Machine Learning In Multi-Wavelength Galaxy/Quasar Evolution: Photometric Redshift Estimation

Walter Silima

A thesis presented to the University of Cape Town in full fulfilment of the degree: Honours in Astrophysics and Space Science.

Supervisor: Prof. Mattia Vaccari.
Co-supervisor: Mr Chaka Mofokeng.

November, 2020

Abstract

Photometric redshift estimation is currently the most powerful and efficient way to estimate the distances to the vast majority of extragalactic sources we observe in the distant universe. The exponential data avalanche continues and this will require low cost, fast and efficient data-driven methods to analyse and make predictions from the data. In this study, we use supervised machine learning algorithms to estimate photometric redshifts of galaxies and quasars found in the spectroscopic sample of the Sloan Digital Sky Survey data release 16 (SDSS DR16). We perform K-Nearest Neighbour (KNN) and Random Forest (RF) regression to estimate the photometric redshifts of galaxies and quasars on the basis of their broad-band photometry and using the spectroscopic redshifts for training the algorithms. The RF algorithm achieved, for galaxies, the NMAD and RMS of 0.0136 and 0.0187 respectively while the KNN algorithm attained the NMAD and RMS of 0.0152 and 0.0204 respectively. For quasars the NMAD and RMS were found to be 0.0368 and 0.0931 respectively by the RF algorithm and 0.0384 and 0.1063 by the KNN algorithm. Our photometric redshift estimates are thus reasonably accurate though the accuracy drops at higher redshifts. The random forest achieved a 3σ outlier rate of $P_0 = 0.763\%$ and $P_0 = 2.025\%$ for galaxies and quasars respectively, and the KNN algorithm obtained a 3σ outlier rate of $P_0 = 0.7385\%$ and $P_0 = 2.137\%$ for galaxies and quasars respectively. The study shows that machine learning algorithms can handle large amounts of data very efficiently and with a high degree of accuracy. Upcoming wide-field surveys with the LSST and the SKA are going to observe billions of galaxies and these vast amounts of data will increasingly require machine learning techniques to analyse the data effectively.

Contents

1 Introduction **1**
- 1.1 Redshifts . . . 1
  - 1.1.1 Spectroscopic redshifts . . . 2
  - 1.1.2 Photometric redshifts . . . 2
  - 1.1.3 Challenges in photometric redshifts . . . 3
- 1.2 Galaxies and Quasars . . . 3
  - 1.2.1 Galaxies . . . 3
  - 1.2.2 Quasars . . . 4

2 Multi-wavelength photometry **4**
- 2.1 SDSS . . . 5
- 2.2 WISE . . . 6
- 2.3 SDSS-WISE data sample . . . 8

3 Machine Learning **10**
- 3.1 Regression . . . 10
  - 3.1.1 K-Nearest Neighbour Regression . . . 11
  - 3.1.2 Random Forest Regression . . . 11

4 Results **12**
- 4.1 K-Nearest Neighbour photometric redshifts . . . 13
- 4.2 Random forest photometric redshifts . . . 16

5 Conclusions **19**

6 Acknowledgements **20**

7 Source Code **20**

List of Tables

1 KNN regression statistics for galaxies and quasars 15
2 RF regression statistics for galaxies and quasars 19

1 Introduction

In astronomy and cosmology, the distance to an extragalactic source must be known or determined before one can infer any physical quantities of the source. Photometric redshift estimation is currently the most powerful and efficient way in which the distances to the vast majority of extragalactic sources we observe can be determined. Photometric redshifts are generally estimated using the colours of galaxies derived from their measured fluxes at different wavelengths, or brod-band photometry. Photometric redshifts, compared to the traditional spectroscopic redshift estimation, is cheap and offers the advantage that all the sources detected in the photometric images can have a distance estimate. The redshifts are important in many sub-fields of astronomy, e.g. in studies of the formation and evolution of galaxies and active galactic nuclei (AGN) as well as in precision cosmology (Salvato et al., 2019).
In this project, we employ supervised machine learning algorithms to estimate the photometric redshifts for galaxies and quasars that are found in the spectroscopic sample of the Sloan Digital Sky Survey (SDSS) Data Release 16 (DR16). The photometric redshifts are estimated based on the photometric catalogues and a training set of spectroscopic redshifts, which due to their accuracy are used as the 'true' redshifts. We also explore how the photometric catalogues complement the photometric redshifts compared to the spectroscopic redshifts.

1.1 Redshifts

Most of the information that we currently have about the universe comes to us as electromagnetic signals, or 'light', emitted from distant astronomical objects (Sparke and Gallagher III, 2007). We must interpret the information that is provided to us correctly and this includes understanding how the light travels through the expanding space-time. When light travels through the expanding space it is 'stretched' so that the wavelength observed is longer than the one that was emitted by the astronomical source and this is known as cosmological redshift. The redshift is dependent on the overall expansion of space over the journey of light from a source to an observer. The redshift is most commonly expressed with the unit-less parameter z and is defined as the fractional change in the observed wavelength,

$$z = \frac{\lambda_{obs} - \lambda_{em}}{\lambda_{em}} \tag{1}$$

where λ_{obs} is the detected/observed wavelength and λ_{em} is the wavelength emitted by the astronomical source. Redshift is useful when we want to measure the distance to astronomical sources (Baldry, 2018). Redshift gives a good estimate of the expansion of the universe and how long light has travelled to reach us.

Edwin Hubble discovered that the universe is expanding in the early 20th century exploiting spectroscopic observations to carry out redshift measurements. The galaxies we observe are receding from us and are also moving away from each other as the cosmic space expands. In his paper, Hubble showed that the galaxies that are further away from us the faster they move away from us, while those closer to us recede more slowly from us (Bahcall, 2015).

$$v = H_0 \times d \tag{2}$$

where v is the recession velocity, $H_0 = 70km/s/Mpc$ is Hubble's constant and d is the distance. The galaxies that are further away thus have a larger redshift due to their higher recession velocity while closer galaxies have a smaller redshift because they are receding more slowly from us. There are two main methods in the literature that are used to measure the redshifts of large numbers of astronomical sources; spectroscopic and photometric (Bahcall, 2015).

1.1.1 Spectroscopic redshifts

Spectroscopic redshifts are measured through cross-correlating a library of galaxy templates with observed spectra (Cunha et al., 2014). Spectroscopy is a traditional and accurate way to measure redshifts. In this case, the redshifts are calculated by observing a shift in the absorption lines or emission lines of the target galaxy spectra when cross-correlated to the template spectra. Since the galaxies are moving away from us the absorption or emission lines are usually shifted to the lower energy side of the spectrum. Although this method is effective in determining the redshifts of the galaxies, it is time-consuming and expensive, which means that when dealing with a larger set of astronomical data we may not be able to obtain a spectroscopic redshift for every object that is in the dataset (Chong and Yang, 2019).

1.1.2 Photometric redshifts

The photometric redshift of an astronomical source is estimated on the basis of broad-band photometric observations. Photometric redshifts are highly efficient, low cost and can be applied to a large number of celestial objects with a minimal telescope observing time (Han et al., 2020). Two classes of methods can be used to compute the photometric redshifts; the template-fitting method and the machine learning method or data-driven method (Chong and Yang, 2019).
The template-fitting method is the one in which the redshift is estimated by modelling the physical processes that drive the light emission. The template-fitting method requires observations that cover the whole electromagnetic spectrum to achieve better accuracy, as such this model is not very robust for the large sky surveys such as SDSS where the number of photometric bands is limited and covers only the optical part of the electromagnetic spectrum.
The machine learning method, also called the data-driven method, utilizes supervised machine learning algorithms to determine the redshift of the celestial object. These algorithms are trained over the accurately measured spectroscopic redshifts with the associated colours to determine the correlation between the colours and the redshifts (Tarrío and Zarattini, 2020). The data-driven method offers the advantage that all the effects of dust and spectral resolutions that are present in the training set are implicitly accounted for. The disadvantage is that the methods only predict accurately the redshifts of the objects that are similar to the ones that are contained in the training set.
Estimating redshifts this way was applied by Beck et al. (2016b), where they created a photometric redshift database of SDSS data release 12 (SDSS DR12). A local linear regression was applied to estimate the redshift and its error by considering a training set of 1,976,978 sources with spectroscopic redshifts.
In this study, we adopt a method similar to the one used by Beck et al. (2016b) and use K-Neighbours Nearest Regressor (KNNR) and Random Forest Regressor (RFR) to estimate

the redshift of galaxies and quasars found in the SDSS DR16.

1.1.3 Challenges in photometric redshifts

The accuracy of photometric redshifts is affected in two ways which are independent of the method that is used to estimates the redshifts. Different galaxy types overlap in the photometric colour space and the measurement errors in photometry. The first factor is purely a physical phenomenon, in that when galaxies of different morphological types possess the same colours at different redshifts, there is not enough information to tell what the redshift is. When this occurs, we have a degeneracy in the colour-redshift relations. Overlapping can be reduced by using observations obtained in a large number of photometric filters covering a large wavelength range. The latter factor is a major problem in photometric redshift estimation, in that when measurements errors are not accurately estimated, the assumption that Gaussian errors are uncorrelated does not hold. Photometric errors can be reduced and more importantly estimated more accurately by using a better camera or telescope (Beck et al., 2016a).

1.2 Galaxies and Quasars

1.2.1 Galaxies

A galaxy is a huge collection of billions of stars, gas and dust that are bound together by gravity. Galaxies appear in the sky as diffuse sources thousands of light-years across in diameter (Sparke and Gallagher III, 2007). Galaxies are in many ways the basic building blocks of the universe, complex systems made up of many separate interacting components such as stars, gas and dust. Galaxies can also exist in galaxy clusters bound together by the force of gravity in space just like stars are bound in star clusters (Karttunen et al., 2016). Galaxies can be classified according to their morphological and/or physical properties. In 1926 Hubble developed an influential classification scheme grouping galaxies into three main groups; *ellipticals*, *spirals* and *irregulars*. This classification scheme, known as Hubble sequence, is based on the overall appearance of the galaxy (Carroll and Ostlie, 2006). *Ellipticals* are galaxies that are round, featureless, smooth and do not have major spiral arms or dust. *Spirals* are the most common galaxies in the universe and are characterised by a rotating disk that contains stars, gas, dust, and a central bulge. They are well known for their continuous spiral arms which contain mostly blue stars. Lastly, *irregulars* shows no regular pattern in their structures and are mostly blue (Sparke and Gallagher III, 2007). Spiral galaxies in the Hubble sequence are further divided into normal spirals and barred spirals. Galaxies somewhat intermediate in shape between ellipticals and spirals are also known as *lenticulars* or S0 (Carroll and Ostlie, 2006). Figure 1 below illustrates Hubble's "tuning fork" diagram where the two types of spiral galaxies are subdivided into two groups.

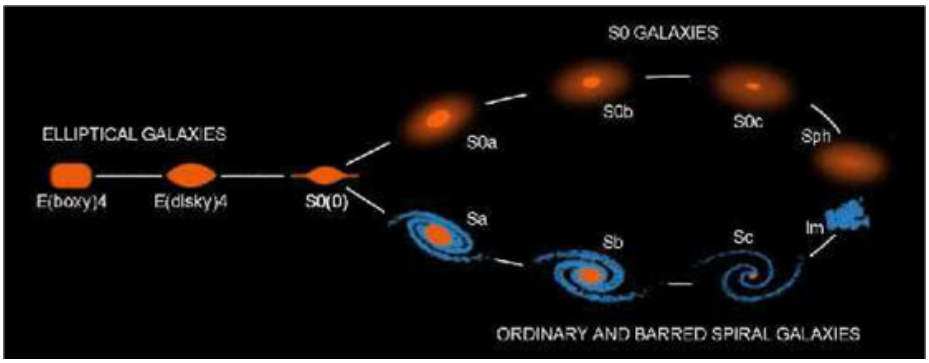


Figure 1: Hubble's tuning-fork diagram of galaxy classifications. The diagram shows a sequence of elliptical galaxies, lenticulars and spiral galaxies. The sequence of ellipticals depends entirely on the shape of the galaxy. The spiral subdivisions are classified as a, b or c depending on how tight are the spiral arms and how large the is bulge relative to the disk. On the far right of the diagram are the irregular galaxies (Kormendy and Bender, 2012).

The physical properties of galaxies are also significant to understanding their formation and evolution. Galaxies are abundant in the universe, as they can be observed over large distances and time scales. Galaxies can therefore be used to understand the evolution of the universe as a whole (Mo et al., 2010).

1.2.2 Quasars

Quasars belong to the class of galaxies hosting an Active Galactic Nucleus (AGN), i.e. a compact region at the centre of the galaxy that has an excess of non-stellar luminosity over at least some part of the electromagnetic spectrum. Quasars were first found in the 1950s, as astronomers were busy classifying the vast number of radio sources that were being discovered at the time. The first identified quasar was a 16-magnitude star-like object whose spectrum displayed a broad emission line that astronomers could not identify with any known molecule or element (later to be classified as a Balmer emission line of hydrogen). Quasars are so bright and far away that in optical images most of them appear as bright star-like objects. The most luminous quasars are of the orders 10^5 more energetic than our Milky Way galaxy (Carroll and Ostlie, 2006).

2 Multi-wavelength photometry

Photometry is a technique used in astronomy that is concerned with the measurement of flux or intensity of light radiated by an astronomical source. (wiki-photometry). Multi-wavelength photometry can be used to describe the general shape of the spectrum of an astronomical source sampling the brightness in several different broad-band filters. The filters are selected in such a way that the maximum amount of light is admitted while still providing valuable astronomical information. Using multi-wavelength photometry we can e.g. extract the temperature, luminosity and the metal content of a star or the stellar mass, star formation rate and distance of an extragalactic source (Chromey, 2016).

2.1 SDSS

The Sloan Digital Sky Survey (SDSS) carried out an imaging and spectroscopic survey covering a large fraction of the sky. The survey makes use of a 2.5-meter telescope. The dedicated telescope is equipped with a large-format mosaic CCD camera to image the sky in five optical bands ($u'g'r'i'z'$) as shown in Figure 2 below. The telescope also uses two digital spectrographs to obtain the spectra of large numbers of galaxies selected from the imaging data. The SDSS telescope is located at the Apache Point Observatory, New Mexico. The telescope achieves a distortion-free wide field of $3^\circ \times 3^\circ$ for images by using a large secondary mirror and two corrector lens (York et al., 2000). The imaging survey which uses the 5 broad-band filters mentioned above has a detection limit of about $u' = 22.0\ mag$, $g' = 22.2\ mag$, $r' = 22.2\ mag$, $r' = 21.3\ mag$, and $r' = 20.5\ mag$ in the AB system (Djorgovski et al., 2013). Figure 2 below shows the response functions of the five bandpasses that are used for photometric imaging.

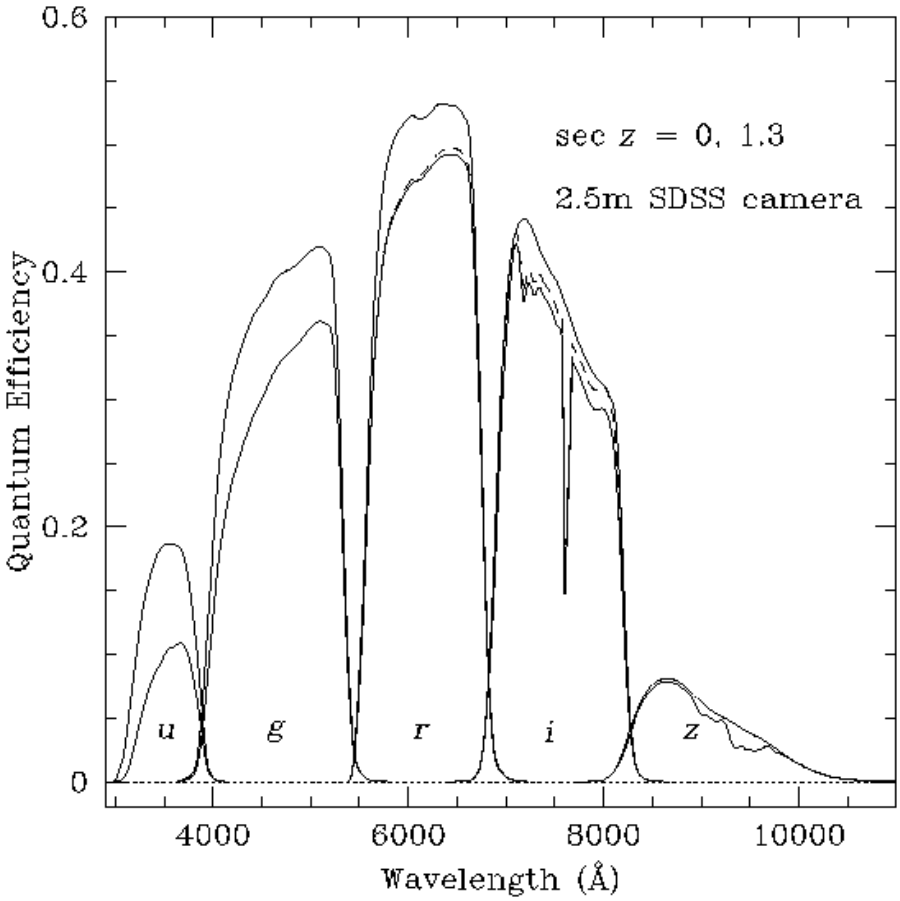


Figure 2: The filter response in u, g, r, i, z of the 2.5 m SDSS telescope. The upper curve in each curve indicates the filter response in the absence of atmospheric extinction while the lower curve assumes an air mass of 1.3. The dashed response curves in the r and i bands represent the effect of the scattering within the thin chips which does not affect the extended objects (Stoughton et al., 2002)

SDSS regular operations began in 2000 and has advanced through the several phases.

The initial phase was SDSS-I (2000-2005) which covered about $8,000 deg^2$ of the sky. The second phase of the survey was SDSS-II (2005-2008) which comprised the *Sloan Legacy Survey* that covered about $8,400 deg^2$ of the sky, the *Sloan Extension for Galactic Understanding and Exploration* (SEGUE) that covered about $3,500 deg^2$ and the *Sloan Supernova Survey* which confirmed about 500 Type Ia SNe (Djorgovski et al., 2013). The surveys that comprise SDSS-III are APO Galactic Evolution Experiment (APOGEE) which surveyed over 100000 red giant stars across the full range of the galactic bulge, bar, disk and Halo, Baryon Oscillation Spectroscopic Survey (BOSS) that covers $10,000 deg^2$, Multi-object APO Radial Velocity Exoplanet Large-area Survey (Marvels) which monitored radial velocities of 11,000 bright stars, and SUGUE-2 which expanded the sky coverage to about $1,317 deg^2$. SDSS-IV (2014-2020) is the recently completed survey which covers about $14,555 deg^s$ area of the sky.
The surveys that comprise SDSS-IV are the Extended Baryon Oscillation Spectroscopic Survey (eBOSS), the APO Galactic Evolution Experiment 2 (APOGEE-2) and Mapping Nearby Galaxies at APO (MaNGA). eBOSS maps the distribution of galaxies and quasars from when the Universe was 3 billion years old, a critical time at which that expansion of the universe started due to dark energy. APOGEE-2 explores the history and evolution of the Galaxy. APOGEE-2 relies on the spectroscopic information of stars using the near-infrared light, which is not absorbed by the interstellar dust. MaNGA obtains spectra across the entire face of target galaxies using custom-designed fibre bundles. Figure 3 shows the eBOSS DR16 spectroscopic coverage in equatorial coordinates.

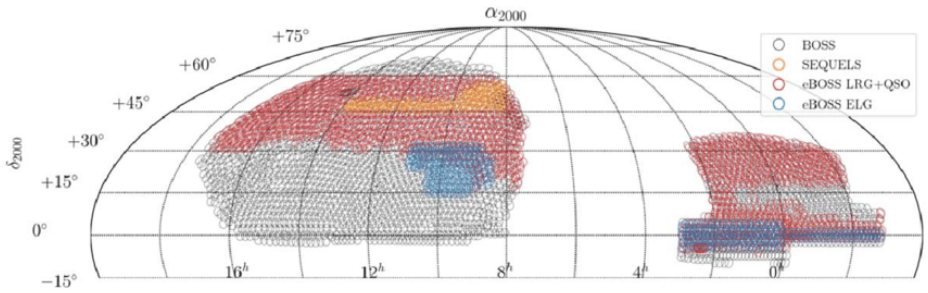


Figure 3: DR16 eBOSS spectroscopic scope in equatorial coordinates, the map is centered at $R.A. = 8h'r$. Every symbol in figure 3 indicates the location of a completed spectroscopic plate and it is scaled to approximate the field of view. The blue, red yellow and grey symbols represent the spectroscopic coverage of eBOSS ELG, SEQUELS, eBOSS LRG+QSO and BOSS respectively (Ahumada et al., 2020).

2.2 WISE

The Wide-field Infrared Survey Explorer (WISE) is a space satellite mission which has surveyed the sky in the mid-infrared starting in 2010. WISE launched on 14 December 2009, concluded its first mid-infrared survey of the entire sky on 17 July 2010 and concluded its cryogenic mission on 30 September 2010. After its hydrogen coolant depleted on 30 September 2010, the satellite continued observations in the W1 and W2 channels as part of the renamed NEOWISE mission until 1 February 2011, when the satellite was put in hybernation mode. WISE was much more sensitive than previous infrared survey missions and imaged the entire sky in four infrared passbands (W1, W2, w3 and W4) which are centred at 3.4, 4.6, 12, and 22 microns respectively. It used a 40 cm telescope with a

field-of-view of 47 arcminutes. The four WISE mid-infrared passbands have an angular resolution of 6.1",6.4", 6.5" and 12.0" respectively. During its cryogenic mission WISE achieved 5-sigma point source sensitivity per band better than 0.08, 0.11, 1 and 6 mJy in unconfused regions on the ecliptic. The WISE all-sky survey greatly improved depth upon the 2MASS all-sky surveys for astronomical sources that have spectra close to the spectra of an A0 star. It also went deeper for averagely red astronomical sources similar to K stars or galaxies which are populated by old stars. The numerous frames covering each part of the sky were stacked and produced by WISE as an image atlas in four colours. The WISE data products were modelled on the 2MASS image atlas and the 2MASS point source catalogue. WISE originally produced three data releases. The first (preliminary) data release on 14 April 2011 covered 55% of the sky to a depth of 20σ. The second (All-Sky) data release on 13 March 2012 covered the entire sky to a depth 5σ. The third (AllWISE) data release on 13 November 2013, which we use in this project, combined data from the WISE cryogenic mission and from the NEOWISE 4-month early mission. The satellite has since been reactivated on 7 December 2013 to continue observations in the W1 and W2 channels and has been in operations since then. Different teams have since worked on improving the imaging and the cataloguing of (NEO)WISE data, e.g. as part of the unWISE and the CatWISE efforts (Schlafly et al., 2020; Eisenhardt et al., 2020).

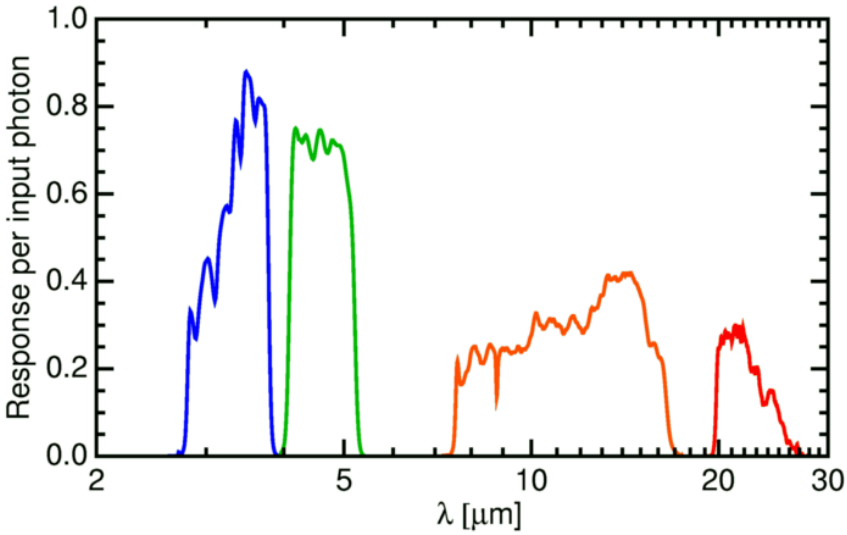


Figure 4: WISE response curves per photon

Figure 4 shows the response curves of WISE four infrared pass-bands W1, W2, W3 and W4. The response curves are constructed using the product of component data and the expected value from the design. The effective wavelengths of the four bands are 3.368, 4.618, 12.082 and 22.194 μm respectively for a signal spectrum with $\nu F\nu = const$. The effective wavelength is defined such that infinitesimal power law changes to the spectrum, pivoting about the effective wavelength, do not change the integrated response (Wright et al., 2010). During its cryogenic survey, WISE imaged the whole sky with multiple, independent exposures. The exposures of all four pass-bands were measured simultaneously. The exposure times were 7.7 seconds in W1 and W2 and the 8.8 seconds in W3 and W4. Figure 5 shows that the survey scanning strategy resulted in 12 to 13 exposures of each point on the ecliptic plane. The coverage increases to about thousands of exposures at the ecliptic poles. Ev-

ery individual exposure that met minimum requirements for image quality and noise levels were combined to form the All-Sky Release Image Atlas and Source Catalog. In Figure 5 there are localized decreases in the coverage in small areas, which took place because of the exclusion of lower quality exposures.

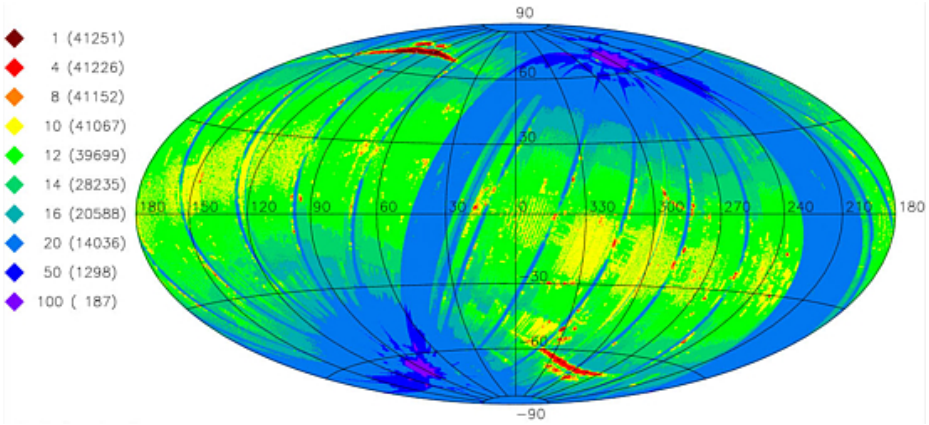


Figure 5: Equatorial Aitoff projection sky map showing the average number of individual 7.7 sec and 8.8-sec exposure frames of (W1 and W2) and (W3 and W4) respectively. Colours encode different frame depths-of-coverage, and the legend on the left gives the cumulative area in square degrees as a function of coverage depth. Most parts of the sky were imaged within 10 to 20 times but the areas closer to the ecliptic poles has most sky imaged at a much higher frequency.

2.3 SDSS-WISE data sample

The SDSS observed mostly the northern part of the sky while WISE covered the entire sky. Most astronomical sources that are detected by SDSS are also detected by the WISE. In this project, we matched the SDSS DR16 spectroscopic sample and the AllWISE Data Release in order to obtain a sample of jointly detected sources. This SDSS-WISE dataset therefore covers the optical and mid-infrared portion of the electromagnetic spectrum with multi-wavelength broad-band photometry in 5 optical bands and 2 mid-infrared bands. The 16th data release from the SDSS (DR16) is the latest data release in a series that began in 2001. It is the cumulative forth data release from SSDS-IV which means that all the previous data releases are included in DR16 (Ahumada et al., 2020). The AllWISE Data Release comprises the data that was taken from WISE cryogenic and NEOWISE post-cryogenic survey phases from 14 January 2010 up to 1 February 2011, to form the most comprehensive view of the full mid-infrared sky currently available. The SDSS-WISE datasets, $sdss\_all\_small\_dr16\_AllWISE\_red$, consists of 2,418,325 astronomical sources of which 1,683,116 sources are galaxies, 412,051 are quasars and 323,158 are stars. In our project, we use the SDSS optical photometry and spectroscopy, and the AllWISE mid-infrared photometry for all sources that are classified as extragalactic objects, i.e. galaxies or quasars, in the DR16 spectroscopic database. The main assumption here is that colours of astronomical sources of the same kind are on average the same or very similar for sources that are at the same redshifts. Thus we compute colours by taking the difference between magnitudes of the extragalactic sources in different filters. The colours associated with the objects in the SDSS-WISE datasets are therefore u-g, g-r, r-i, i-z, z-w1 and w1-w2. In this

study, we used SDSS-WISE extragalactic sources with a reliable spectroscopic measurement. The sample comprises 285,685 galaxies and 124,688 quasars. The spectroscopic redshifts of galaxies range up to about $z \approx 0.3$ while the redshifts of quasars range up tp about $z \approx 3.5$. The redshifts of galaxies and quasars were estimated independently using machine learning regression algorithms. For both galaxies and quasars, 20% of the sample was selected as a testing set while 80% was kept as the training set for the algorithms. About 57,137 galaxies were used for testing and 228,548 galaxies were used for training the algorithm. On the other hand, 24,938 quasars were used for testing while 99,750 quasars were used to train the algorithm. This is a form of 'blind' testing because none of the objects in the testing sample was used for training the regression algorithms. The six dimensions that correspond to the colours of the galaxies and quasars used are u-g, g-r, r-i, i-z, z-w1 and w1-w2. The colours were derived by the extinction corrected model magnitudes. The figure below shows the colour to colour diagrams for galaxies and quasars under study.

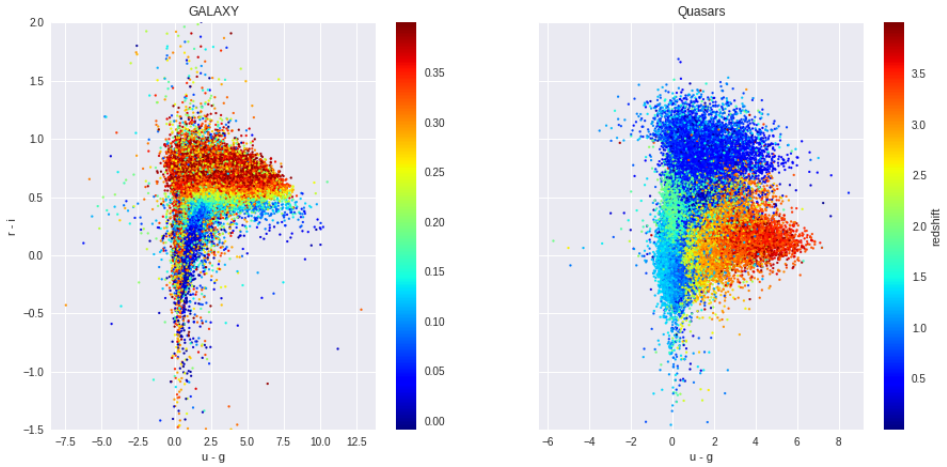


Figure 6: The colour-colour diagram showing the distribution of the observed galaxies and quasars with the colour bar of true spectroscopic redshifts. The left plot indicates the observed galaxy count distribution on the 2D grid of r-i vs u-g, and the right plot represents the distribution of observed quasars on the same r-i vs u-g grid.

Figure 6 shows an example of a colour-colour diagram which depicts the distribution of galaxies and quasars as a function of redshift. The left plot shows the observed galaxy count distribution on the 2D grid of r-i vs u-g, and the right plot represents the distribution of quasars on the same r-i vs u-g grid. The left plot of figure 6, shows that most of the galaxies are concentrated in the redshift range $0 \leq z \leq 0.3$. The galaxies at these redshifts also appear to be at the r-i band magnitude of -0.5 mag and extend to about 1.5 mag. There is a small set of galaxies that lies within the redshift range $0.25 < z < 0.35$. and there are a very few galaxies which lie at the redshift greater than 0.35. Galaxies that are observed usually have a trend that the galaxies that are having a similar magnitude in r-i and u-g bands appear to be having similar redshifts with very few galaxies overlapping. From this plot we can infer that most of the galaxies at lower redshift have the lower magnitude in r-i band. The observed galaxies have a maximum redshift of about 0.4. On the other hand, the right plot of figure 6 represents the distribution of observed quasars on the same r-i vs

u-g grid. In a case of the observed quasars shown in figure 6, most of the quasars are found within the redshift range $0 \leq z \leq 3.5$. There are very few quasars that are found at higher redshift than z = 3.5. The right plot indicates the three regions that are dominated by three different average redshifts. The plot depict a relationship between the redshifts and colours, and how some sources are overlapping in the colour-colour diagram which would mean that it will be a bit challenging to estimate the photometric redshifts of quasars compared to the galaxies. Because some galaxies or quasars with the same the redshifts may be found to have different colours or vice versa, we utilize the spectroscopic measurements to assist when differentiating the sources.

3 Machine Learning

Machine learning is the field of computer science dealing with algorithms that provide systems with the ability to automatically learn and improve from experience without being explicitly programmed (Danysz et al., 2019). Two common applications of machine learning in astronomy deal with classification and regression. Classification and regression are supervised machine learning techniques to learn from the data discrete labels (for classification) or continuous values (for regression) that already exist (training set) and then assign the labels to new "unseen" data (testing set). This can be achieved with techniques such as random forests, neural networks or decision trees (Borne, 2013). In this study, we are going to make use of two regression methods, Random Forest Regression (RFR) and K-Nearest Neighbour Regression (KNNR), to estimate the photometric redshifts of galaxies and quasars. The effectiveness of supervised methods depend on the availability of training sets for which the desired properties are known with confidence and on the training set being representative of the sample for which the properties of interest must be estimated. In otjer words, the parameter space covered by the input attributes must cover the same parameter space over which the algorithm is to be used to make predictions (Ball and Brunner, 2010).

3.1 Regression

Regression is one of the most important fields in statistics and machine learning and is a useful technique for estimating the photometric redshifts of astronomical sources. Regression aims to establish a relation between a set of independent variables x, and the dependent variable y, that describes the expected value of y given x. Regression techniques have to make several simplifying assumptions about the nature of the data, the uncertainties of the measurements, and the complexity of the models (Ivezić et al., 2014). In this project we adopted a couple of supervised machine learning techniques (Random Forest and K-Nearest Neighbour) to describe how the redshifts of galaxies and quasars depend on the broad-band colours and magnitude of the source. The two methods determine the relationship between the colours and the spectroscopic redshifts. Since Random Forest and K-Nearest Neighbour regression are both supervised machine learning tools, they rely solely on the training set at hand to map the relationship between the broad-band colours and the spectroscopic redshift to model the output redshifts given the colours as an input, the input values must be the properties that are known with confidence. Random Forest and K-Nearest Neighbour Regressors are very robust, simple to use and often the most accurate algorithms for low- as well as high-dimensional data.

3.1.1 K-Nearest Neighbour Regression

K-Nearest Neighbour (KNN) regression is one of the most simple regression algorithms in machine learning. KNN regression approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighbourhood. The size of the neighbourhood needs to be set by the analyst or can be chosen using cross-validation to select the size that minimises the mean-squared error. The algorithm depends on the user's ability to calculate the distance in any multivariate parameter space between database objects. The value of K in this regression is usually chosen to be the square root of the total number of objects in the training set, and it should be an odd number to avoid the tie. The algorithm is expensive because all the distances to the object are recalculated for each new item, and the list of distances must be sorted every time. KNN uses the test data to identify similar points within the training set while the parameter k is kept fixed (Borne, 2013). This is a powerful method because all the information available for each object is utilized without interpolation or approximation. KNN regression is then used in this study to map the relationship between the spectroscopic redshifts and the multi-wavelength photometric data of galaxies and quasars in the training set. The memorised relationship is then used to predict the most probable redshifts of the galaxies or quasars on the testing data set. This algorithm estimates the photometric redshift by considering the average of K-nearest neighbours with similar properties to the unknown object from the training dataset.

3.1.2 Random Forest Regression

Random Forest as mentioned above is one of the most powerful supervised machine learning algorithms used for classification and regression. It is very popular because it is easy to use and understand. Similarly, the algorithms work in two stages; first, the algorithm determines the map from the spectroscopic redshifts and the multi-dimensional colour space in the training set, and second, the algorithm uses the map to estimate the unknown redshifts from the test set. Random forest is a modification to the standard decision tree model-building and rule learning algorithm. Random forest compensates for the bias of decision tree towards the specific data on which they are trained, and also for the risk of overfitting by combining many decision trees into one model (Ivezić et al., 2014).

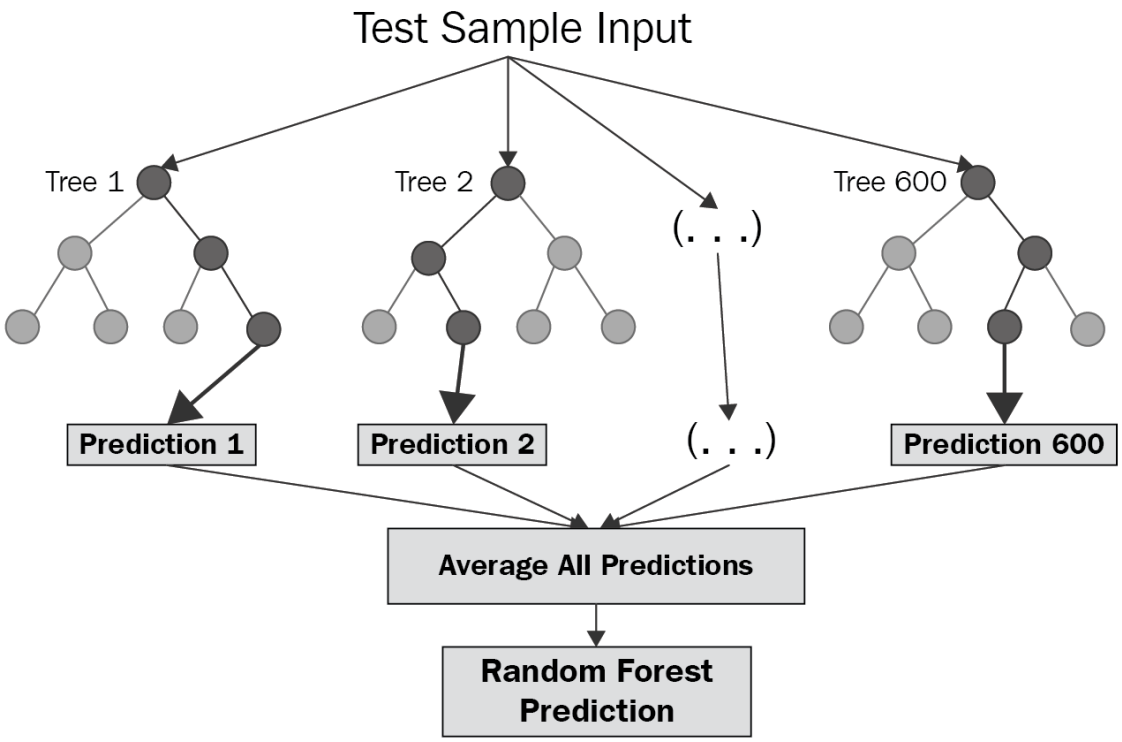


Figure 7: Random Forest Structure.

Random forest building algorithm considers the total number of different measured quantities, called attributes, to build the number of decision trees. Consider that, n, gives the number of trees in the forest and, m, is the number of attributes that are split at each level of the node of the tree. Each decision tree contains a sub-sample that is selected from the data. At the decision tree node, All the variables of attributes are mixed up randomly and then reassigned to the objects in the source list. A different set of m attributes are used for each node. Each one of $n - trees$ in the forest uses all of the attributes, but with a different one of the attributes randomized in each distinct tree.
We use random forest regression to estimate the photometric redshifts of galaxies and quasars.

4 Results

In this section, we provide the detailed results of our photometric redshift estimation for galaxies and quasars using the the K-Nearest Neighbour and Random Forest regressors. The photometric redshifts of galaxies and quasars were estimated separately, i.e. we split galaxies and quasars into two separate samples and ran the two algorithms on the two samples separately. This is an important point, as we implicitly assumed that sources have been correctly classified as galaxies or quasars before we embarked on their photometric redshift estimates. Classification into a finite set of classes is a separate problem in astronomy and in science in general, and while the machine learning algorithms we adopted are also useful to deal with classification, we did tackle this problem here but rather assumed that classification of our sample was correctly carried out beforehand. We divide this section into KNN regression and RF regression so that we can easily compare the performance of the

two algorithms at estimating the photometric redshifts of galaxies or quasars. For both algorithms, we calculated the biases, errors and outlier rates so we could understand to what degree we can trust our results. These uncertainty estimates play an important role in assessing the performance of the algorithms used. We determine the statistical error of the measurements and the fractions of outliers for each algorithm used. We begin by calculating the Δz_{norm} parameter, called the normalized redshift estimation error and defined as

$$\Delta z_{norm} = \frac{z_{phot} - z_{spec}}{1 + z_{spec}} \tag{3}$$

where z_{spec} is the spectroscopic redshift and z_{phot} is the photometric redshift of an astronomical source. Using Δz_{norm} we calculate the mean and the median as mean(Δz_{norm}) and median(Δz_{norm}) respectively. It is necessary to calculate both the mean and the median because the mean is usually skewed by large values of z or small values of z while the median can not be skewed.
There are two commonly used estimators to determine the accuracy of the photometric redshifts; the *Root Mean Square (RMS)* and the *Normalised Mean Absolute Deviation (NMAD)*. The RMS yields are the standard deviation of the residuals, thus RMS measures how these residuals are spread out, and NMAD is a robust alternative to RMS and it is not sensitive to outliers. The RMS and NMAD are calculated as follows:

$$RMS = std(\Delta z_{norm}) \tag{4}$$

where std stands for standard deviation,

$$NMAD = 1.4826 \times MAD(\Delta z_{norm}) \tag{5}$$

where MAD is given by

$$MAD = median(|\Delta z_{norm} - median(\Delta z_{norm})|) \tag{6}$$

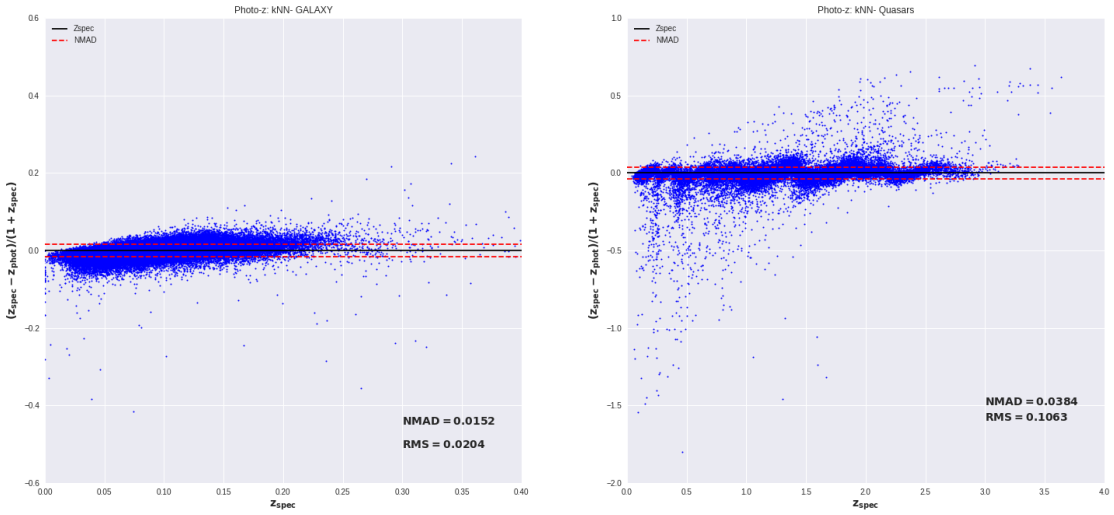
The outlier fractions are calculated in the form of percentages. Sigma1 ($\sigma 1$) is defined to be

$$1\sigma = n[|\Delta z_{norm}| > (1 \times RMS)]/N \times 100 \tag{7}$$

where $n(|\Delta z_{norm}| > (1 \times RMS))$ gives the total number of objects in the set Δz_{norm} that are greater than $1 \times RMS$ and N is the total number of objects in the set Δz_{norm}. Similarly 2σ, 3σ and 4σ were defined, where $4\sigma = \Delta z_{norm} > 0.15$.

4.1 K-Nearest Neighbour photometric redshifts

We use KNN regression to determine the photometric redshifts of our test samples, which is made up by 57, 137 galaxies and 24, 938 quasars respectively. To maximise the performance of the algorithm 80% of each sample of the galaxies and quasars was selected as a training set, to train the algorithm. Figure 8 below shows the plot of the Δz_{norm} *vs* z_{spec}.

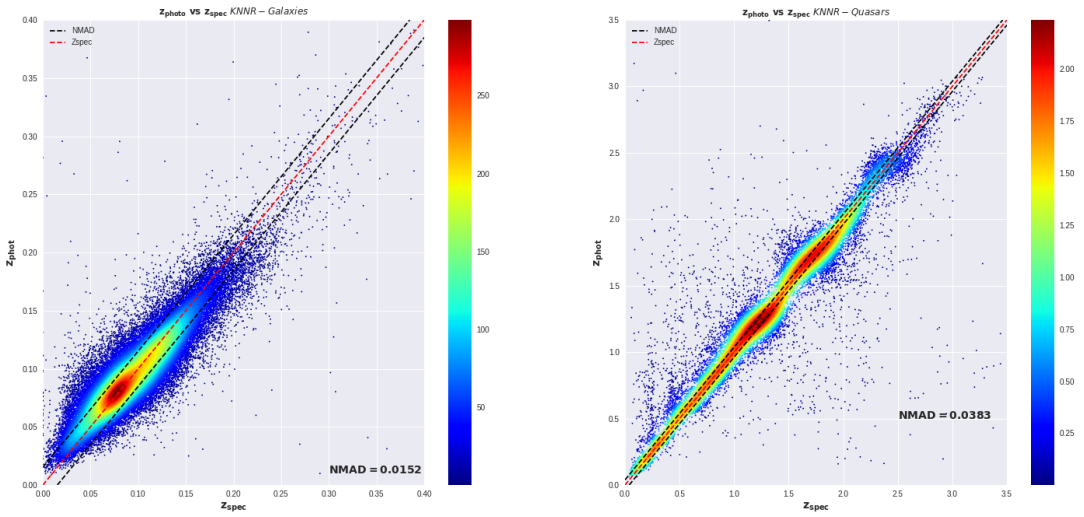


(a) Δz_{norm} *vs* z_{spec} galaxies.

(b) Δz_{norm} *vs* z_{spec} quasars.

Figure 8: normalized redshift estimation error as a function of spectroscopic redshift by KNN algorithm.

Figure 8 shows the normalised redshift estimation error, Δz_{norm} defined above, against the spectroscopic redshift of galaxies. The left plot represents Δz_{norm} for galaxies calculated from the KNN regression while the right plot shows the Δz_{norm} for quasars calculated from the same KNN algorithm. The black solid line represents the spectroscopic redshifts for which the normalised redshift error is equal to zero while the red dashed line shows the spectroscopic redshifts for which the normalised redshift error is equal to plus or minus the Normalised Mean Absolute Deviation (NMAD) deviation. Using the Δz_{norm} K-Nearest Neighbour regression achieves the NMAD and Root Mean Square (RMS) of 0.0152 and 0.0202 for galaxies respectively. On the other hand, the same algorithm achieves the NMAD and RMS of 0.0384 and 0.1063 respectively for quasars. In both figure 8 plots About 50% of the points in the plots are concentrated within the NMAD range, ie $-NMAD \leq \Delta z_{norm} \leq +NMAD$. The RMS and NMAD are fairly close for galaxies which makes sense because few outliers are seen in figure 8a. On Figure 8, the RMS is greater and it also evident from the plot that there a lot of residuals. As mentioned above the NMAD is not sensitive to the outliers which is why it remains a lower value even when there are outliers observed.
We now make a plot of photometric redshifts as a function of the spectroscopic redshifts. The figure below shows the density plot of z_{phot} *vs* z_{spec} for the galaxies and quasar. The photometric redshifts for both astronomical sources are estimated by the KNN regressor.



(a) z_{phot} *vs* z_{spec} galaxies.

(b) z_{phot} *vs* z_{spec} quasars.

Figure 9: The predicted redshifts as a function of the spectroscopic redshift by the K-Nearest Neighbour algorithm. The left plot represents the z_{phot} *vs* z_{spec} for galaxies and the right plot shows the z_{phot} *vs* z_{spec} for quasars. The colour bar shows the number of galaxies that correspond to a colour on the plot. The density plots become redder where there is a larger concentration of objects and bluer where there is less concentration of objects. The solid black straight line represents a function of photometric redshifts against the spectroscopic redshifts for which the photometric redshift is equal to the spectroscopic redshift. The dashed red straight line in this case shows represents $\Delta z = \pm NMAD$.

Figure 9a shows the plot of galaxies photometric redshifts versus the corresponding spectroscopic redshifts and figure 9b shows the plot of quasars photometric redshift against the spectroscopic redshift. Both galaxies and quasars photometric redshifts are estimated by the KNN algorithm. Figure 9b shows that most of the galaxies extend to the redshift of about $z = 0.25$ with a lot of galaxies at a redshift of about $z \approx 0.8$. The plot also shows that a vast number of galaxies found below and above the z_{spec} line shows a good match with the few outliers observed as the redshift increases. The match between most source above and below the z_{spec} line indicates that the KNN algorithm has made a food estimate of the photometric redshifts of galaxies. In figure 9b, quasars extend redshifts to about $z \approx 3.5$ with some outliers observed as the redshift increases. The RMS value in this case is bigger compared to that of galaxies. The photometric redshifts and spectroscopic redshifts agree and this is evident because most of the sources lie within the $\Delta z = \pm NMAD$, represented by the nice red colour increasing with the redshifts in figure 9b.
Table 1 below summarises the statistical calculations for the KNN regression.

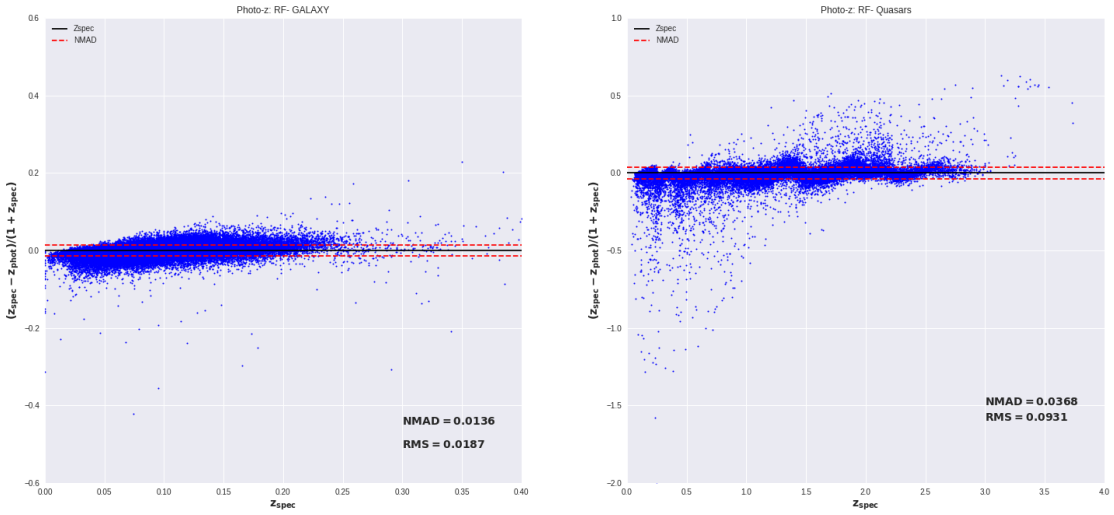
Table 1: KNN regression statistics for galaxies and quasars

Class	Mean	Median	RMS	NMAD	$1\sigma(\%)$	$2\sigma(\%)$	$3\sigma(\%)$	$\Delta z_{norm} > 0.15(\%)$
Galaxies	0.00021	3.83025	0.02042	0.01519	21.39944	3.43560	0.73857	0.11026
Quasars	0.00818	0.00073	0.10626	0.03842	7.77929	3.32023	2.13730	5.50040

Table 1 shows the summarized KNN results for both galaxies and quasars. All the calculations in Table 1 are estimated from the Normalized redshift error between the KNN photometric redshifts and the spectroscopic redshifts, thus the mean, median, RMS, NMAD, etc represents the mean(Δz_{norm}), median(Δz_{norm}), RMS (Δz_{norm}), NMAD (Δz_{norm}), etc respectively. From the KNN algorithm for galaxies, the calculated mean bias, median bias, and RMS, NMAD, 1σ, 2σ, 3σ and $\Delta z_{norm} > 0.15$ errors are 0.00021, 3.83025, 0.02042, 0.01519, 21.39944, 3.43560, 0.73857 and 0.11026 respectively. For the quasars, the same parameters were found to be 0.00818, 0.00073, 0.10626, 0.03842, 7.77929, 3.32023, 2.13730 and 5.5.00040 respectively. In Table 1 the RMS for galaxies compared to the one for quasars and also from FigureZ9a, the galaxies have few outliers compared to the outliers found in quasars shown in Figure 9b.
The RMS (Δz_{norm}) and NMAD (Δz_{norm}) are fairly small for both galaxies and quasars this means that the there are fewer uncertainties in the photometric redshift estimation. The biases and higher errors in the redshift estimation are strongly dependent on the position of a given galaxy in the space of broad-band colours. The observed larger uncertainties in the quasars may be as a result that at higher redshifts it is difficult to distinguish the quasars as shown in figure 6 and also due to the spectroscopic redshift limitation of the training dataset. Thus we have the KNN giving best estimates of photometric redshifts for the galaxies than quasars.

4.2 Random forest photometric redshifts

In a similar way to the KNN algorithm above, the random forest algorithm is applied to determine the photometric redshifts of galaxies and quasars. Figure 10 below shows the plot of normalized redshift estimation error as a function of redshifts for the galaxies (left) and the quasar (right). To determine the map between the redshifts and the multidimensional colour-colour space, as subsample that contains about 80% of the galaxies and quasars was used as training sets for galaxies and quasars respectively.



(a) Δz_{norm} *vs* z_{spec} galaxies.

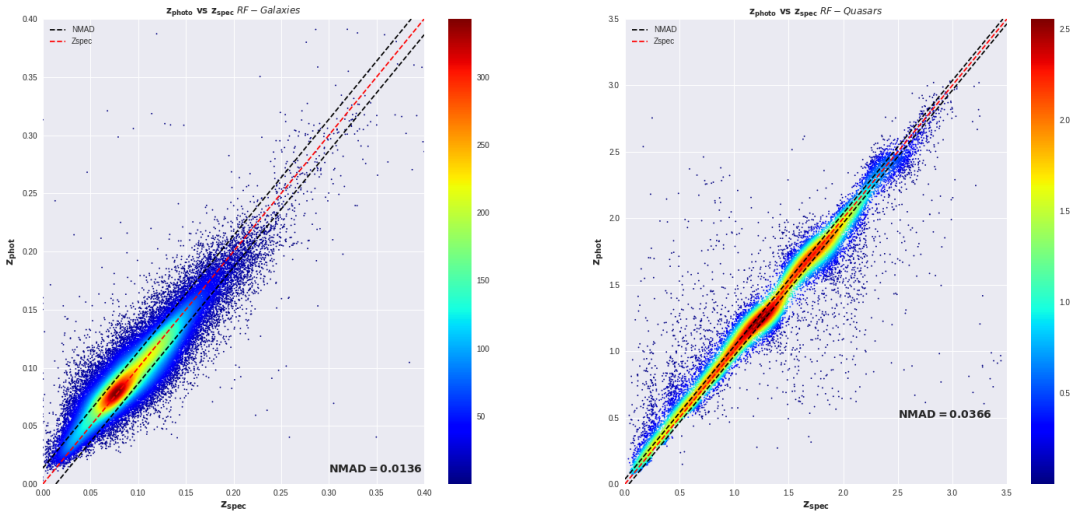
(b) Δz_{norm} *vs* z_{spec} quasars.

Figure 10: normalized redshift estimation error as a function of spectroscopic redshift by RF algorithm.

Figure 10 represents the normalized redshift estimation error (Δz_{norm}), calculated from the photometric redshifts estimated by the random forest algorithm, as a function of spectroscopic redshifts. The black solid line represents the spectroscopic redshifts for which the normalised redshift error is equal to zero while the red dashed line shows the spectroscopic redshifts for which the normalised redshift error is equal to plus or minus the Normalised Mean Absolute Deviation (NMAD). Figure 10a depicts the Δz_{norm} versus the spectroscopic redshifts for the galaxies. The RMS and NMAD of the galaxies were found to be 0.0187 and 0.0136 by the RF algorithm. The outliers in figure 10a dominate at higher spectroscopic redshifts thus the galaxies photometric redshifts are well estimated at the lower redshift. The small amount of the outliers that are observed in figure 10a are accounted for by having a small RMS value which is closer to the NMAD. Most of the galaxies are concentrated to a redshift of about $z = 0.25$ and about 50% of the points in the plots are concentrated within the NMAD range, ie $-NMAD \leq \Delta z_{norm} \leq +NMAD$. Figure 10b shows the Δz_{norm} against the corresponding spectroscopic redshifts for the quasars. The RF algorithm for quasars achieves the RMS and NMAD of 0.0931 and 0.0368 respectively. These two values are higher compared to the ones that were found by the algorithm when estimating the photometric redshifts of the galaxies. The RMS of quasars should be higher because the outliers extend to $\Delta z_{norm} \approx -1.7$ at the lower quasars redshifts and the NMAD may be higher. After all, the algorithm has a difficulty to separate the sources at higher redshifts. Figure 10b shows that the redshift of the quasar extends to about $z = 3.5$. Most of the quasars outliers lie below the -NMAD line for $0 \leq z \leq 1.5$ and some outliers are found above the +NMAD line from $z = 1.5$ (for z the spectroscopic redshift). This leads to the conclusions that the RF algorithm photometric redshifts are not good at the lower the redshift.

The photometric redshifts and the spectroscopic redshifts are best compared to one another through a plot of photometric redshifts as a function of the corresponding spectroscopic

redshift. The photometric redshift against spectroscopic redshifts shows a nice trend that increases as both photometric and spectroscopic redshifts increases. Figure 11 shows the comparison between the photometric redshifts, estimated by the RF algorithm, and the spectroscopic redshift of galaxies and quasars.



(a) z_{phot} *vs* z_{spec} galaxies.

(b) z_{phot} *vs* z_{spec} quasars.

Figure 11: The predicted redshifts as a function of the spectroscopic redshift by the random forest algorithm. The left plot represents the comparison between the photometric redshifts and spectroscopic redshifts for galaxies and the right plot shows the same comparison for quasars. The colour bar shows the number of galaxies that correspond to a colour on the plot thus the plots become redder where there is a larger concentration of objects and bluer where there is less concentration of objects. The solid black straight line represents a function of photometric redshifts against the spectroscopic redshifts for which the photometric redshift is equal to the spectroscopic redshift. The dashed red straight line in this case shows represents $\Delta z = \pm NMAD$.

Figure 11a shows the comparison between the galaxies redshifts from the RF algorithm with the spectroscopic redshifts. Similarly most galaxies lies within $z = 0.25$. The residuals increases as the redshift increases and most are found at higher redshifts. The plot indicates that most galaxies are located at the redshift (z) equal to 0.8. The plot also shows that a vast number of galaxies found below and above the z_{spec} line shows a good match. The match between most sources photometric redshifts and spectroscopic redshifts above and below the z_{spec} line indicates that the RF algorithm has made a food estimate of the photometric redshifts of galaxies. Figure 11b shows the comparison between the photometric redshifts and the spectroscopic redshifts for the quasars. As explained above there are a handful of residuals that are accounted for by the larger RMS value. This means that the random forest algorithm makes good photometric approximation on the galaxies than quasars. This result was also found by the KNN algorithm. The photometric redshifts and spectroscopic redshifts of agree and this is shown by most of the sources lying within the $\Delta z = \pm NMAD$, represented by the nice red colour increasing with the redshifts in figure

11b.
Table 2 below summarises the statistical calculations for the RF regression.

Table 2: RF regression statistics for galaxies and quasars

Class	Mean	Median	RMS	NMAD	$1\sigma(\%)$	$2\sigma(\%)$	$3\sigma(\%)$	$\Delta z_{norm} > 0.15(\%)$
Galaxies	0.00049	0.00043	0.01874	0.01358	21.49220	3.54586	0.76308	0.08751
Quasars	0.00973	0.00361	0.09313	0.03683	9.25094	3.79742	2.02502	5.01243

Table 2 shows that the galaxies and quasars have a mean error of 0.00047 and 0.01175 respectively. The median error was found to be 0.00049 for galaxies and 0.00973 for the quasars. The RMS and NMAD were found to be 0.0187 and 0.0136 for galaxies and were found to be 0.0931 and 0.0368 for the quasars. The outlier fractions for galaxies were also calculated; 1σ, 2σ, 3σ and $\Delta z_{norm} > 0.15$ were found to be 21.49220%, 3.54586%, 0.76308%, and 0.08751% respectively. For the quasars the outlier fractions (1σ, 2σ, 3σ and $\Delta z_{norm} > 0.15$) are 9.25094%, 3.79742%, 2.02502% and 5.01243% respectively. All the calculations in Table 2 are also estimated from the Normalized redshift estimation error between the RF photometric redshifts and the spectroscopic redshifts, thus the mean, median, RMS, NMAD, etc represents the mean(Δz_{norm}), median(Δz_{norm}), RMS (Δz_{norm}), NMAD (Δz_{norm}), etc respectively.

5 Conclusions

Using the K-nearest neighbour and random forest regression supervised machine learning algorithms, we estimated the photometric redshifts of galaxies and quasars jointly found in the SDSS DR16 spectroscopic sample and in the AllWISE Data Release. The two algorithms provided accurate photometric redshift estimates by considering the photometric measurements of the galaxies and quasars in seven optical to mid-infrared filters. We thus demonstrated that simple machine learning algorithms can handle a large amount of data and are very efficient when estimating photometric redshifts. The two algorithms have comparable performance and both produce excellent photometric redshifts. Both algorithms achieve a better RMS and NMAD for galaxies while giving larger uncertainties when calculating photometric redshifts for quasars. The results show that the random forest algorithm performs slightly better than the K-nearest neighbour algorithm. The RMS and NMAD of the RF algorithm are slightly smaller than those of the KNN algorithm for both galaxies and quasars. This means that the random forest algorithm made better estimates of the photometric redshifts of galaxies and quasars under study. The larger uncertainties derived by both algorithms for the photometric redshift estimates of quasars can be attributed to a combination of factors, such as: quasar spectra are 'flatter', and therefore broad-band photometry is less effective in contraining the redshifting of their main spectral features; quasars are found over a large range of redshifts than galaxies for a fixed sensitivity limit, and thus there is greater colour degeneracy; quasars are often highly variable, and therefore multi-wavelength observations taken at different times are less effective when used together. These difficulties can partly be overcome by improved data taking and pre-processing as well as by more sophisticated machine learning algorithms. For the photometric redshifts of galaxies, the RF algorithm obtained NMAD and RMS of 0.0136 and 0.0187 respectively

while the KNN algorithm achieved NMAD and RMS of 0.0152 and 0.0204 respectively. The NMAD and RMS are fairly close for galaxies though the random forest algorithm achieved the smaller values of those. The quasars NMAD and RMS were found to be 0.0368 and 0.0931 respectively by the RF algorithm and were found to be 0.0384 and 0.1063 by the KNN algorithm. Our photometric redshift estimates are reasonably accurate across our full redshift range, though the accuracy drops at higher redshifts. The random forest achieved the acceptable 3σ outlier rate of $P_0 = 0.763\%$ and $P_0 = 2.025\%$ for galaxies and quasars respectively, and the KNN algorithm obtained the 3σ outlier rate of $P_0 = 0.739\%$ and $P_0 = 2.137\%$ for galaxies and quasars respectively. Machine learning algorithms have proven very effective and efficient in estimating the photometric redshifts of extragalactic sources. The advances in computing technology and digital detectors has led astronomy to become a data-rich science. The data avalanche continues as large sky surveys such as the Large Synoptic Sky Survey (LSST) and the Square Kilometer Array (SKA) surveys will get going in the 2020s. These surveys will require low-cost and accurate machine learning algorithms to be able to determine the photometric redshifts and other properties for billions of astronomical sources.

6 Acknowledgements

I would like to thank my supervisor, Prof. Mattia Vaccari for his support throughout the project. I am also grateful for the effort that has been put forward by Chaka Mofokeng in helping with the code write up. I also appreciate the support from my colleague Yaaseen Jones, friends and family. The work uses the SDSS DR16 data which is the fourth data release of the fourth phase of Sloan Digital Sky Survey and AllWISE data products from the Wide-field Infrared Survey Explorer. Without the two surveys this work would have been impossible.

7 Source Code

The python codes used for this project can be accessed on the GitHub link attached below.
Username: pfunzowalter
Acount holder: Walter Silima
GitHub Account: https://github.com/pfunzowalter/REGRESSION

References

Ahumada, R., Prieto, C. A., Almeida, A., Anders, F., Anderson, S. F., Andrews, B. H., Anguiano, B., Arcodia, R., Armengaud, E., Aubert, M., et al. (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. *The Astrophysical Journal Supplement Series*, 249(1):3. 6, 8

Bahcall, N. A. (2015). Hubble's law and the expanding universe. *Proceedings of the National Academy of Sciences*, 112(11):3173–3175. 1, 2

Baldry, I. K. (2018). Reinventing the slide rule for redshifts: the case for logarithmic wavelength shift. 1

Ball, N. M. and Brunner, R. J. (2010). Data mining and machine learning in astronomy. *International Journal of Modern Physics D*, 19(07):1049–1106. 10

Beck, R., Dobos, L., Budavári, T., Szalay, A. S., and Csabai, I. (2016a). Photometric redshifts for the sdss data release 12. *Monthly Notices of the Royal Astronomical Society*, 460(2):1371–1381. 3

Beck, R., Dobos, L., Budavári, T., Szalay, A. S., and Csabai, I. (2016b). Photometric redshifts for the sdss data release 12. *Monthly Notices of the Royal Astronomical Society*, 460(2):1371–1381. 2

Borne, K. (2013). *Virtual Observatories, Data Mining, and Astroinformatics*, pages 403–443. Springer Netherlands, Dordrecht. 10, 11

Carroll, B. W. and Ostlie, D. A. (2006). *An introduction to modern astrophysics and cosmology*. 3, 4

Chong, De Wei, K. and Yang, A. (2019). Photometric Redshift Analysis using Supervised Learning Algorithms and Deep Learning. In *European Physical Journal Web of Conferences*, volume 206 of *European Physical Journal Web of Conferences*, page 09006. 2

Chromey, F. R. (2016). *To measure the sky: an introduction to observational astronomy*. Cambridge University Press. 4

Cunha, C. E., Huterer, D., Lin, H., Busha, M. T., and Wechsler, R. H. (2014). Spectroscopic failures in photometric redshift calibration: cosmological biases and survey requirements. *Monthly Notices of the Royal Astronomical Society*, 444(1):129–146. 2

Danysz, K., Cicirello, S., Mingle, E., Assuncao, B., Tetarenko, N., Mockute, R., Abatemarco, D., Widdowson, M., and Desai, S. (2019). Artificial intelligence and the future of the drug safety professional. *Drug safety*, 42(4):491–497. 10

Djorgovski, S. G., Mahabal, A., Drake, A., Graham, M., and Donalek, C. (2013). Sky surveys. *Planets, Stars and Stellar Systems*, page 223–281. 5, 6

Eisenhardt, P. R. M., Marocco, F., Fowler, J. W., Kirkpatrick, J. D., Meisner, A. M., Garcia, N., Schlafly, E. F., Stanford, S. A., Caselden, D., Cushing, M. C., Cutri, R. M., Faherty, J. K., Gelino, C. R., Gonzalez, A. H., Jarrett, T. H., Koontz, R., Mainzer, A., Marchese, E. J., Mobasher, B., Schlegel, D. J., Stern, D., Teplitz, H. I., Wright E. L., and CatWISE Team (2020). VizieR Online Data Catalog: The CatWISE2020 catalog (Eisenhardt+, 2020). *VizieR Online Data Catalog*, page II/365. 7

Han, B., Qiao, L.-N., Chen, J.-L., Zhang, X.-D., Zhang, Y., and Zhao, Y. (2020). GeneticKNN: A Weighted KNN Approach Supported by Genetic Algorithm for Photometric Redshift Estimation of Quasars. *arXiv e-prints*, page arXiv:2009.08608. 2

Ivezić, Ž., Connolly, A. J., VanderPlas, J. T., and Gray, A. (2014). *Statistics, data mining, and machine learning in astronomy: a practical Python guide for the analysis of survey data*, volume 1. Princeton University Press. 10, 11

Karttunen, H., Kröger, P., Oja, H., Poutanen, M., and Donner, K. J. (2016). *Fundamental astronomy*. Springer. 3

Kormendy, J. and Bender, R. (2012). A Revised Parallel-sequence Morphological Classification of Galaxies: Structure and Formation of S0 and Spheroidal Galaxies. *ApJS*, 198(1):2. 4

Mo, H., Van den Bosch, F., and White, S. (2010). *Galaxy formation and evolution*. Cambridge University Press. 4

Salvato, M., Ilbert, O., and Hoyle, B. (2019). The many flavours of photometric redshifts. *Nature Astronomy*, 3(3):212–222. 1

Schlafly, E. F., Meisner, A. M., and Green, G. M. (2020). VizieR Online Data Catalog: The band-merged unWISE Catalog (Schlafly+, 2019). *VizieR Online Data Catalog*, page II/363. 7

Sparke, L. S. and Gallagher III, J. S. (2007). *Galaxies in the universe: an introduction*. Cambridge University Press. 1, 3

Stoughton, C., Lupton, R. H., Bernardi, M., Blanton, M. R., Burles, S., Castander, F. J., Connolly, A., Eisenstein, D. J., Frieman, J. A., Hennessy, G., et al. (2002). Sloan digital sky survey: early data release. *The Astronomical Journal*, 123(1):485. 5

Tarrío, P. and Zarattini, S. (2020). Photometric redshifts for the Pan-STARRS1 survey. *A&A*, 642:A102. 2

Wright, E. L., Eisenhardt, P. R., Mainzer, A. K., Ressler, M. E., Cutri, R. M., Jarrett, T., Kirkpatrick, J. D., Padgett, D., McMillan, R. S., Skrutskie, M., et al. (2010). The wide-field infrared survey explorer (wise): mission description and initial on-orbit performance. *The Astronomical Journal*, 140(6):1868. 7

York, D. G., Adelman, J., Anderson Jr, J. E., Anderson, S. F., Annis, J., Bahcall, N. A., Bakken, J., Barkhouser, R., Bastian, S., Berman, E., et al. (2000). The sloan digital sky survey: Technical summary. *The Astronomical Journal*, 120(3):1579. 5