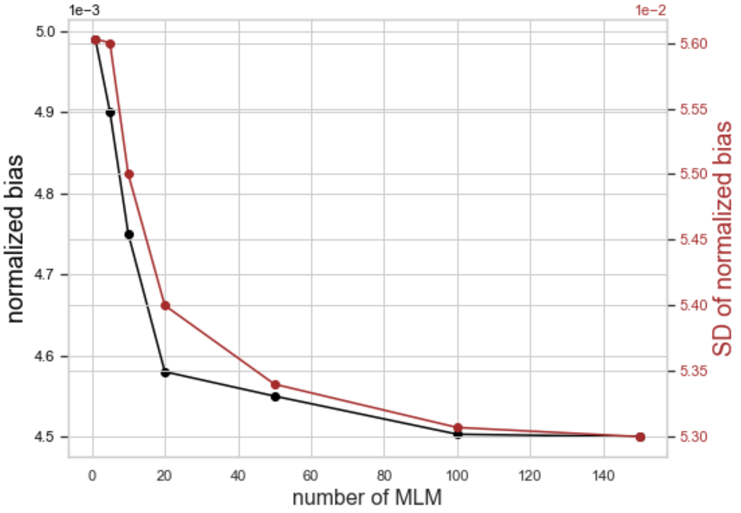


Figure 5.2: Normalized bias for photo$_z$ and its standard deviation as a function of the number of machine learning methods.

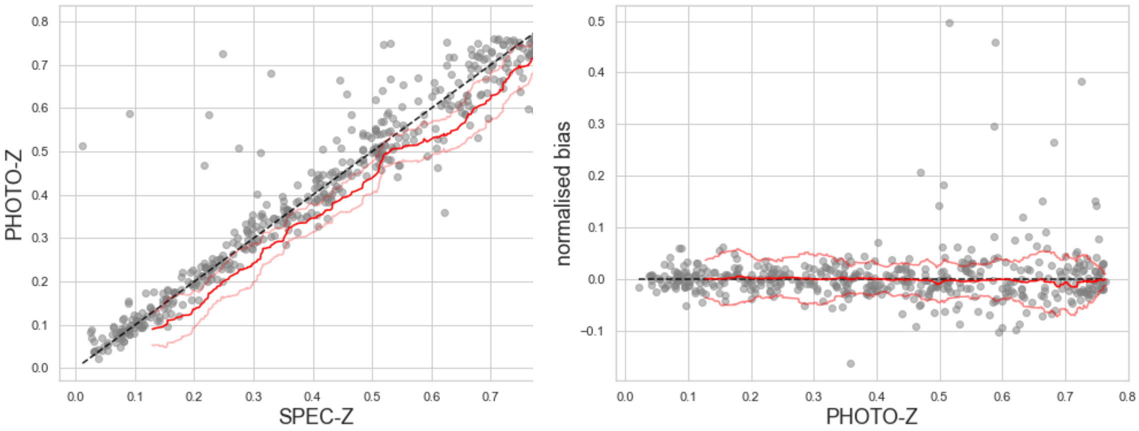
experiment with NNk to see which color space yielded better results.

All color combinations were tested, and among them, Table 5.2 displays those that performed the best when used as input variables for the neural network. As observed, $C = U, (U-G), (G-R), (R-I), (z)$ yields the best results in the majority of metrics. For this reason, this configuration will be used in the training for the rest of the experiments related to NNk.

NNk **Color space**	**st.dev. of** $n\delta_z$	**SMAD of** $n\delta_z$	**loss of** $n\delta_z$	**RMS of** $n\delta_z$	**% of** O
$(U), (U-G), (G-R), (R-I), (z)$	0.054	0.039	0.075	0.054	2.12
$(U-G), (G-R), (R-I), (I-Z), (z)$	0.057	0.041	0.074	0.051	2.15
$(U), (U-G), (G-R), (I), (z)$	0.053	0.038	0.077	0.059	2.16
$(U), (G), (G-R), (I), (z)$	0.052	0.042	0.076	0.052	2.2

Table 5.2: Metrics of photometric redshift performance of NNk for different color combinations.

Having performed the above tests, we compared the redshift determination method of the SDSS against spectroscopic data. The graphical comparison can be seen in figure 5.5, while the metrics are presented in Table 5.4, along with the rest of the metrics for the redshift methods presented earlier. The experiment with the best set of metrics was reported in the table. For ANNz2, this corresponds to the use



(a)

(b)

Figure 5.3: Performance of ANNz2 using neural networks as compared to the spectroscopic samples, the red solid line shows the moving median while the slim lines shows the scatter (SMAD). (a) Direct $\text{spec}_z - \text{photo}_z$ comparison. (b) Normalised bias as a function of photo_z.

of 100 neural networks, and for NNK, it corresponds to training using the colors $C = U, (U - G), (G - R), (R - I), (z)$. The results are presented in bins of photo_z.

Table 5.4 provides a summary of the results in this section, presenting the method with the best performance for each metric across different intervals of photo_z. It is observed that the local linear model performs better across almost all metrics when redshift ranges from $0 < z < 0.4$. However, for higher redshifts, the method that outperforms is ANNz2. For the last two redshift bins, the SDSS method significantly deviates from the performance of the two neural networks. These results suggest that redshift determination using neural networks is particularly robust for redshifts between $0.6 < z < 0.8$ compared to other empirical methods.

5.2 WISE experiments

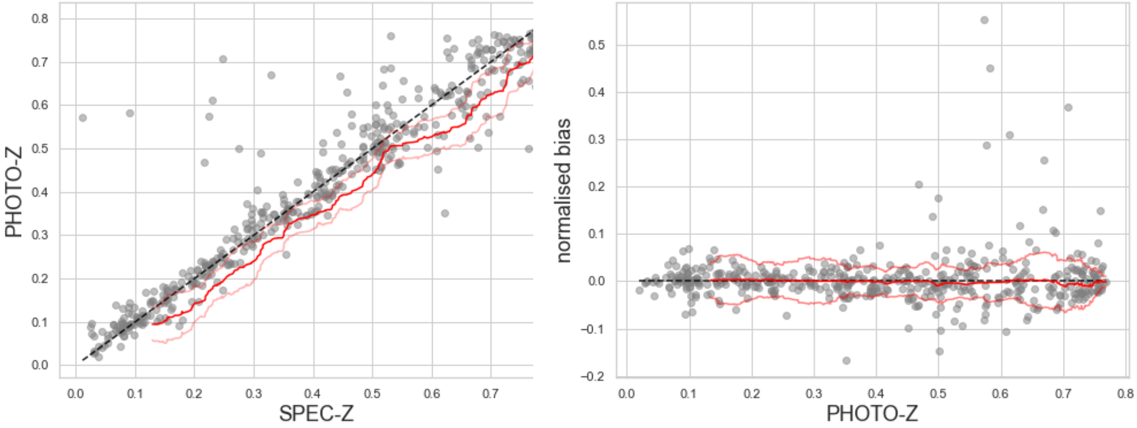
The second phase of this research focuses on the two databases discussed in Section 4.2. The goal is to determine if it's possible to enhance accuracy in more distant redshift ranges using an empirically-based

photo$_z$ **bin**	**Method**			**st. dev of**	**SMAD of**	**Loss of**	**RMS of**
		δ_z	$n\delta_z$	$n\delta_z$	$n\delta_z$	$n\delta_z$	$n\delta_z$
$0 < z < 0.1$	ANNz2	0.029	0.028	0.094	0.015	0.056	0.097
	SDSS method	0.018	0.017	0.063	0.012	0.034	0.065
	NNk	0.044	0.041	0.057	0.022	0.086	0.070
$0.1 < z < 0.2$	ANNz2	3.9×10^{-4}	3.5×10^{-4}	0.019	0.016	0.015	0.019
	SDSS method	-0.001	-0.001	0.016	0.015	0.011	0.016
	NNk	0.011	0.010	0.029	0.019	0.038	0.031
$0.2 < z < 0.3$	ANNz2	0.022	0.018	0.069	0.020	0.057	0.070
	SDSS method	0.012	0.009	0.047	0.019	0.044	0.047
	NNk	0.025	0.020	0.060	0.021	0.064	0.063
$0.3 < z < 0.4$	ANNz2	0.012	0.009	0.046	0.021	0.044	0.047
	SDSS method	0.002	0.001	0.026	0.013	0.022	00.025
	NNk	0.009	0.007	0.046	0.020	0.051	0.046
$0.4 < z < 0.5$	ANNz2	0.009	0.006	0.036	0.024	0.043	0.036
	SDSS method	0.010	0.007	0.049	0.024	0.045	0.049
	NNk	-0.011	-0.007	0.035	0.030	0.048	0.036
$0.5 < z < 0.6$	ANNz2	0.010	0.006	0.041	0.027	0.055	0.041
	SDSS method	-0.017	-0.011	0.047	0.023	.059	0.048
	NNk	-0.036	-0.023	0.040	0.035	0.079	0.046
$0.6 < z < 0.7$	ANNz2	-0.014	-0.009	0.042	0.040	0.062	0.043
	SDSS method	-0.103	-0.062	0.065	0.046	0.188	0.090
	NNk	-0.070	-0.042	0.047	0.044	0.121	0.063
$0.7 < z < 0.8$	ANNz2	-0.036	-0.020	0.029	0.022	0.046	0.035
	SDSS method	-0.123	-0.070	0.086	0.072	0.239	0.111
	NNk	-0.083	-0.047	0.047	0.028	0.114	0.066

Table 5.3: Metrics of photometric redshift performance for the first database $0 < z < 0.8$. Results are presented per bin of NNK, ANNz2 and SDSS method photo$_z$, respectively.

photo$_z$ **bin**			**st. dev of**	**SMAD of**	**Loss of**	**RMS of**
	δ_z	$n\delta_z$	$n\delta_z$	$n\delta_z$	$n\delta_z$	$n\delta_z$
$0 < z < 0.1$	SDSS	SDSS	NNK	SDSS	SDSS	SDSS
$0.1 < z < 0.2$	ANNz2	ANNz2	SDSS	SDSS	SDSS	SDSS
$0.2 < z < 0.3$	SDSS	SDSS	SDSS	SDSS	SDSS	SDSS
$0.3 < z < 0.4$	SDSS	SDSS	SDSS	SDSS	SDSS	SDSS
$0.4 < z < 0.5$	ANNz2	ANNz2	NNK	SDSS	ANNz2	ANNz2
$0.5 < z < 0.6$	ANNz2	ANNz2	NNK	SDSS	ANNz2	ANNz2
$0.6 < z < 0.7$	ANNz2	ANNz2	ANNz2	ANNz2	ANNz2	ANNz2
$0.7 < z < 0.8$	ANNz2	ANNz2	ANNz2	ANNz2	ANNz2	ANNz2

Table 5.4: Comparison of Redshift Determination Methods for Each Metric using the $0 < z < .8$ database. The method with the best performance is presented by photo$_z$ bin.



(a)

(b)

Figure 5.4: Performance of ANNz2 using boosted decision trees as compared to the spectroscopic samples, the red solid line shows the moving median while the slim lines shows the scatter (SMAD). (a) Direct $\text{spec}_z - \text{photo}_z$ comparison. (b) Normalised bias as a function of photo_z.

model supported by artificial intelligence. This will be achieved by incorporating WISE bands along with the $ugriz$ bands from SDSS.

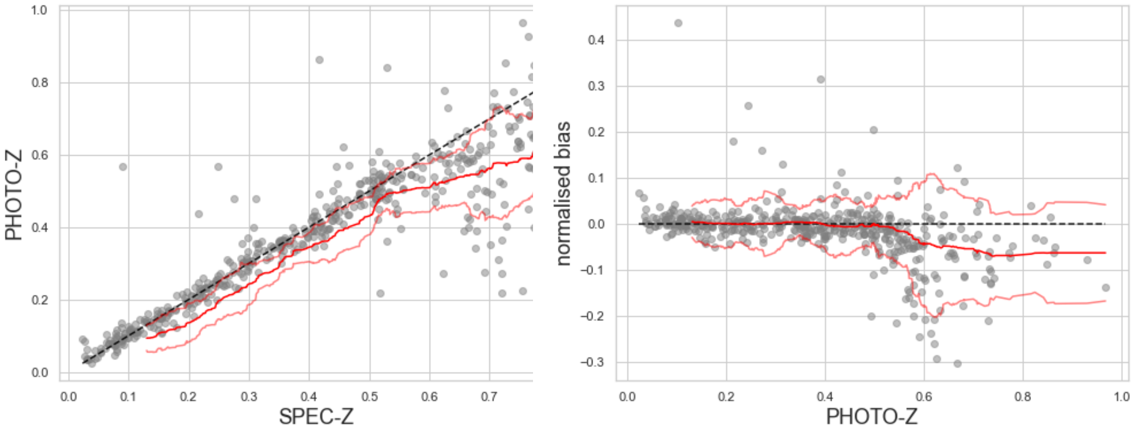
Both databases were divided into training and evaluation sets, ensuring they were statistically equivalent. However, due to differences in the nature of the data, the percentage of galaxies allocated to each evaluation set was different for each one.

The database containing 166,999 galaxies, corresponding to the W1 and W2 bands, was divided, reserving 0.03% for the evaluation set. Meanwhile, the database with 21,302 galaxies, encompassing all the WISE bands, was divided, reserving 0.1%.

These percentages were chosen to ensure that the evaluation sets, on which the metrics will be assessed, contain a sufficient number of redshifts within the intervals of 0.8 to 1.5.

Experiments using W1 and W2 bands

ANNz2 and NNk were trained using the recommendations derived from section 5.1, with the only difference being the inclusion of two additional neurons in the input layer of the neural networks. The



(a)

(b)

Figure 5.5: Performance of the SDSS method as compared to the spectroscopic samples, the red solid line shows the moving median while the slim lines shows the scatter (SMAD). (a) Direct spec$_z$ $-$ photo$_z$ comparison. (b) Nnormalised bias as a function of photo$_z$.

performance of NNk with respect to the spectroscopic data of the galaxy is shown in figure 5.6, while the performance of ANNz2 is displayed in figure 5.8. In Table 5.5, the metrics of both methods are compared against the metrics derived from the SDSS method. It can be observed that empirical models based on neural networks do experience a decrease in precision as the redshift interval increases. However, the decrease in precision is much more catastrophic when it comes to the SDSS method.

Table 5.6 provides a clearer view of the results by presenting the method with the best performance for each metric within the bin interval. Similar to the SDSS experiments, the local linear model performs better within the $z < 0.6$ range. When the redshift exceeds 0.6, NNk emerges as the method with the most promising results. This suggests that the architecture of NNk tends to outperform ANNz2 for larger redshifts.

Experiments using full WISE bands

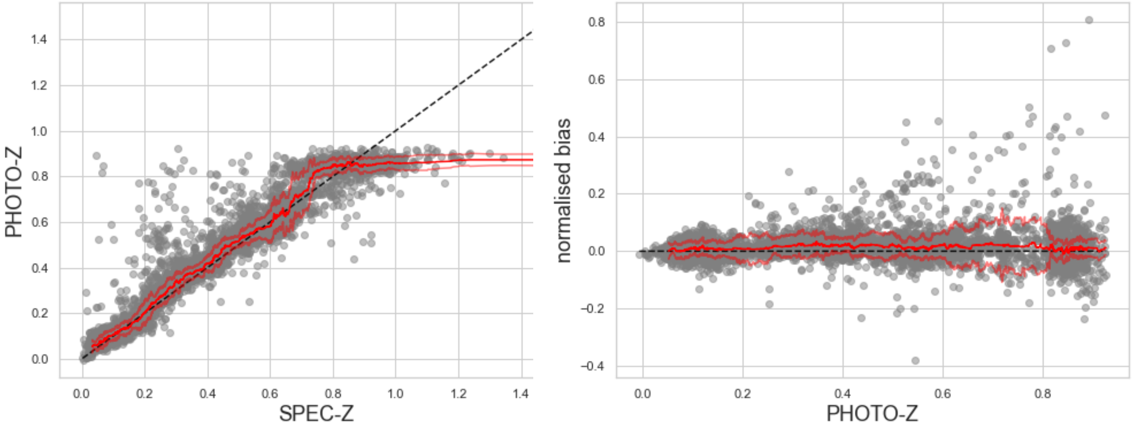
The latest experiments involve a database containing 21,302 galaxies, each with information from both the $ugriz$ bands of SDSS and the W1, W2, W3, and W4 bands corresponding to WISE. figures 5.9 and

photo$_z$ **bin**	**Method**			**st. dev of**	**SMAD of**	**Loss of**	**RMS of**
		δ_z	$n\delta_z$	$n\delta_z$	$n\delta_z$	$n\delta_z$	$n\delta_z$
$0 < z < 0.2$	ANNz2	0.007	0.006	0.045	0.015	0.026	0.046
	SDSS method	0.004	0.004	0.038	0.015	0.022	0.039
	NNk	0.018	0.017	0.059	0.024	0.047	0.061
$0.2 < z < 0.4$	ANNz2	0.019	0.014	0.055	0.021	0.052	0.057
	SDSS method	0.011	0.008	0.041	0.017	0.040	0.042
	NNk	0.047	0.037	0.066	0.025	0.093	0.076
$0.4 < z < 0.6$	ANNz2	0.002	0.002	0.040	0.025	0.046	0.040
	SDSS method	-0.009	-0.006	0.037	0.024	0.045	0.038
	NNk	0.030	0.020	0.041	0.025	0.058	0.046
$0.6 < z < 0.8$	ANNz2	0.013	0.007	0.048	0.047	0.082	0.049
	SDSS method	-0.118	-0.068	0.081	0.057	0.209	0.106
	NNk	0.037	0.022	0.045	0.047	0.088	0.050
$0.8 < z < 1.0$	ANNz2	-0.041	-0.021	0.041	0.035	0.063	0.046
	SDSS method	-0.256	-0.136	0.095	0.095	0.408	0.166
	NNk	-0.025	-0.013	0.034	0.032	0.046	0.037
$1.0 < z < 1.2$	ANNz2	-0.205	-0.098	0.027	0.026	0.298	0.102
	SDSS method	-0.480	-0.232	0.094	0.091	0.650	0.250
	NNk	-0.195	-0.094	0.025	0.024	0.280	0.097
$1.2 < z < 1.4$	ANNz2	-0.350	-0.142	0.012	0.010	0.528	0.159
	SDSS method	-0.797	-0.351	0.067	0.034	0.837	0.356
	NNk	-0.358	-0.159	0.018	0.004	0.526	0.160
$1.4 < z < 1.5$	ANNz2	-0.775	-0.312	0.119	0.124	0.786	0.323
	SDSS method	-0.954	-0.388	0.008	0.008	0.870	0.388
	NNk	-0.764	-0.308	0.101	0.106	0.789	0.316

Table 5.5: Metrics of photometric redshift performance for the database with W1 and W2 wise bands. Results are presented per bin of NNK, ANNz2 and SDSS method photo$_z$, respectively.

photo$_z$ **bin**			**st. dev of**	**SMAD of**	**Loss of**	**RMS of**
	δ_z	$n\delta_z$	$n\delta_z$	$n\delta_z$	$n\delta_z$	$n\delta_z$
$0 < z < 0.2$	SDSS	SDSS	SDSS	SDSS	SDSS	SDSS
$0.2 < z < 0.4$	SDSS	SDSS	SDSS	SDSS	SDSS	SDSS
$0.4 < z < 0.6$	SDSS	SDSS	SDSS	SDSS	SDSS	SDSS
$0.6 < z < 0.8$	ANNz2	ANNz2	NNK	NNK	ANNz2	ANNz2
$0.8 < z < 1.0$	NNK	NNK	NNK	NNK	NNK	NNK
$1.0 < z < 1.2$	NNK	NNK	NNK	NNK	NNK	NNK
$1.2 < z < 1.4$	ANNz2	ANNz2	NNK	NNK	NNK	ANNz2
$1.4 < z < 1.5$	NNK	NNK	SDSS	SDSS	ANNz2	NNK

Table 5.6: Comparison of Redshift Determination Methods for Each Metric using the W1 and W2 WISE database. The method with the best performance is presented by photo$_z$ bin.



(a)

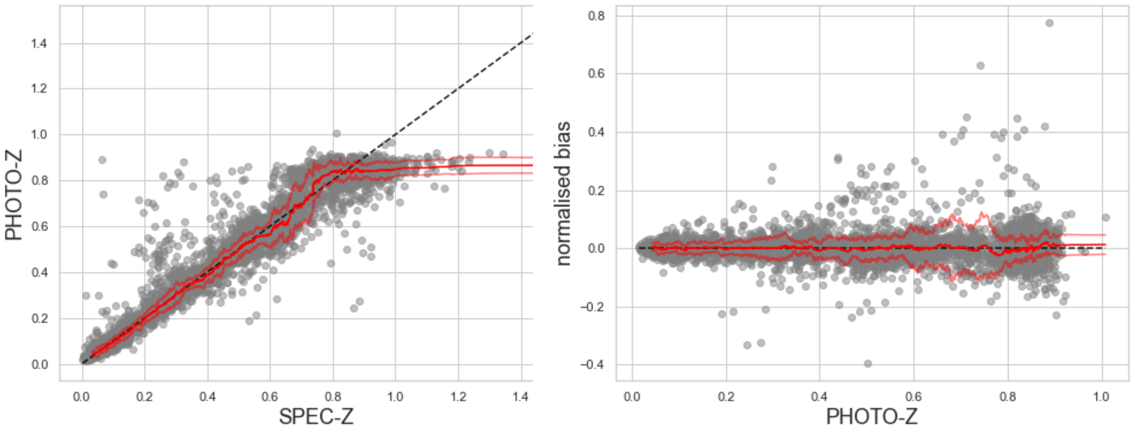
(b)

Figure 5.6: Performance of NNk with W1 and W2 bands as compared to the spectroscopic samples, the red solid line shows the moving median while the slim lines shows the scatter (SMAD). (a) Direct $\text{spec}_z - \text{photo}_z$ comparison. (b) Normalised bias as a function of photo_z.

5.10 illustrate the performance of ANNz2 and NNk against the spectroscopic data, respectively. Table 5.8 presents the performance of both methods compared to the SDSS method, which matches the results obtained from experiments related to W1 and W2. This is because we are working with the same sample. Similarly, Table 5.9 shows the method that demonstrated the best performance for each metric.

The objective of splitting the sample 2 obtained from SDSS according to the WISE bands was to determine whether the additional spectral information provided by W3 and W4 had a greater impact on the performance of a neural network, even if obtaining these values would result in a drastic reduction in the number of observations with which the model would be trained.The comparison between the performance of the neural networks based on the training bands, especially over the redshift interval $0 < z < 1.5$, is shown in Table 5.7.

The Table 5.7 provides very interesting information about the differences between the neural networks. ANNz2 performed better across all metrics (except for the O_c percentage) using all four bands provided by WISE, while NNk showed greater predictive capacity in the database that only included the first two WISE bands.



(a)

(b)

Figure 5.7: Performance of ANNz2 with W1 and W2 bands as compared to the spectroscopic samples, the red solid line shows the moving median while the slim lines shows the scatter (SMAD). (a) direct $\mathrm{spec}_z - \mathrm{photo}_z$ comparison. (b) normalised bias as a function of photo_z.

database	**method**	**st.dev. of** $n\delta_z$	**SMAD of** $n\delta_z$	**loss of** $n\delta_z$	**RMS of** $n\delta_z$	**% of** O	**% of** O_c
SDSS sample 2 with W1 and W2 bands	ANNz2	0.050	0.024	0.054	0.050	1.85	0.239
	NNk	0.056	0.029	0.069	0.059	2.43	0.319
SDSS sample 2 with full WISE bands	ANNz2	0.046	0.022	0.047	0.046	1.70	0.281
	NNk	0.059	0.037	0.079	0.061	2.53	0.375

Table 5.7: Comparison of the photometric redshift performance between the use of different WISE bands and between the two neural networks. The metrics are calculated over the entire interval $0 < z < 1.5$.

Another objective of including the WISE bands in our artificial intelligence-based models was to increase the precision of the methods across the entire color range. Since we know that in redshifts from 0 to 0.8 the $ugriz$ bands are sufficient to obtain good results (see 5.1), the final analysis in this research work was to determine if training a neural network with the $ugriz$ bands as input can improve its precision if is additionally trained with the WISE bands.

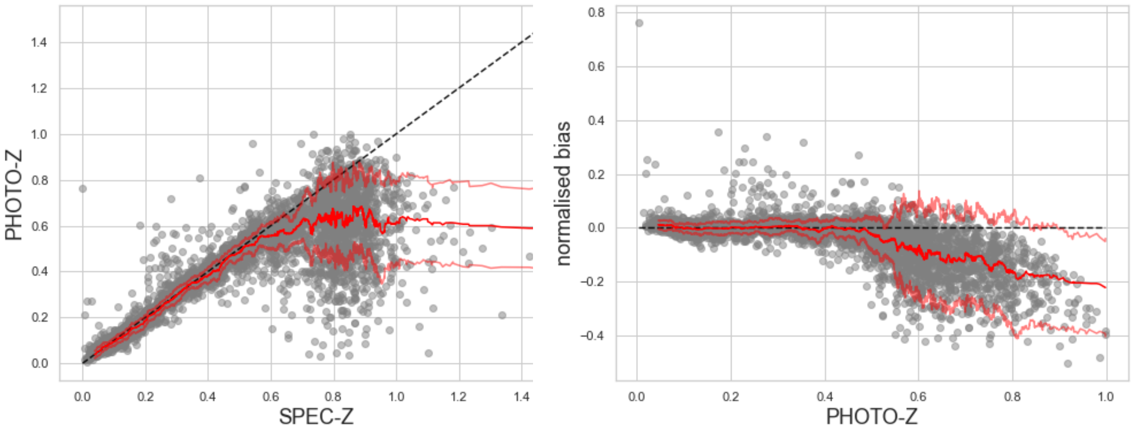
In Table 5.10, we are comparing the performance of our two artificial intelligence models over the redshift range of $0 < z < 0.8$. The presented results were trained and evaluated using the sample 1 discussed in Section 5.1. The outcomes are presented based on the photometric bands with which they were trained.

photo$_z$ **bin**	**Method**			**st. dev of**	**SMAD of**	**Loss of**	**RMS of**
		δ_z	$n\delta_z$	$n\delta_z$	$n\delta_z$	$n\delta_z$	$n\delta_z$
$0 < z < 0.2$	ANNz2	0.004	0.004	0.031	0.016	0.020	0.032
	SDSS method	0.004	0.004	0.038	0.015	0.022	0.039
	NNk	0.024	0.022	0.048	0.028	0.056	0.054
$0.2 < z < 0.4$	ANNz2	0.023	0.018	0.070	0.032	0.102	0.072
	SDSS method	0.011	0.008	0.041	0.017	0.040	0.042
	NNk	0.031	0.024	0.089	0.066	0.161	0.092
$0.4 < z < 0.6$	ANNz2	0.020	0.013	0.051	0.040	0.086	0.053
	SDSS method	-0.009	-0.006	0.037	0.024	0.045	0.038
	NNk	0.051	0.034	0.050	0.041	0.112	0.061
$0.6 < z < 0.8$	ANNz2	-0.006	-0.003	0.046	0.034	0.065	0.046
	SDSS method	-0.118	-0.068	0.081	0.057	0.209	0.106
	NNk	-0.014	-0.008	0.045	0.035	0.068	0.046
$0.8 < z < 1.0$	ANNz2	-0.070	-0.037	0.075	0.037	0.101	0.084
	SDSS method	-0.256	-0.136	0.095	0.095	0.408	0.166
	NNk	-0.090	-0.047	0.074	0.047	0.128	0.088
$1.0 < z < 1.2$	ANNz2	-0.174	-0.084	0.036	0.040	0.240	0.091
	SDSS method	-0.480	-0.232	0.094	0.091	0.650	0.250
	NNk	-0.198	-0.095	0.026	0.019	0.286	0.099
$1.2 < z < 1.4$	ANNz2	-0.365	-0.160	0.022	0.011	0.529	0.161
	SDSS method	-0.797	-0.351	0.067	0.034	0.837	0.356
	NNk	-0.431	-0.189	0.027	0.024	0.609	0.190
$1.4 < z < 1.5$	ANNz2	-0.552	-0.227	0.010	0.010	0.695	0.227
	SDSS method	-0.954	-0.388	0.008	0.008	0.870	0.388
	NNk	-0.576	-0.236	9.3×10^{-4}	9.8×10^{-4}	0.713	0.236

Table 5.8: Metrics of photometric redshift performance for the database with full WISE bands. Results are presented per bin of NNK, ANNz2 and SDSS method photo$_z$, respectively.

photo$_z$ **bin**			**st. dev of**	**SMAD of**	**Loss of**	**RMS of**
	δ_z	$n\delta_z$	$n\delta_z$	$n\delta_z$	$n\delta_z$	$n\delta_z$
$0 < z < 0.2$	ANNz2	ANNz2	ANNz2	SDSS	ANNz2	ANNz2
$0.2 < z < 0.4$	SDSS	SDSS	SDSS	SDSS	SDSS	SDSS
$0.4 < z < 0.6$	SDSS	SDSS	SDSS	SDSS	SDSS	SDSS
$0.6 < z < 0.8$	ANNz2	ANNz2	NNK	ANNz2	ANNz2	ANNz2
$0.8 < z < 1.0$	ANNz2	ANNz2	NNK	ANNz2	ANNz2	ANNz2
$1.0 < z < 1.2$	ANNz2	ANNz2	NNK	NNK	ANNz2	ANNz2
$1.2 < z < 1.4$	ANNz2	ANNz2	ANNz2	ANNz2	ANNz2	ANNz2
$1.4 < z < 1.5$	ANNz2	ANNz2	NNK	NNK	ANNz2	ANNz2

Table 5.9: Comparison of Redshift Determination Methods for Each Metric using the full WISE bands. The method with the best performance is presented by photo$_z$ bin.



(a)

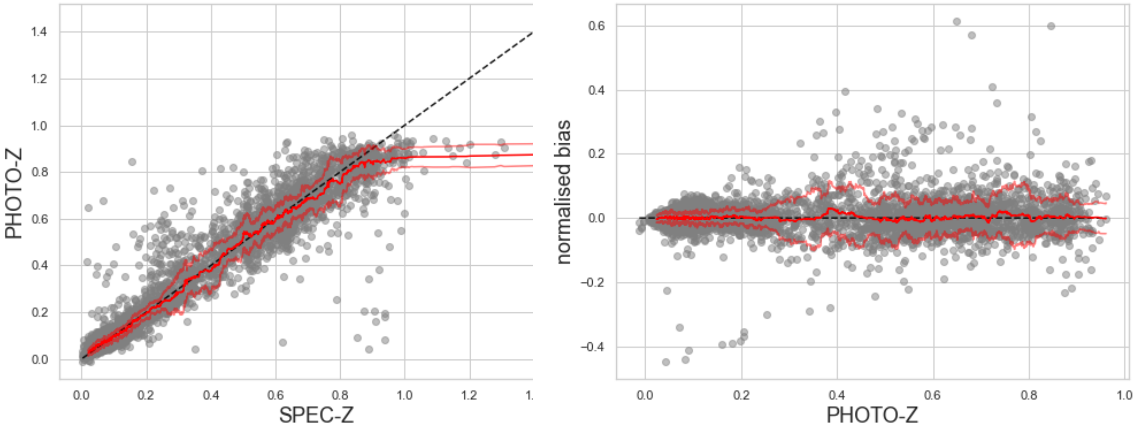
(b)

Figure 5.8: Performance of the SDSS method with W1 and W2 bands as compared to the spectroscopic samples, the red solid line shows the moving median while the slim lines shows the scatter (SMAD). (a) direct spec$_z$ − photo$_z$ comparison. (b) normalised bias as a function of photo$_z$.

From table 5.10, it is deduced that ANNz2 exhibits the most outstanding performance when trained with the four WISE bands. This finding suggests the importance of including this additional information in the training process. Additionally, it can be observed that the SDSS method shows the most modest performance in terms of redshift estimation, presenting inferior results compared to the other evaluated methods.

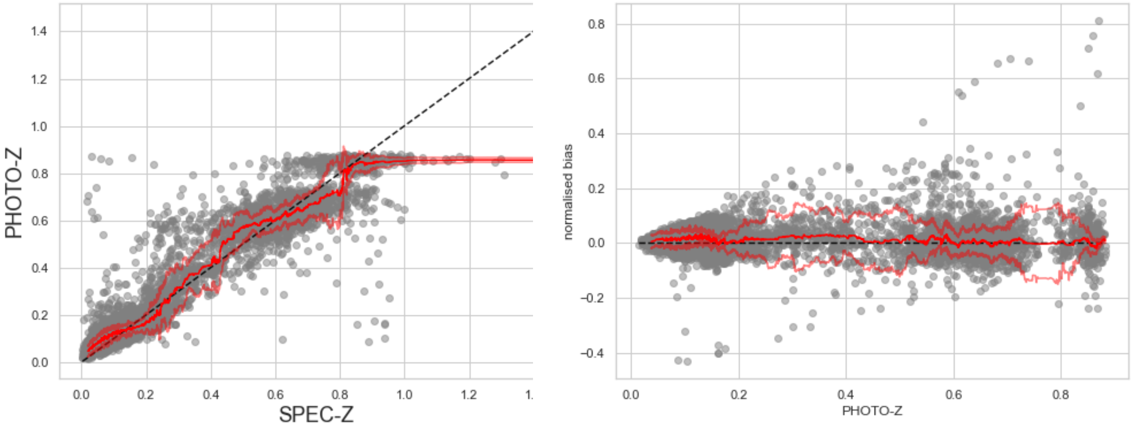
bands	**method**	**st.dev. of**	**SMAD of**	**loss of**	**RMS of**	**% of**
		$n\delta_z$	$n\delta_z$	$n\delta_z$	$n\delta_z$	O
$ugriz$	ANNz2	0.053	0.025	0.047	0.053	2.12
	NNk	0.054	0.039	0.075	0.054	2.12
$ugriz$	ANNz2	0.048	0.022	0.048	0.048	1.80
W1 and W2	NNk	0.047	0.028	0.070	0.060	2.10
$ugriz$	ANNz2	0.042	0.021	0.043	0.043	1.70
W1,W2,W3,W4	NNk	0.056	0.035	0.076	0.059	1.95
	SDSS	0.063	0.026	0.083	0.064	4.93

Table 5.10: Performance of photometric redshift determination methods using NN for various training variables in the $0 < z < 0.8$ range. The results are presented alongside the performance of the SDSS method.



(a) (b)

Figure 5.9: Performance of ANNz2 with the full WISE bands as compared to the spectroscopic samples, the red solid line shows the moving median while the slim lines shows the scatter (SMAD). (a) Direct $\text{spec}_z - \text{photo}_z$ comparison. (b) Normalised bias as a function of photo_z.



(a) (b)

Figure 5.10: Performance of NNk with the full WISE bands as compared to the spectroscopic samples, the red solid line shows the moving median while the slim lines shows the scatter (SMAD). (a) Direct $\text{spec}_z - \text{photo}_z$ comparison. (b) Normalised bias as a function of photo_z.

Chapter 6

Conclusions

Redshifts have proven to be a fundamental tool in the exploration and understanding of the universe. Their study has allowed us to investigate the evolution of galaxies, the expansion of the universe, and the nature of dark energy, as well as gather information about the composition, structure, and age of the cosmos. The determination of redshifts through photometric data has revolutionized our ability to map and comprehend the distribution of galaxies, but it has introduced a new challenge to the scientific community, related to handling large databases and obtaining models that can accurately predict photometric redshifts.

This research presented the analysis of two different techniques for determining photometric redshifts using machine learning methods, with a specific focus on the use of neural networks. The performance of both neural networks was compared with a standard procedure for redshift determination obtained from the Sloan Digital Survey.

In the first part of the results, it was demostrated that the photometric bands $ugriz$ are sufficient for determining redshifts in the range of $0 < z < 0.8$. It was found that the optimal configuration for ANNz2 is to train 100 < machine learning methods based on neural networks. Additionally, the color space that serves best as input for a neural network is composed of the following combination of colors and bands: $(U), (U - G), (G - R), (R - I), (z)$. It can be concluded that neural networks are a superior model for redshifts above $z = 0.4$.

In the second section of the results, the impact of adding some WISE bands in the training of a neural network was determined. The findings indicated that the bands $ugriz$ + WISE are not sufficient to achieve a high level of precision for redshifts greater than 0.8. However, the precision of the neural networks in this interval proved to be significantly superior to the local linear model.

In conclusion, ANNz2 emerged as the best method for determining redshifts, followed by NNk, especially when these algorithms were trained using the $ugriz$+WISE bands. It was demonstrated that this combination of bands yields superior results when applied over a redshift range of $0 < z < 0.8$.

It should be emphasized that despite obtaining promising results within this redshift range, these experiments are contingent on the quantity of galaxies found in the surveys of SDSS and WISE. Therefore, future lines of research may unfold as more spectral information becomes available for a greater number of galaxies.

Future prospects for these types of experiments include expanding the training set, not only in terms of the quantity of observations but also in the number of photometric bands near the infrared range. Experiments with $ugriz$+WISE could benefit from also having information from the five near-IR bands of the Kilo-degree Infrared Galaxy survey (VIKING) (Edge et al., 2013)

In conclusion, the determination of redshifts through neural networks based on photometric data has proven to yield accurate redshift values and can be confidently used to investigate other properties of the cosmos. While the redshift estimation may vary depending on the nature of the survey used and the type of parameters employed, they remain the most convenient way to estimate redshifts today.

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv e-prints*, page arXiv:1603.04467.

Abdalla, F. B., Banerji, M., Lahav, O., and Rashkov, V. (2011). A comparison of six photometric redshift methods applied to 1.5 million luminous red galaxies. , 417(3):1891–1903.

Ahumada, R., Allende Prieto, C., Almeida, A., Anders, F., Anderson, S. F., Andrews, B. H., Anguiano, B., Arcodia, R., Armengaud, E., Aubert, M., Avila, S., Avila-Reese, V., Badenes, C., Balland, C., Barger, K., Barrera-Ballesteros, J. K., Basu, S., Bautista, J., Beaton, R. L., Beers, T. C., Benavides, B. I. T., Bender, C. F., Bernardi, M., Bershady, M., Beutler, F., Bidin, C. M., Bird, J., Bizyaev, D., Blanc, G. A., Blanton, M. R., Boquien, M., Borissova, J., Bovy, J., Brandt, W. N., Brinkmann, J., Brownstein, J. R., Bundy, K., Bureau, M., Burgasser, A., Burtin, E., Cano-Díaz, M., Capasso, R., Cappellari, M., Carrera, R., Chabanier, S., Chaplin, W., Chapman, M., Cherinka, B., Chiappini, C., Doohyun Choi, P., Chojnowski, S. D., Chung, H., Clerc, N., Coffey, D., Comerford, J. M., Comparat, J., da Costa, L., Cousinou, M.-C., Covey, K., Crane, J. D., Cunha, K., Ilha, G. d. S., Dai, Y. S., Damsted, S. B., Darling, J., Davidson, James W., J., Davies, R., Dawson, K., De, N., de la Macorra, A., De Lee, N., Queiroz, A. B. d. A., Deconto Machado, A., de la Torre, S., Dell'Agli, F., du Mas des Bourboux, H., Diamond-Stanic, A. M., Dillon, S., Donor, J., Drory, N., Duckworth, C., Dwelly, T., Ebelke, G., Eftekharzadeh, S., Davis Eigenbrot, A., Elsworth, Y. P., Eracleous, M., Erfanianfar, G., Escoffier, S., Fan, X., Farr, E., Fernández-Trincado, J. G., Feuillet, D., Finoguenov, A., Fofie, P., Fraser-McKelvie, A., Frinchaboy, P. M., Fromenteau, S., Fu, H., Galbany, L., Garcia, R. A., García-Hernández, D. A., Garma Oehmichen, L. A., Ge, J., Geimba Maia, M. A., Geisler, D., Gelfand, J., Goddy, J., Gonzalez-Perez, V., Grabowski, K., Green, P., Grier, C. J., Guo, H., Guy, J., Harding, P., Hasselquist, S., Hawken, A. J., Hayes, C. R., Hearty, F., Hekker, S., Hogg, D. W., Holtzman, J. A., Horta, D., Hou, J., Hsieh, B.-C., Huber, D., Hunt, J. A. S., Ider Chitham, J., Imig, J., Jaber, M., Jimenez Angel, C. E., Johnson, J. A., Jones, A. M., Jönsson, H., Jullo, E., Kim, Y., Kinemuchi, K., Kirkpatrick, Charles C., I., Kite, G. W., Klaene, M., Kneib, J.-P., Kollmeier, J. A., Kong, H., Kounkel, M., Krishnarao, D., Lacerna, I., Lan, T.-W., Lane, R. R., Law, D. R., Le Goff, J.-M., Leung, H. W., Lewis, H., Li, C., Lian, J., Lin, L., Long, D., Longa-Peña, P., Lundgren, B., Lyke, B. W., Mackereth, J. T., MacLeod, C. L., Majewski, S. R., Manchado, A., Maraston, C., Martini, P., Masseron, T., Masters, K. L., Mathur, S., McDermid, R. M., Merloni, A., Merrifield, M., Mészáros, S., Miglio, A., Minniti, D., Minsley, R., Miyaji, T., Mohammad, F. G., Mosser, B., Mueller, E.-M., Muna, D., Muñoz-Gutiérrez, A., Myers, A. D., Nadathur, S., Nair, P., Nandra, K., Correa do Nascimento, J., Nevin, R. J., Newman, J. A., Nidever, D. L., Nitschelm, C., Noterdaeme, P., O'Connell, J. E., Olmstead, M. D., Oravetz, D., Oravetz, A., Osorio, Y., Pace, Z. J., Padilla, N., Palanque-Delabrouille, N., Palicio, P. A., Pan, H.-A., Pan, K., Parker, J., Paviot, R., Peirani, S., Ramŕez, K. P., Penny, S., Percival, W. J., Perez-Fournon, I., Pérez-Ràfols, I., Petitjean, P., Pieri, M. M., Pinsonneault, M., Poovelil, V. J., Povick, J. T., Prakash, A., Price-Whelan, A. M., Raddick, M. J., Raichoor, A., Ray, A., Rembold, S. B., Rezaie, M., Riffel, R. A., Riffel, R., Rix, H.-W., Robin, A. C., Roman-Lopes, A., Román-Zúñiga, C., Rose, B., Ross, A. J., Rossi, G., Rowlands, K., Rubin, K. H. R., Salvato, M., Sánchez, A. G., Sánchez-Menguiano, L., Sánchez-Gallego, J. R., Sayres, C., Schaefer, A., Schiavon, R. P., Schimoia, J. S., Schlafly, E., Schlegel, D., Schneider, D. P., Schultheis, M., Schwope, A., Seo, H.-J., Serenelli, A., Shafieloo, A., Shamsi, S. J., Shao, Z., Shen, S., Shetrone, M., Shirley, R., Silva Aguirre, V., Simon, J. D., Skrutskie, M. F., Slosar, A., Smethurst, R., Sobeck, J., Sodi, B. C., Souto, D., Stark, D. V., Stassun, K. G., Steinmetz, M., Stello, D., Stermer, J., Storchi-Bergmann, T., Streblyanska, A., Stringfellow, G. S., Stutz, A., Suárez, G., Sun, J., Taghizadeh-Popp, M., Talbot, M. S., Tayar, J., Thakar, A. R., Theriault, R., Thomas, D., Thomas, Z. C., Tinker, J., Tojeiro, R., Toledo, H. H., Tremonti, C. A., Troup, N. W., Tuttle, S., Unda-Sanzana, E., Valentini, M., Vargas-González, J., Vargas-Magaña, M., Vázquez-Mata, J. A., Vivek, M., Wake, D., Wang, Y., Weaver, B. A., Weijmans, A.-M., Wild, V., Wilson, J. C., Wilson, R. F., Wolthuis, N., Wood-Vasey, W. M., Yan, R., Yang, M., Yèche, C., Zamora, O., Zarrouk, P., Zasowski, G., Zhang, K., Zhao, C., Zhao, G., Zheng, Z., Zheng, Z., Zhu, G., and Zou, H. (2020). The 16th Data Release of the Sloan Digital Sky Surveys: First Release from the APOGEE-2 Southern Survey and Full Release of eBOSS Spectra. , 249(1):3.

Ball, N. M., Loveday, J., Fukugita, M., Nakamura, O., Okamura, S., Brinkmann, J., and Brunner, R. J. (2004). Galaxy types in the Sloan Digital Sky Survey using supervised artificial neural networks. , 348(3):1038–1046.

Beck, R., Dobos, L., Budavári, T., Szalay, A. S., and Csabai, I. (2016). Photometric redshifts for the SDSS Data Release 12. , 460(2):1371–1381.

Benítez, N. (2000). Bayesian Photometric Redshift Estimation. , 536(2):571–583.

Bilicki, M., Hoekstra, H., Brown, M. J. I., Amaro, V., Blake, C., Cavuoti, S., de Jong, J. T. A., Georgiou, C., Hildebrandt, H., Wolf, C., Amon, A., Brescia, M., Brough, S., Costa-Duarte, M. V., Erben, T., Glazebrook, K., Grado, A., Heymans, C., Jarrett, T., Joudaki, S., Kuijken, K., Longo, G., Napolitano, N., Parkinson, D., Vellucci, C., Verdoes Kleijn, G. A., and Wang, L. (2018). Photometric redshifts for the Kilo-Degree Survey. Machine-learning analysis with artificial neural networks. , 616:A69.

Bolzonella, M., Miralles, J. M., and Pelló, R. (2000). Photometric redshifts based on standard SED fitting procedures. , 363:476–492.

Brunner, R. J., Connolly, A. J., Szalay, A. S., and Bershady, M. A. (1997). Toward More Precise Photometric Redshifts: Calibration Via CCD Photometry. , 482(1):L21–L24.

Bruzual A., G. and Charlot, S. (1993). Spectral Evolution of Stellar Populations Using Isochrone Synthesis. , 405:538.

Budding, E. and Demircan, O. (2007). *Introduction to Astronomical Photometry*, volume 6.

Coleman, G. D., Wu, C. C., and Weedman, D. W. (1980). Colors and magnitudes predicted for high redshift galaxies. , 43:393–416.

Collister, A. A. and Lahav, O. (2004). ANNz: Estimating Photometric Redshifts Using Artificial Neural Networks. , 116(818):345–351.

Connolly, A. J., Csabai, I., Szalay, A. S., Koo, D. C., Kron, R. G., and Munn, J. A. (1995). Slicing Through Multicolor Space: Galaxy Redshifts from Broadband Photometry. , 110:2655.

Csabai, I., Dobos, L., Trencséni, M., Herczegh, G., Józsa, P., Purger, N., Budavári, T., and Szalay, A. S. (2007). Multidimensional indexing tools for the virtual observatory. *Astronomische Nachrichten*, 328(8):852.

de Diego, J. A., Nadolny, J., Bongiovanni, Á., Cepa, J., Lara-López, M. A., Gallego, J., Cerviño, M., Sánchez-Portal, M., Ignacio González-Serrano, J., Alfaro, E. J., Pović, M., Pérez García, A. M., Pérez Martínez, R., Padilla Torres, C. P., Cedrés, B., García-Aguilar, D., González, J. J., González-Otero, M., Navarro-Martínez, R., and Pintos-Castro, I. (2021). Nonsequential neural network for simultaneous, consistent classification, and photometric redshifts of OTELO galaxies. , 655:A56.

de Jong, J. T. A., Verdoes Kleijn, G. A., Erben, T., Hildebrandt, H., Kuijken, K., Sikkema, G., Brescia, M., Bilicki, M., Napolitano, N. R., Amaro, V., Begeman, K. G., Boxhoorn, D. R., Buddelmeijer, H., Cavuoti, S., Getman, F., Grado, A., Helmich, E., Huang, Z., Irisarri, N., La Barbera, F., Longo, G., McFarland, J. P., Nakajima, R., Paolillo, M., Puddu, E., Radovich, M., Rifatto, A., Tortora, C., Valentijn, E. A., Vellucci, C., Vriend, W.-J., Amon, A., Blake, C., Choi, A., Conti, I. F., Gwyn, S. D. J., Herbonnet, R., Heymans, C., Hoekstra, H., Klaes, D., Merten, J., Miller, L., Schneider, P., and Viola, M. (2017). The third data release of the Kilo-Degree Survey and associated data products. , 604:A134.

Edge, A., Sutherland, W., Kuijken, K., Driver, S., McMahon, R., Eales, S., and Emerson, J. P. (2013). The VISTA Kilo-degree Infrared Galaxy (VIKING) Survey: Bridging the Gap between Low and High Redshift. *The Messenger*, 154:32–34.

Firth, A. E., Lahav, O., and Somerville, R. S. (2003). Estimating photometric redshifts with artificial neural networks. , 339(4):1195–1202.

Fukugita, M., Ichikawa, T., Gunn, J. E., Doi, M., Shimasaku, K., and Schneider, D. P. (1996). The Sloan Digital Sky Survey Photometric System. , 111:1748.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Gunn, J. E., Siegmund, W. A., Mannery, E. J., Owen, R. E., Hull, C. L., Leger, R. F., Carey, L. N., Knapp, G. R., York, D. G., Boroski, W. N., Kent, S. M., Lupton, R. H., Rockosi, C. M., Evans, M. L., Waddell, P., Anderson, J. E., Annis, J., Barentine, J. C., Bartoszek, L. M., Bastian, S., Bracker, S. B., Brewington, H. J., Briegel, C. I., Brinkmann, J., Brown, Y. J., Carr, M. A., Czarapata, P. C., Drennan, C. C., Dombeck, T., Federwitz, G. R., Gillespie, B. A., Gonzales, C., Hansen, S. U., Harvanek, M., Hayes, J., Jordan, W., Kinney, E., Klaene, M., Kleinman, S. J., Kron, R. G., Kresinski, J., Lee, G., Limmongkol, S., Lindenmeyer, C. W., Long, D. C., Loomis, C. L., McGehee, P. M., Mantsch, P. M., Neilsen, Eric H., J., Neswold, R. M., Newman, P. R., Nitta, A., Peoples, John, J., Pier, J. R., Prieto, P. S., Prosapio, A., Rivetta, C., Schneider, D. P., Snedden, S., and Wang, S.-i. (2006). The 2.5 m Telescope of the Sloan Digital Sky Survey. , 131(4):2332–2359.

Géron, A. (2017). *Hands-on machine learning with scikit-learn and tensorflow : concepts, tools, and techniques to build intelligent systems.* O'Reilly Media.

Hildebrandt, H., Arnouts, S., Capak, P., Moustakas, L. A., Wolf, C., Abdalla, F. B., Assef, R. J., Banerji, M., Benítez, N., Brammer, G. B., Budavári, T., Carliles, S., Coe, D., Dahlen, T., Feldmann, R., Gerdes, D., Gillis, B., Ilbert, O., Kotulla, R., Lahav, O., Li, I. H., Miralles, J. M., Purger, N., Schmidt, S., and Singal, J. (2010). PHAT: PHoto-z Accuracy Testing. , 523:A31.

Hoecker, A., Speckmayer, P., Stelzer, J., Therhaag, J., von Toerne, E., Voss, H., Backes, M., Carli, T., Cohen, O., Christov, A., Dannheim, D., Danielowski, K., Henrot-Versille, S., Jachowski, M., Kraszewski, K., Krasznahorkay, A., J., Kruk, M., Mahalalel, Y., Ospanov, R., Prudent, X., Robert, A., Schouten, D., Tegenfeldt, F., Voigt, A., Voss, K., Wolter, M., and Zemla, A. (2007). TMVA - Toolkit for Multivariate Data Analysis. *arXiv e-prints*, page physics/0703039.

Jones, E., Do, T., Boscoe, B., Singal, J., Wan, Y., and Nguyen, Z. (2023). Photometric Redshifts for Cosmology: Improving Accuracy and Uncertainty Estimates Using Bayesian Neural Networks. *arXiv e-prints*, page arXiv:2306.13179.

Karttunen, H., Kröger, P., Oja, H., Poutanen, M., and Donner, K. J. (2017). *Fundamental Astronomy*.

Koo, D. C. (1999). Overview - Photometric Redshifts: A Perspective from an Old-Timer[!] on their Past, Present, and Potential. In Weymann, R., Storrie-Lombardi, L., Sawicki, M., and Brunner, R., editors, *Photometric Redshifts and the Detection of High Redshift Galaxies*, volume 191 of *Astronomical Society of the Pacific Conference Series*, page 3.

Li, L.-L., Zhang, Y.-X., Zhao, Y.-H., and Yang, D.-W. (2007). Estimating Photometric Redshifts with Artificial Neural Networks and Multi-Parameters. , 7(3):448–456.

Oyaizu, H., Lima, M., Cunha, C. E., Lin, H., and Frieman, J. (2008). Photometric Redshift Error Estimators. , 689(2):709–720.

Pasquet, J., Bertin, E., Treyer, M., Arnouts, S., and Fouchez, D. (2019). Photometric redshifts from SDSS images using a convolutional neural network. , 621:A26.

Pietrinferni, A., Cassisi, S., Salaris, M., Percival, S., and Ferguson, J. W. (2009). A Large Stellar Evolution Database for Population Synthesis Studies. V. Stellar Models and Isochrones with CNONa Abundance Anticorrelations. , 697(1):275–282.

Puschell, J. J., Owen, F. N., and Laing, R. A. (1982). Near-infrared photometry of distant radio galaxies - Spectral flux distributions and redshift estimates. , 257:L57–L61.

Sadeh, I., Abdalla, F. B., and Lahav, O. (2016). ANNz2: Photometric Redshift and Probability Distribution Function Estimation using Machine Learning. , 128(968):104502.

Searle, L., Sargent, W. L. W., and Bagnuolo, W. G. (1973). The History of Star Formation and the Colors of Late-Type Galaxies. , 179:427–438.

Singal, J., Shmakova, M., Gerke, B., Griffith, R. L., and Lotz, J. (2011). The Efficacy of Galaxy Shape Parameters in Photometric Redshift Estimation: A Neural Network Approach. , 123(903):615.

Tinsley, B. M. (1972). Stellar Evolution in Elliptical Galaxies. , 178:319–336.

Walcher, J., Groves, B., Budavári, T., and Dale, D. (2011). Fitting the integrated spectral energy distributions of galaxies. , 331:1–52.

Weaver, W. B. (2000). Automatic Spectral Classification of Unresolved Binary Stars. In *American Astro-*

nomical Society Meeting Abstracts, volume 197 of *American Astronomical Society Meeting Abstracts*, page 15.17.

Weiss, A. and Schlattl, H. (2008). GARSTEC—the Garching Stellar Evolution Code. The direct descendant of the legendary Kippenhahn code. , 316(1-4):99–106.

Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., Ressler, M. E., Cutri, R. M., Jarrett, T., Kirkpatrick, J. D., Padgett, D., McMillan, R. S., Skrutskie, M., Stanford, S. A., Cohen, M., Walker, R. G., Mather, J. C., Leisawitz, D., Gautier, Thomas N., I., McLean, I., Benford, D., Lonsdale, C. J., Blain, A., Mendez, B., Irace, W. R., Duval, V., Liu, F., Royer, D., Heinrichsen, I., Howard, J., Shannon, M., Kendall, M., Walsh, A. L., Larsen, M., Cardon, J. G., Schick, S., Schwalm, M., Abid, M., Fabinsky, B., Naes, L., and Tsai, C.-W. (2010). The Wide-field Infrared Survey Explorer (WISE): Mission Description and Initial On-orbit Performance. , 140(6):1868–1881.

York, D. G., Adelman, J., Anderson, John E., J., Anderson, S. F., Annis, J., Bahcall, N. A., Bakken, J. A., Barkhouser, R., Bastian, S., Berman, E., Boroski, W. N., Bracker, S., Briegel, C., Briggs, J. W., Brinkmann, J., Brunner, R., Burles, S., Carey, L., Carr, M. A., Castander, F. J., Chen, B., Colestock, P. L., Connolly, A. J., Crocker, J. H., Csabai, I., Czarapata, P. C., Davis, J. E., Doi, M., Dombeck, T., Eisenstein, D., Ellman, N., Elms, B. R., Evans, M. L., Fan, X., Federwitz, G. R., Fiscelli, L., Friedman, S., Frieman, J. A., Fukugita, M., Gillespie, B., Gunn, J. E., Gurbani, V. K., de Haas, E., Haldeman, M., Harris, F. H., Hayes, J., Heckman, T. M., Hennessy, G. S., Hindsley, R. B., Holm, S., Holmgren, D. J., Huang, C.-h., Hull, C., Husby, D., Ichikawa, S.-I., Ichikawa, T., Ivezić, Ž., Kent, S., Kim, R. S. J., Kinney, E., Klaene, M., Kleinman, A. N., Kleinman, S., Knapp, G. R., Korienek, J., Kron, R. G., Kunszt, P. Z., Lamb, D. Q., Lee, B., Leger, R. F., Limmongkol, S., Lindenmeyer, C., Long, D. C., Loomis, C., Loveday, J., Lucinio, R., Lupton, R. H., MacKinnon, B., Mannery, E. J., Mantsch, P. M., Margon, B., McGehee, P., McKay, T. A., Meiksin, A., Merelli, A., Monet, D. G., Munn, J. A., Narayanan, V. K., Nash, T., Neilsen, E., Neswold, R., Newberg, H. J., Nichol, R. C., Nicinski, T., Nonino, M., Okada, N., Okamura, S., Ostriker, J. P., Owen, R., Pauls, A. G., Peoples, J., Peterson, R. L., Petravick, D., Pier, J. R., Pope, A., Pordes, R., Prosapio, A., Rechenmacher, R., Quinn, T. R., Richards, G. T., Richmond, M. W., Rivetta, C. H., Rockosi, C. M., Ruthmansdorfer, K., Sandford, D., Schlegel, D. J., Schneider, D. P., Sekiguchi, M., Sergey, G., Shimasaku, K., Siegmund, W. A., Smee, S., Smith, J. A., Snedden, S., Stone, R., Stoughton, C., Strauss, M. A., Stubbs, C., SubbaRao, M., Szalay, A. S., Szapudi, I., Szokoly, G. P., Thakar, A. R., Tremonti, C., Tucker, D. L., Uomoto, A., Vanden Berk, D., Vogeley, M. S., Waddell, P., Wang, S.-i., Watanabe, M., Weinberg, D. H., Yanny, B., Yasuda, N., and SDSS Collaboration (2000). The Sloan Digital Sky Survey: Technical Summary. , 120(3):1579–1587.

Zhang, Y. and Zhao, Y. (2007). Preselect Quasar Candidates by Automated Methods. In Ho, L. C. and Wang, J. W., editors, *The Central Engine of Active Galactic Nuclei*, volume 373 of *Astronomical Society of the Pacific Conference Series*, page 734.