Università degli Studi di Napoli "Federico II"

Scuola Politecnica e delle Scienze di Base

Area Didattica di Scienze Matematiche Fisiche e Naturali

Dipartimento di Fisica

Laurea triennale in Fisica

Characterization of Outliers in the Quasar Photometric Redshift Space

Relatori: Prof. Giuseppe Longo
Dott. Stefano Cavuoti
Dott. Massimo Brescia

Candidato: Michele Delli Veneri
Matricola: N85000365

A.A. 2015/2016

Introduction

One of the most important problem in observational cosmology is to build large samples of galaxies with well measured physical parameters. In order to achieve such goal it is crucial to obtain the estimate of their distances: a task which is usually performed using the well known correlation between distance and recession velocity discovered by Edwin Hubble in 1929 also known as Hubble's law:

$$d = H_0 \times z$$

where z is the so called redshift. As we shall see in what follows this crucial parameter can be measured by exploiting the Doppler effect induced by the recession velocity on both the position of the spectral lines (spectroscopic redshifts) or on the shape of the continuum (photometric redshifts). This work focuses on the derivation of photometric redshifts for a large sample of quasars using a machine learning method and, in particular, tries to better characterize the so called "catastrophic outliers", i.e. the objects for which the method fails in making a correct prediction. The Machine learning method we used is the so called "Multi Layer Perceptron with Quasi Newton Algorithm" or MLPQNA which will be detailed in what follows. To train and test our network, we used a multi-band quasars data-set built by cross-matching four different astronomical surveys.

This paper is organized as it follows. In the first chapter we present a brief introduction of the astronomical objects for which we try to measure the photometric redshift. In the second chapter we summarize the various techniques actually used to measure redshifts of astronomical object and we explain why using Neural Networks at this stage of the exploration of the universe is almost mandatory. In the third chapter, we give a small introduction to Neural Networks and then explain the structure of the neural model used in this work and its working principles. In the fourth chapter

we describe the data-set creation process, the actual setup of the model for this particular problem and the results we obtained. In the fifth chapter we analyse the output and try to characterize the outliers. In the conclusions we discuss some hypothesis that couldn't be proved in this work and could pave the road for future experiments and we summarize all the discoveries made in this work.

Contents

Chapter 1

Active Galactic Nuclei

In general the term "Active Galactic Nucleus", or AGN, refers to the existence of energetic phenomena which take place in the nuclei, or the central regions, of galaxies and cannot be attributed clearly and directly to stars. The largest subclasses of AGNs are Seyfert galaxies and quasars, and the distinction between them is, to some degree, a matter of semantics. The fundamental difference between these two subclasses is in the amount of radiation emitted by the compact central source. Some of the other differences are due more to the *way* we observe them than the to concrete differences between the various types.

From a physical point of view, AGNs are extremely massive and dense objects (almost certainly blach holes) situated in the central region of galaxies that, through accretion processes, release enormous quantities of energy in the radio, optic, X, γ wavelengths and through cosmic rays.
We will see that AGNs have luminosities hundred or thousand times higher than those of "normal" galaxies and have spectra that differ from those of "normal" galaxies for the intensity of the emitted light in the different bands and for the presence of emission lines with widths of up to $10^4 kms^{-1}$ that imply the presence of high-velocity moving gas.
The highly energetic phenomena render AGNs very bright sources especially in the X band and the radio band.

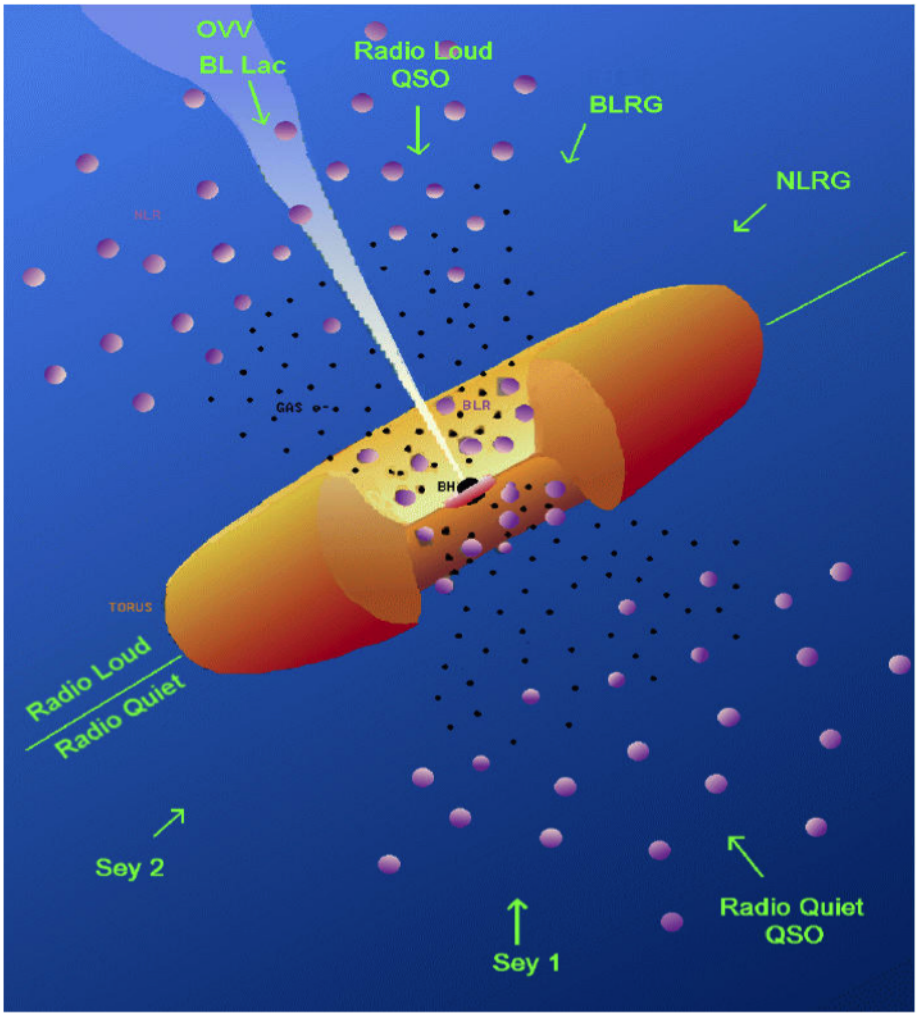

Figure 1.1: This illustration shows the geometric dependency of the unified AGN model. The broad-line (BLRG) and narrow-line (NLRG) regions are shown, as well as the "obscuring torus". A number of other AGN types are named as well. From Urry & Padovani 1995, [21]

1.1 The Unified Model

Since the discovery of AGNs (Seyfert 1943, [20]) our knowledge of their phenomenology has become much deeper and more complete. For almost fifty years after the discovery, it was believed that the differences in the phenomenologies of AGNs corresponded to the existence of several families of objects with different physical properties. This led to the development of many theories and models that were, most of the times, in conflict with one another.

The breakthrough arrived in 1993 with the work of Antonucci [1], revisited later in 1995 by Urry and Padovani [21], in which it was proposed that all these categories with all their different phenomenologies were in reality a single type of object viewed from different orientation: the so called "unified model", which can be summarized as it follows:

- In the center of the active galactic nuclei there is a massive black hole with mass between 10^6 and 10^{10} solar masses surrounded by an infalling matter disk (for the most part made of gas and dusts) of toroidal shape called the accretion disk. The matter present in the accretion disk is captured by the supermassive black hole (SMBL) and converted in electromagnetic energy with an extremely high efficiency ($\leq 10\%$). Furthermore it can be shown that, if the disk is rotating around its axis, it can accelerate the gas present in the accretion disk and expels it through two jets perpendicular to the rotating axis.

- At a distance of $\sim$ 100 light years from the center of the singularity there is an optically thick dust torus that, if oriented toward the observer line of sight, obscures most of the matter present in the inner regions of the AGN. In fact in the latter case the spectrum of the AGN is made of the emission lines of the torus (mainly in the infrared) and the absorption and emission lines of the outer regions of the galaxy.

- Inside the torus there is a a fast rotating medium made of clouds of gas and dust, called the BLR or "Broad Lines Region", responsible for the emission of the broad component of the permitted lines (these broad lines have amplitudes up to several thousands of Km/s) it is considered that the BLR may be caused by the photo-ionization of the

extremely hot accretion disk around the supermassive black hole. It is thought that the BLR is located at a distance from the nucleus $r \sim 1pc$.

- Gas clouds located at a distance $r \sim 10 - 1000pc$ from the black hole, form instead the "Narrow Lines Region" (or NLR), which is responsible for the emission of the forbidden lines and the narrow component of the permitted lines of the spectrum (their amplitudes are not larger than $1000km/s$).

To summarize, each component of the AGN is responsible for a particular type of emission: the relativistic jets are responsible for the gamma and radio emissions; the NLR and BLR are responsible for most of the optical emission; the accretion disk is responsible for the UV emission and partially of the optical one; the dust torus produces most of the IR emission.

1.1.1 Taxonomy of Active Galaxy Nuclei

From an observational point of view, the unified model leads to three main phenomenologies:

1. **Line of sight almost perpendicular to the torus plane**.
   The central engine is oriented face-on with one of the jets pointed towards the observer. The jets matter moves at relativistic speeds and so the emitted radiation is extremely collimated and can vary over a very short time span. In some cases the luminosity of the nucleus can surpass the luminosity of the host galaxy and the object is called Quasar (QSO) or quasi stellar object because at "first glance" it appears as an isolated star and not as an extended object like a galaxy (spatial resolution less than $7''$).
   Quasars were initially discovered due to their high radio emission and only later their optical counterparts were properly characterised. Quasars are usually divided in two categories RLQ (radio loud quasars) and RQQ (radio quiet quasars) depending on the presence or absence of a strong radio emission.
   Inside the RQQ class 5% - 10% of the objects show very high absorptions in the blue band of the resonant emission lines (Broad Absorption Lines); this phenomenon is due to the material expelled from the nucleus in the jet aimed toward the observer. Quasars that have flat

spectra (flat because they are saturated by the jet emissions) apart from the radio variability, tend to vary on even shorter time spans and present variability also in the optical emission.
These properties are common in Optical Violent Variable too, also called blazars, in which a high and variable degree of polarization is found in the optical continuum both in intensity and direction. The blazar class include both objects with the usual emission lines and objects with a continuum spectrum. The latter are called BL Lac after the prototype Bl Lacertae.

2. **Line of sight almost parallel to the torus plane**
When the nucleus is only partially obscured and its light is intercepted only partially by the observer line of sight, it is possible to see directly the inner regions of the AGN and in the spectra are present both the broad line and narrow line components as well as the accretion disk emissions. In this case the objects are "classified" as type I Seyfert galaxies, BLR and type I Quasars.

3. **Line of sight in an intermediate position (Type II Seyfert Galaxies)**
If the object are observed edge-on, the central black hole, the accretion disk and the broad lines region are obscured by the dust torus and so the only component visible is the Dust torus and its infrared emission spectrum. In the spectra, therefore, will be present only the narrow component of the permitted lines.

AGNs and quasars are quite common. Actually, it is commonly believed that all massive galaxies host in their center a massive black hole which, in particular phases of accretion becomes luminous originating the AGN phenomenology.

Chapter 2

Photometric Redshifts

As mentioned earlier, a way to measure the distance of an extragalactic astronomical source, is to measure its redshift; this is the shift of the source's spectral lines due to the expansion of the universe. It is possible, in fact, to obtain the distance of the source, knowing its redshift, from the solution of the Friedman equation.
Historically, redshifts have been measured with spectroscopy and several spectroscopic surveys have been done in the past and some are still active nowadays.
Spectroscopic surveys, however, are very much time consuming (in terms of precious telescope observing time and data reduction) and cannot match the requirements of modern precision cosmology which is based on samples of many millions of galaxies.
A viable alternative is provided by "Photometric redshifts" (or photo-z's) which are based on photometry rather than spectroscopy. At the price of lower accuracy, photo-z's offer several advantages over their spectroscopic counterparts:

1. being derived from intermediate/broadband imaging, photo-z's are much more effective in terms of observing time. That's because spectrographs diffract light into narrow wavelength bins, thus longer exposition times are required to achieve an acceptable signal to noise ratio;

2. depending on the particular method used, they allow to probe objects much fainter than the spectroscopic flux limit;

3. they allow, under specific conditions, to correct some biases, like the

ones encountered at high redshifts, where spectroscopy is pushed to its limit by the low signal to noise ratio in the spectra and by the fact that, even if a good signal to noise ratio is achieved, the lack of features in the observed spectral range make the estimation of the redshifts not trustworthy.

All these aspects render them ideal to produce large samples of candidate quasars. In fact, quasar samples are mostly constructed via a two-step process in which the first step is to identify quasar candidates through color[1] selection from multi-wavelength surveys and then, in the second step, to validate these candidates via a spectroscopic follow up. In practice, due to the large amount of observing time required by spectroscopy and the huge number of these candidates that modern survey supply, this method has become unusable.
Thankfully, it has been demonstrated in the last few years that with an accurate photometry and wavelength coverage, it is possible to obtain photometrically selected quasar samples with the low contamination and high completeness required from modern surveys.
The evaluation of photo-z's is then made possible by the existence of a rather complex correlation existing between the fluxes, as measured in broad band photometry, the types of galaxies and their distance. The search for such a correlation (a nonlinear mapping between the photometric parameter space and the redshift values) is particularly suited to data mining methods.
The existing methods for the evaluation of photometric redshifts can be broadly divided in two categories: SED template fitting and interpolative. In both cases the starting point is the availability of a catalogue of photometric multiband data for a large sample of objects, while in the case of interpolative methods a certain amount of spectroscopic counterparts is also required as training set.

2.1 Template fitting

These methods use libraries of galaxy spectra obtained either from real spectra or from synthetic Spectral Energy Distributions (SED). These templates

[1] In astronomy the terms color stands for the difference of the two magnitudes measured through two different bands. Since a magnitude is in first approximation the logarithm of a flux, colors correspond to the ratio between two fluxes measured at different wavelengths.

can be shifted to any redshift and then convolved with the transmission curves of the filter used in the photometric survey to create the template set for the redshift estimators. In general these methods are preferred when exploring new regimes (in terms of depth)in a survey or when a large set of spectroscopic observations is not available. To quote just a few, typical examples of such methods are Le Phare, [4], and HyperZ, [5].

2.2 Empirical methods

When spectroscopically determined redshifts are available for a fairly large subset of objects (hereafter knowledge base or KB) it is possible to use this information to uncover the hidden empirical correlation (often a non linear function) between the photometric observables of the sources and their redshifts, through a mapping function derived from the set of objects in the KB. Empirical methods have the advantage over the theoretical ones because they do not need accurate templates, being the dataset from which the mapping function is extracted made by real sky objects, and because they intrinsically include effects such as the filter bandpass and flux calibration. Obviously the KB needs to be large enough (about several thousands of objects) to provide a good coverage of the magnitudes and color space. Regarding the extension of this coverage (one of the limits of the empirical method), they can hardly extend out of the boundaries imposed by the magnitude and color limits of the KB.
In the years many estimators have been used to determine this empirical function, from linear and non linear fitting in the last decade to the use of Support Vector Machines (SVM) and Artificial Neural Networks in the last few years. An Artificial Neural Network is the empirical method used in this work.

Chapter 3

Neural Networks

Artificial Neural Networks (ANNs) are a family of models inspired by biological neural networks like the the nervous system in the human brain. They are generally presented as a system of interconnected "neurons" than can exchange information with each other.
The first artificial neuron called Threshold Logic Unit (TLU), the basic constituent of ANNs, was proposed by McCulloch & Pits in 1943 [9]. Its operating principle was very similar to how biological neurons actually work.
The basic structure of the TLU is shown in figure 3.1.

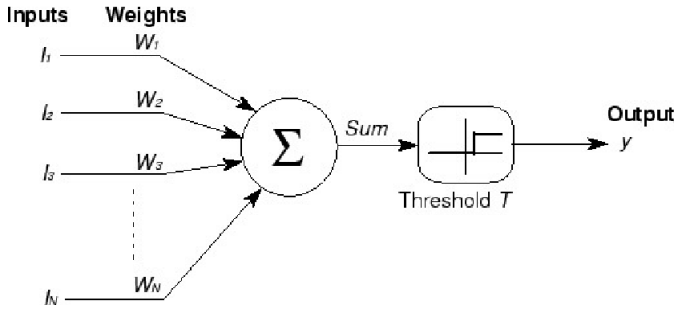

Figure 3.1: TLU scheme

It takes n input, each one with an associated weight w_i, then it performs the weighted sum of these input, if the sum is bigger than a previously set threshold, it gives as output 1, else 0 (in this case the activation function of the neuron is the simple step function). This is exactly how dendrons work.

Rosenblatt(1958) [17] introduced the first model of neural network: a two layer learning network called perceptron which had the ability to change the weights assigned to the input after each 'learning cycle'. This change was performed trying to minimize the difference between the actual output of the algorithm and the expected one. The architecture had its limits and Minsky & Papert (1969)[10] demonstrated its inability to solve simple non linearly separable functions like the XOR function.

These limitations were overcome by Werbos 1974, [22], and its Multi Layer Perceptron (MLP). This architecture had a hidden layer between the input and output layer and thanks to the backpropagation algorithm [18] was capable to solve non linearly separable functions.

3.1 Multi Layer Perceptron

In this section we will explain in few words the general structure and the operating principles of the Multi Layer Perceptron (MLP).
The MLP architecture is one of the most typical feed forward neural network model. The expression feed forward identify the fact that in this neural network model, the impulse is always propagated in the same direction, e.g. from the input layer to the output layer, passing through one or more hidden

layers, by combining the sum of weights associated to all neurons except the input ones.
The output of each neuron is obtained through an **activation function** applied to the weighted sum of input. The shape of the activation function can differ substantially from model to model, from the simplest linear function, to the hyperbolic tangent, which is the one used in this work by the model MLPQNA.
Concerning the training phase of the network, the weights are changed according to the particular learning rule in use, and until a predetermined distance between the network output and the desired know output is reached (usually this distance is decided a priori by the user and it is commonly known as Error Threshold).
Feed-forward neural networks provide a general framework for representing nonlinear functional mappings between a set of input variable and a set of output variables (Bishop 2006, [2]). One can achieve this goal by representing the nonlinear function of many variables by a composition of nonlinear activation functions of one variable, which formally describe the mathematical representation of a feed-forward neural network with two computational layers

$$y_k = \sum_{j=0}^{M} w_{jk}^{(2)} g(\sum_{i=0}^{d} w_{ji}^{(1)} x_i)$$

A multi -layer perceptron is made by several hierarchical layers:

1. the input layer (x_i) made by a number of nodes equal to the number of input variables (d)

2. the output layer made by a number of perceptrons (neurons) equal to the number of output variables (K) (1 in case of a regression)

3. one or more hidden layers, each one composed by an arbitrary number of perceptrons (M).

In a fully connected feed forward network, each node of a layer is connected to all the nodes in the adjacent layers. Each connection is represented by an adaptive weight, which is the strength of the synaptic connection between two neurons (w_{jk}^{l}). The response of each perceptron to the input is, as said before, represented by a nonlinear function g, referred to as the activation function. The above equation assumes a linear activation function for neurons in the output layer.

We refer to the topology of the MLP and the weight matrix of its connections as the **model**. To train the model that best fits the data, we have to provide the network with a set of examples. This set of examples is the training set extracted from the Knowledge Base (KB).

3.1.1 Backpropagation Algorithm

In this section we will discuss the most known learning rule used to train a MLP network, also known as the Backpropagation algorithm.
The algorithm allows to modify the weights associated to the neural connections in order to minimize a certain error function E. This function depends from the h-th output vector $out_k^h = f(x^h)$ given the x-th input vector x^h and the h-th target vector y^h. The training set is made up of all the couples (x^h, y^h) for $h = 1, ..., N$.
The most simple error function E, can be written as:

$$E = \frac{1}{2} \sum_h \sum_k (out_k^h - y_k^h)^2 + \frac{1}{2} \|W\|^2 \lambda$$

where the k index represents the value of the k-th output vector. The error E depends on the weights and to be minimized, different strategies (algorithms) can be used. One of the most used is the gradient-descent algorithm. The algorithm starts from a generic point x^0 and evaluate the error function gradient in that point $\nabla f(x^0)$. The gradient gives the direction of the maximum increment of the function (decrement if one considers $-\nabla$). Defined the direction, one moves from x^0 of a previously defined distance η (learning rate) landing in a new point x^1 on which the gradient is recalculated. The process is iterated until the gradient is null.
The algorithm stops after a chosen number of iterations or, as we said earlier, if the error becomes smaller than a chosen threshold. When the learning/training phase is concluded, the trained network can be used like a simple function.

3.1.2 Quasi Newton Algorithm

The algorithm used in our MLP to find the minimum is the Quasi Newton algorithm which is, compared with the GDA, more efficient in avoiding local

minima and more accurate in the error function trend follow-up, thus revealing a natural capability to find the absolute minima of the error (Shanno 1990, [19]).
QNA differs from the Newton algorithm in terms of the calculation of the Hessian of the error function. The traditional Newton method uses the Hessian of a function to find the stationary point of a quadratic form. The Hessian of a function is not always available and in many cases it is far too complex to be computed. More often we can only calculate the function gradient, which can be used to derive the Hessian via N consequent gradient calculations. The gradient in every point w is in fact given by

$$\nabla E = H \times (w - w^*)$$

where w corresponds to the minimum of the error function, which satisfies the condition

$$w^* = w - H^{-1} \times \nabla E$$

The vector $-H^{-1} \times \nabla E$ is known as the Newton Direction and it is the traditional base for a variety of optimization strategies. The step of this traditional method is defined as the product of the inverse Hessian matrix and a function gradient. If the function is a positive definite quadratic form, the minimum can be reached in just one step, while for an indefinite quadratic form (which has no minimum), we will reach either the maximum or a saddle point. To solve this problem, quasi Newton methods proceed with a positive definite Hessian approximation. So far, if the Hessian is positive, we take the step using the newton method. If, instead it is indefinite, we first modify to make it definite positive, and then perform a step using the Newton, method which is always calculated in the direction of the function decrement.
Instead of calculating the H matrix and its inverse, the QNA uses a series of intermediate steps of lower computational cost to generate a sequence of matrices that are more and more accurate approximations of the inverse Hessian. During the exploration of the parameter space and in order to find the minimum error direction, QNA starts in the wrong direction. This direction is chosen because at the first step the method has to follow the error gradient, so it takes the direction of steepest descent. However, in subsequent steps, it incorporates information from the gradient. By using the second derivatives, QNA is able to avoid local minima and to follow the error function trend more precisely, revealing a "natural" capability to find the absolute minimum error of the optimization problem.

3.2 Implementation of the MLPQNA

After the network is trained, it has to be tested to evaluate the overall performance of the model. Before testing there is, usually, an intermediate step, the validation phase, in which the model is checked against loss of generalization capabilities (a phenomenon also known as over-fitting).
In fact, in statistics and machine learning, one of the most common tasks is to fit a "model" to a set of training data, so as to be able to make reliable predictions on general untrained data. Overfitting occurs when a statistical model describes random error or noise instead of the underlying relationship. This generally occurs when a model is excessively complex, such as having too many parameters relative to the number of observations. A model that occur in overfitting will generally have poor predictive performance, as it can exaggerate minor fluctuations in the data.
The possibility of overfitting exists because the criterion used for training the model is not the same as the criterion used to judge the efficiency of a model. In particular, a model is typically trained by maximizing its performance on some set of training data. However, its efficiency is determined not by its performance on the training data but by its ability to perform well on unseen data. Overfitting occurs when a model begins to "memorize" training data rather than "learning" to generalize from trends. As an extreme example, if the number of parameters is the same as or greater than the number of observations, a simple model or learning process can perfectly predict the training data simply by memorizing the training data in its entirety, but such a model will typically fail drastically when making predictions about new or unseen data, since the simple model has not learned to generalize at all.
To avoid over-fitting a procedure called k-fold cross validation [11] has been used in this work; in the k-fold cross-validation, the original sample is randomly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is used as the validation set for testing the model, and the remaining k - 1 subsamples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation set. The k results from the folds can then be averaged (or otherwise combined) to produce a single estimation.
The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once. In this way the possibility of overfitting

the data is greatly reduced. A 10-fold cross-validation is commonly used, but in general k remains a free parameter.

The algorithm for MLP with QNA learning rule is the following.
Let us consider a generic MLP with $w^{(t)}$ being the weight vector at time (t).

1. Initialize all weights $w^{(0)}$ with small random values (typically normalized in [-1,1]), set the constant error tolerance ε and t = 0
2. present to the network all training set and calculate $E(w^{(t)})$ as the error function for the current weight configuration
3. if t = 0 then $d^{(t)} = -\nabla E^{(t)}$
4. else $d^{(t)} = -\nabla E^{(t-1)} + Ap + B\nu$ where $p = w^{(t+1)} - w^{(t)}$ and $\nu = g^{(t+1)} - g^{(t)}$
5. calculate $w^{(t+1)} = w^{(t)} - \alpha^{(t)}$, where α is obtained by line search equation
6. calculate A and B for the next iteration
7. if $E(w^{(t+1)}) > \varepsilon$ then $t = t + 1$ and goto (2), else STOP

The MLPQNA model used in this work was introduced in astronomy and widely tested on a variety of science cases by Brescia et al. 2013, [6]. It was also made available through the Data Mining & Exploration Web Application REsource (DAMEWARE, Brescia et al. 2012, [7]).

3.3 Statistical Estimators

The statistical estimators used in this work to evaluate and compare the regression performance are the same used in Brescia et al. 2013, [6]):

$$bias(x) = \frac{\sum_{i=1}^{N} x_i}{N}$$

$$\sigma(x) = \sqrt{\frac{\sum_{i=1}^{N}\left[x_i - \frac{\sum_{i=1}^{N} x_i}{N}\right]^2}{N}}$$

$$MAD(x) = Median(|x|)$$

$$NMAD(x) = 1.48 \times Median(|x|)$$

$$RMS(x) = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N}}$$

where σ is the standard deviation, MAD is the Median Absolute Deviation, NMAD is the normalized MAD and RMS is the Root Mean Square. In this work the variable x could be either $\Delta z = (z_{spec} - z_{phot})$) or

$$\Delta z_{norm} = (z_{spec} - z_{phot})/(1 + z_{spec})$$

the use of Δz_{norm} takes into account the fact that the error on z is a function increasing with z itself.

Chapter 4

Photometric redshifs for quasars with MLPQNA

4.1 The Knowledge Base

As a rule of thumb, in machine learning supervised methods it is a common practice to populate the training, validation and test set with respectively the 60%, 20% and 20% of the KB (Keams 1996, [8]).
However, by taking into account the results obtained by Brescia ad Cavuoti in their works,(Brescia et al. 2013, [6]) we reduced the training + validation set from 80% of the KB to 60% in order to have a larger set of objects on which to verify the performance of the model as well as faster execution of the training phase.
Furthermore, in order to ensure a proper coverage of the KB in the Parameter Space (PS), the data objects were divided using a random extraction, and this process was iterated several times to minimize the possible biases induced by fluctuations in the coverage of the PS, such as small differences in the redshift distribution of training and test samples used in the experiments.
The automatized process of the cross-validation was done by performing ten different training runs with the following procedure:

- splitting of the training/validation set into ten random subsets, each one composed by 10% of the dataset;
- at each training run we applied the 90% of the dataset and the excluded 10% for validation.

This work uses as KB a multi-band catalogue of quasars built by following criteria very similar to those used in Brescia et al. 2013, [6]. The catalogue was obtained starting from spectroscopically selected quasars extracted from the Sloan Digital Sky Survey (SDSS) Data Release 9 (DR9). These objects were crossmatched with data sets from the Galaxy Evolution Explorer (GALEX) GR6/7, UKIRT Infrared Deep Sky Survey (UKIDSS) Data Release 10 (DR10) and the Wide-Field Infrared Survey Explorer (WISE). The crossmatching procedure was done with a maximum search radius of $1.5''$ and only sources with a unique counter-part in the search circle for each one of the four surveys were retained.
The KB obtained consists of $16,446$ objects.

4.2 The Four astronomical Surveys

Below we briefly describe the four surveys, the data extraction process and the pruning process that we performed on the data to create the KB.

- **SDSS**: the SDSS DR9, [13], catalogue contains five band photometry for $1,231,051,050$ objects and $1,843,200$ spectra with a total area covered of $31,637$ square degrees (including overlaps) with an angular precision of 1.3 arcsec.

u	g	r	i	z
354 nm	475 nm	622 nm	763 nm	905 nm
22.0	22.2	22.2	21.3	20.5

Table 4.1: SDSS effective wavelengths and magnitude limits (95% completeness for point sources)

- **GALEX**: is a two bands UV survey (fuv and nuv coverage $[140 - 170]$ nm and $[180 - 275]$ nm wavelength ranges respectively, [14]
  GALEX includes an all-sky imaging survey of $26,000$ square degrees, a medium imaging survey of 1000 sq. deg. and a deep imaging survey of 80 sq. deg., with a limiting magnitude in the nuv band of 20.5, 23.5

and 25 respectively and an angular resolution of 4″.5 (fuv) and 6″.5 (nuv).

- **UKIDSS**:UKIDSS is a set of five surveys, [15]. The LAS (Large Area Survey) covers an area of 4000 square degrees at high Galactic latitudes in the $YJHK$ filters (cf. Table 4.2)to a depth $K = 18.4$.

Y	J	H	K
1020 nm	1220 nm	1630 nm	2190 nm
20.5	20.0	18.8	18.4

Table 4.2: UKIDSS effective wavelengths and magnitude limits

- **WISE**: WISE is an all-sky mid-infrared survey in the four bands W1, W2, W3, W4 centered at 3.4, 4.6, 12 and 22 μm respectively, with angular resolutions of 6″.1, 6″.4, 6″.5 and 12″.0 and photometric sensitivity of 16.5, 15.5, 11.2 and 7.9 (in the Vega magnitude system), [16].

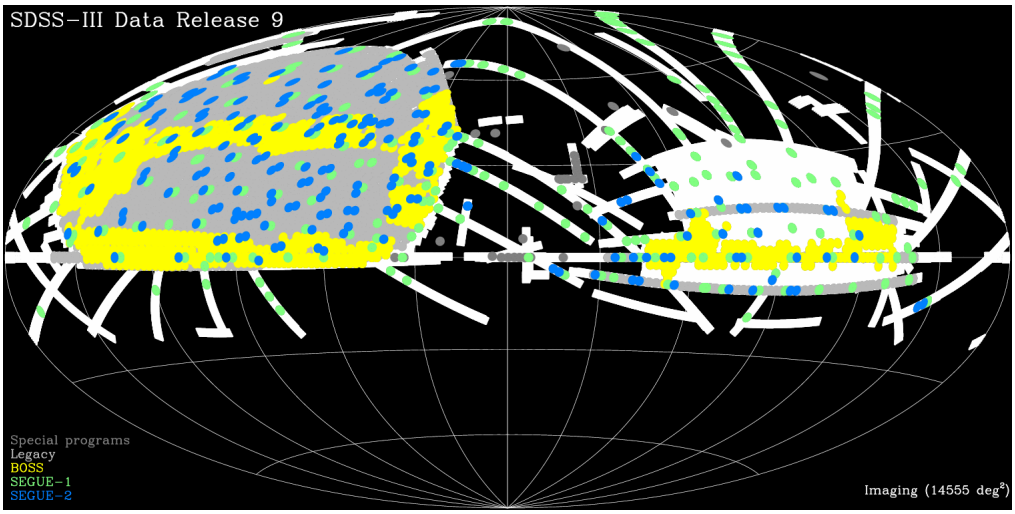

Figure 4.1: DR9 imaging and spectroscopic coverage in Equatorial coordinates (plot centered at RA = 6h, or 90 deg.)

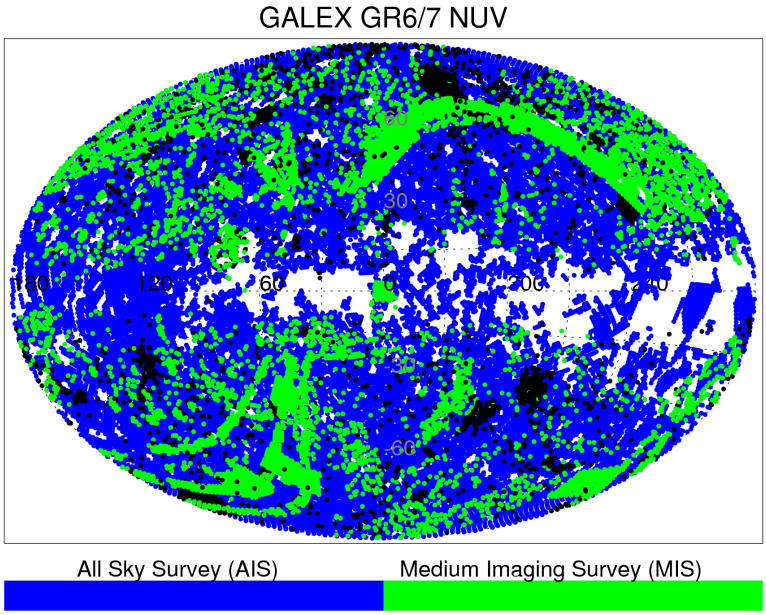

Figure 4.2: GALEX sky coverage map

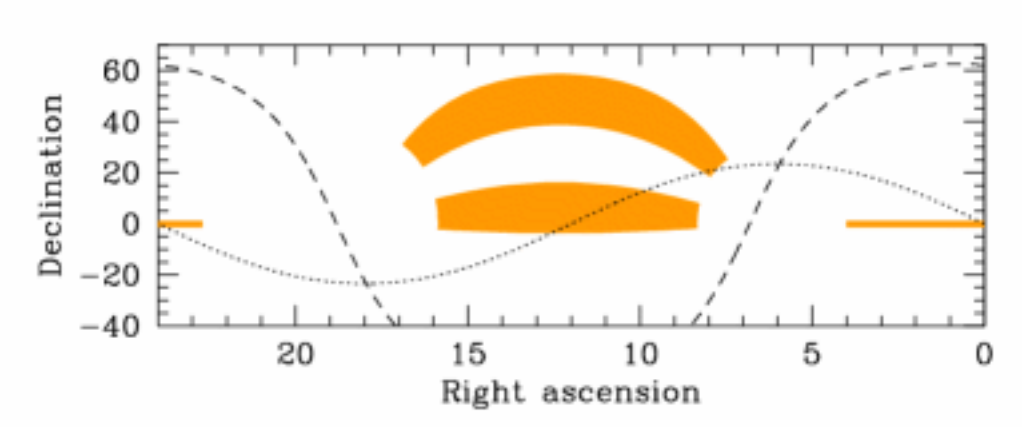

Figure 4.3: UKIDSS LAS sky coverage map

4.2.1 Data Pruning

In this section we describe the details of the performed queries for the surveys and the pruning criteria adopted to clean the data.

The query was:

- **SDSS Query**:

```
SELECT  sp.specObjID, sp.ra, sp.dec,
  sp.z, sp.zErr, sp.class, sp.subClass, sp.zWarning,
  sp.sciencePrimary, sp.objID, sp.type,
  sp.psfMag_u, sp.psfMag_g, sp.psfMag_r, sp.psfMag_i, sp.psfMag_z,
  sp.psfMagErr_u, sp.psfMagErr_g, sp.psfMagErr_r,
  sp.psfMagErr_i, sp.psfMagErr_z,
  sp.petroMag_u, sp.petroMag_g, sp.petroMag_r,
  sp.petroMAg_i, sp.petroMag_z,
  sp.petroMagErr_u, sp.petroMagErr_g, sp.petroMagErr_r,
  sp.petroMagErr_i, sp.petroMagErr_z,
  sp.modelMag_u, sp.modelMag_g, sp.modelMag_r,
  sp.modelMag_i,sp.modelMag_z,
  sp.modelMagErr_u, sp.modelMagErr_g, sp.modelMagErr_r,
  sp.modelMagErr_i, sp.modelMagErr_z,
  sp.cModelMag_u, sp.cModelMag_g, sp.cModelMag_r,
  sp.cModelMag_i, sp.cModelMag_z,
  sp.cModelMagErr_u, sp.cModelMagErr_g,
  sp.cModelMagErr_r,sp.cModelMagErr_i, sp.cModelMagErr_z,
  sp.type, sp.zWarning, sp.flags
  into mydb.SDSS_5 from SpecPhotoAll as sp

  where sp.z > 0 and class = 'QSO' and sp.objID > 0 and sp.zWarning = 0
ORDER BY specObjID
```

From this query we obtained $231,198$ objects on which we performed the following cleaning process:

- discarded all the points with $ObjID \leq 0$;
- discarded all the points with $zErr < 0$ and $zErr > 0.01$;

- discarded all the points with $zWarning \neq 0$ in order to retain only objects with no redshift problems detected;
- eliminated all the object with the same objID leaving inside the table the one with the lowest zErr;
- eliminated all the points with magnitude errors greater than 1;

The cleaning process left us with $174,676$ points.

- **GALEX Query**:

```
SELECT n.ra, n.dec, n.objID, n.matched_id,
       p.objID, p.ra, p.dec, p.fuv_mag, p.fuv_magerr,
       p.FUV_MAG_ISO, p.FUV_MAG_AUTO,
       p.FUV_MAGERR_AUTO, p.FUV_MAGERR_ISO,
       p.FUV_MAG_APER_1, p.FUV_MAG_APER_2, p.FUV_MAG_APER_3,
       p.FUV_MAGERR_APER_1, p.FUV_MAGERR_APER_2, p.FUV_MAGERR_APER_3,
       p.nuv_mag, p.nuv_magerr, p.NUV_MAG_ISO, p.NUV_MAG_AUTO,
       p.NUV_MAGERR_AUTO, p.NUV_MAGERR_ISO,
       p.NUV_MAG_APER_1, p.NUV_MAG_APER_2, p.NUV_MAG_APER_3,
       p.NUV_MAGERR_APER_1, p.NUV_MAGERR_APER_2, p.NUV_MAGERR_APER_3,
       p.fuv_kron_radius, p.nuv_kron_radius,\\
       p.fuv_artifact, p.nuv_artifact, p.fuv_ambg, p.nuv_ambg,
       p.nuv_maskpix, p.fuv_maskpix, p.fuv_nc_r1, p.nuv_nc_r1
       into mydb.Galex_II from mydb.SDSS_N as n
       LEFT OUTER JOIN PhotoObjAll as p on n.matched_id = p.objID
```

To avoid exceeding Galex output limit, we used the results of the SDSS query to search for specific objects in the Galex database. First we extracted all the points in a $1''.5$ radius around the SDSS points (these are individuated by their SDSS objID and their right ascension and declination) and their matched id (the Galex database ids corresponding to the SDSS ids).
Then we used these matched id's to make a cross-match in the Galex database and search all the points with all the features of our interest and with objID equal to the corresponding matched id. (Neighbors Search).

Our SDSS - GALEX neighbours search gave us $191,737$ points on which we performed the following cleaning process:

- eliminated all the objects presenting a value of the nuv-maskpix and fuv-maskpix flags different from 0;
- eliminated, as suggested by GALEX clean Photometry guide, all the points presenting the fuv-artifact and nuv-artifact flags with values 2, 4 and 32;
- as done with SDSS, we have eliminated all the points with magnitude errors greater than 1;

This cleaning process left us with $91,373$ points.
To use the magnitudes for the training process we modified them with their ZPV:

```
Fuv_mag = ZPV + FUV_MAG_AUTO
Nuv_mag = ZPV + FUV_MAG_AUTO
```

where ZPM (Zero Point Value) is 18.82 for the first and 20.08 for the latter.

- **UKIDSS Query**:

```
SELECT sourceID, ra, dec,
      yHallMag, yHallMagErr, yPetroMag, yPetroMagErr, yPsfMag, yPsfMagErr,
      yAperMag3, yAperMag3Err, yAperMag4, yAperMag4Err,
      yAperMag6, yAperMAg6Err, yErrBits, yppErrBits,

      j_1HallMag, j_1HallMagErr, j_1PetroMag, j_1PetroMagErr, j_1PsfMag,
      j_1PsfMagErr,
      j_1AperMag3, j_1AperMag3Err, j_1AperMag4, j_1AperMag4Err,
      j_1AperMag6, j_1AperMAg6Err, j_1ErrBits, j_1ppErrBits,

      j_2HallMag, j_2HallMagErr, j_2PetroMag, j_2PetroMagErr, j_2PsfMag,
      j_2PsfMagErr,
      j_2AperMag3, j_2AperMag3Err, j_2AperMag4, j_2AperMag4Err,
      j_2AperMag6, j_2AperMAg6Err, j_2ErrBits, j_2ppErrBits,
```

```
hHallMag, hHallMagErr, hPetroMag, hPetroMagErr, hPsfMag, hPsfMagErr,
hAperMag3, hAperMag3Err, hAperMag4, hAperMag4Err,
hAperMag6, hAperMAg6Err, hErrBits, hppErrBits,

kHallMag, kHallMagErr, kPetroMag, kPetroMagErr, kPsfMag, kPsfMagErr,
kAperMag3, kAperMag3Err, kAperMag4, kAperMag4Err,
kAperMag6, kAperMAg6Err, kErrBits, kppErrBits  FROM lasSource
```

The SDSS - UKIDSS neighbors search gave us $65,892$ points on which we performed the following cleaning process:

- discarded all the points presenting null values on the magnitude measurements;
- discarded all the points with magnitude errors greater than 1;
- discarded all the points presenting a ppErr flag value different from 0;

This cleaning process left us with $41,128$ points.
All the calculations made by UKIDSS to elaborate magnitudes use the Vega system fluxes in the different bands as zero point values. To use the magnitudes in our experiments, they must be converted in AB magnitudes. This correction is performed by adding these offset values to the UKIDSS magnitudes:

Y	J	H	K
0.634	0.938	1.379	1.900

Table 4.3: Conversion to the AB system ($mAB = mVega+?m$)

- **WISE Query**: The data was obtained from a neighbours search in the ALLWISE catalogue using a circular search of radius $1''.5$ around the SDSS points.
  The SDSS - WISE neighbours search gave us $150,711$ points on which we performed the following cleaning process:
  - discarded from our dataset all the objects with $ccflag \neq 0000$ (4% of the data) and $extflg \neq 0$ (1% of the data);

– discarded all the points with magnitude errors greater than 1;

This cleaning process left us with $142,849$ points;
The in-band fluxes of the Vega spectrum were adopted to define zero magnitude for WISE bands W1, W2, W3 and W4; this means that we have to correct them. This correction is performed by adding these offset values to the WISE magnitudes:

W1	W2	W3	W4
2.683	3.319	5.242	6.604

Table 4.4: Conversion to the AB system ($mAB = mVega + \Delta m$)

4.2.2 Cross-matches and KB finalization

All the flags acronyms refer to the flags in the Appendix A

Table 4.5: Data Cleaning Process Summarize Table

Survey	**Original Data**	**Clean Data**	**Used Flags**	**Unused Flags**
SDSS	231,198	174,676	S4, S5, S8, S10, S13.1, S14.1, S15.1, S16.1	S9, S12
GALEX	191,737	91,373	G4 - G27, G29, G32, G33	
UKIDSS	65,892	41,128	U2, U2.1, U3, U3.1, U5, U5.1, U6, U6.1, U7, U7.1, U9	
WISE	150,711	142,849	W2, W9, W10	W11

Table 4.6: Cross-matches Table

Cross-match	**Points**
SDSS - GALEX	58,799
SDSS - UKIDSS	41,128
SDSS - WISE	137,506
SDSS - WISE - UKIDSS	36,612
SDSS - WISE - GALEX	53,507
SDSS - UKIDSS - GALEX	17,711
SDSS - UKIDSS - GALEX - WISE	16,452

After the cross-match was completed and, unfortunately after the Network training was completed as well, we discovered in the dataset of $16,452$ objects, 6 objects with null values in the W3 band. We removed them and so all the results of this work are obtained from a dataset of $16,446$ objects and are based on the realistic assumption that these five points did not have a strong effect on the network training.

Table 4.7: Available Magnitudes per band Table

Survey	**Bands**	**Available Magnitudes**
SDSS	u, g, r, i, z	psfMag, petroMag, ModelMag, cModelMag
GALEX	fuv, nuv	Mag, MagIso, MagAuto, magAper1, MagAper2, MagAper3
UKIDSS	Y, J, H, K	psfMag, PetroMag, HallMag, AperMag3, AperMag4, AperMag6
WISE	W1, W2, W3, W4	W1mpro, W2mpro, W3mpro, W4mpro

4.3 Model Set-up

In terms of the internal parameter set-up of the MLPQNA, we need to consider the following topological parameters:

- input neurons: a variable number of neurons, corresponding to the number of input parameters of the PS used in the experiments;
- first hidden layer neurons: a variable number of hidden neurons, depending on the number N of input neurons (features in the dataset), equal to $2N + 1$ as rule of thumb;
- second hidden layer neurons: a variable number of hidden neurons ranging from 0 (without second hidden layer) to $N - 1$;
- output neurons: one neuron (regression problem), returning the predicted zPhot value;

For the QNA learning rule, we fixed the following values as best parameters as it was done in (Brescia et al. 2013, [6]):

- step: 0.0001 (one of the two stopping criteria. The algorithm stops if the approximation error step size is less than this value. A step value equal to zero means to use the parameter MaxIt as the unique stopping criterion);
- res : 40 (number of restarts of Hessian approximation from random positions, per-formed at each iteration);

- dec : 0.1 (regularization factor for weight decay. The Tikhonov regularization term $dec \times ||networkweights||^2$ is added to the error function, where network weights is the total number of weights in the network, directly depending on the total number of neurons inside. When properly chosen, the generalization performance of the network is highly improved);
- MaxIt: 8000 (max number of iterations of Hessian approximation. If zero the step parameter is used as stopping criterion); (in some experiments we used $10,000$ instead of $8,000$);
- CV (k): 10 (k-fold cross validation, with k = 10);
- Error evaluation: Least Square Error + Tikhonov regularization (between target and network output).

At this point we had to decide which parameters to use from the parameter space available. Following, as usual, the recipes in Brescia et al. 2013, we used as input parameters the following combination of magnitudes and colors to obtain a set of 15 input parameters (in their work they proved it was the best possible combination of features on which to train the network):

1. 4 reference magnitudes (1 for each survey): rpsfMag for visible band, nuvMag for UV band, KHallMag for NIR band and W4mpro for IR band;
2. 11 colors: fuv - nuv, u - g, g - r, r - i, i - z, Y - J, J - H, K - H, W1 - W2, W2 - W3, W3 - W4;

Having 15 input features, we had 15 neurons in the input layer, 31 in the first hidden one, 14 in the second and 1 output parameter. With these parameters, we obtained the statistical results reported in Section (4.5).

4.4 Catastrophic Outliers

Following the definition commonly adopted in the specialised literature, we define an object as a catastrophic outlier if it satisfies the condition:

$$|x| > 0.15$$

where x might be Δz or Δz_{norm} depending on the statistical indicator used. Of course, in the hypothesis that the distribution of Δz is Gaussian, one could decide to use other definitions of outliers; $|x| > \sigma$ or $|x| > 2\sigma$ or $|x| > 1.48 \times \sigma$. The decision on which one of this boundaries define outliers is a choice more based on experience than on physics itself.

4.5 Results

In this section we present the results of our four survey experiments. All the statistics are calculated on the blind test set, in order to evaluate the performance of our model on new previously "unseen" data.
We did not run just a single experiment, but ten of them for the following reasons:

1. in order to confine and, ultimately, try to exclude the effects of quasar variability on our photometric redshift evaluation.

2. in order to better understand the effect of the training process on the network performance

3. to try to separate the outliers that are statistical in nature from those that are physical in nature (see Section 5)

The following table contains the statistical estimators derived for:

- the first experiment out of ten (named as *Single*);

- the same experiment removing from the test set all the points with standard deviation higher than 0.08 (named as *Single low SD*)(for an explanation see Section 5.2);

- the statistics calculated using, instead of a single experiment photometric redshift evaluation, the average of the ten evaluations done in this work (named as *Average*);

- the same as in the previous item but removing from the test set all the points with standard deviation higher than 0.08 (named as *Average low SD*).

Table 4.8: Statistical estimators calculated on Δz_{norm}. The values are relative to the test set only consisting of $6,580$ objects in the *Single* and *Average* experiments and $5,581$ in the *low SD* ones.

	Single	**Single low SD**	**Average**	**Average low SD**
$\sigma(\Delta z_{norm})$	0.1168	0.0406	0.0768	0.0348
MAD(Δz_{norm})	0.0231	0.0210	0.0184	0.0166
RMS(Δz_{norm})	0.1170	0.0406	0.0772	0.0349
NMAD(Δz_{norm})	0.0343	0.0311	0.0273	0.0245
$\Delta z_{norm} > 0.15$	3.24%	0.43%	2.36%	0.32%

Chapter 5

Characterization of Outliers

One of the main goals of this work was to try to understand why some objects are catastrophic outliers and other are not, with the ultimate goal of trying to see whether this may provide insights on the flaws of the model or, maybe, on its hidden potentialities. In the following chapter we will present some useful considerations.

5.1 Intersecting Outliers and Average

In our ten experiments we had more or less the same number of outliers per experiment, in particular these are the outliers percentages per experiment:

Exp ID	**Outliers Percentage**
1	3.24%
2	3.39%
3	2.99%
4	3.14%
5	3.40%
6	2.93%
7	2.99%
8	3.46%
9	3.74%
10	3.22%

First of all we need to remember that the training/test runs were done using always the same train and test set. The mean fraction of outliers is therefore 3.252%. After that we have calculated our statistical estimators using, as the photometric redshift measure, the average of the single experiment redshift estimations ($zPhot_i$, $i = 1, ..., 10$) (the average photometric redshift estimation, hereafter *zPhotMed*), finding a percentage of outliers in the test set of 2.36%. This means that the average number of outliers per experiment is higher than the number of outliers one would find using the average of the redshift estimations as true estimate of the redshift, or, in other words, that running the same experiment several times and using the average of the redshift estimates as true value, improves the performance of the model.

Then we studied the outliers frequency in the experiments, i.e. of the redshift estimations as the true estimate of the redshift, or, in other words, we have verified if one or more sources are labeled as outliers in more than one experiment. We found a common core of 26 sources that are labeled as outliers in all ten experiments (frequency = 10) and, summing all the outliers from all the experiments without repetition, a set of 752 outliers (frequency = 1).

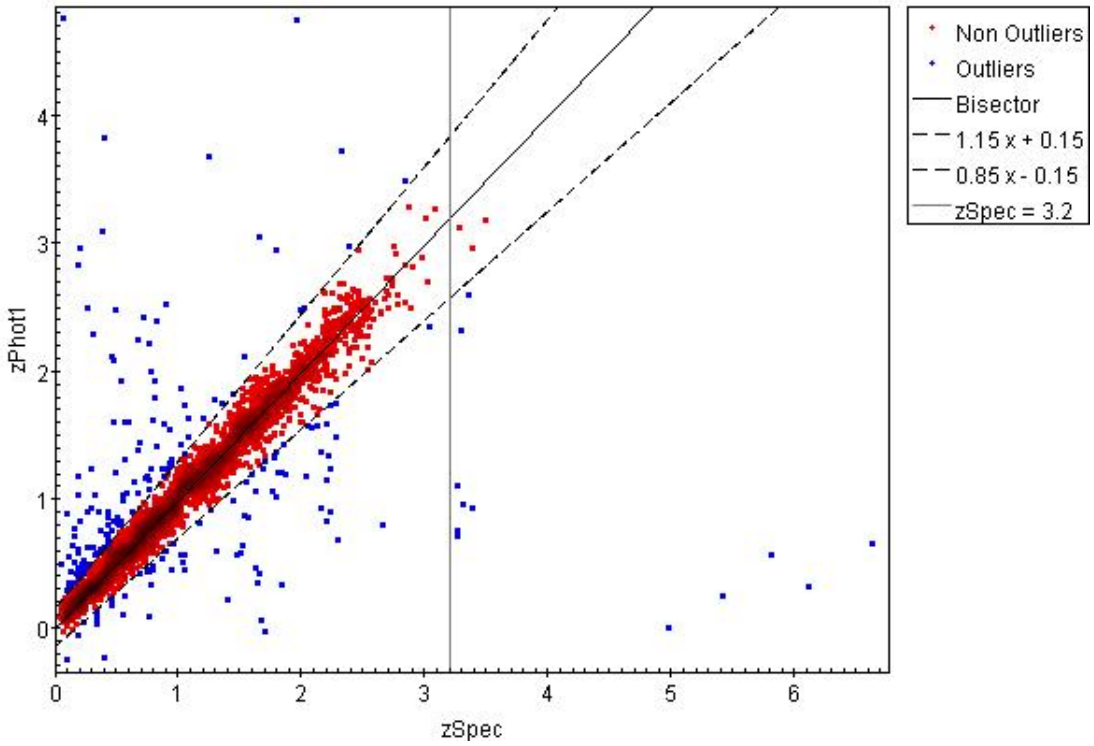

Figure 5.1: zSpec vs zPhot scatter plot of the first experiment. The points in blue are the outliers, the two lines that separate outliers from non outliers are the two solution of the "outliers definition" equation: $|(z_{spec} - z_{phot})/(1 + z_{spec})| > 0.15$

As it clear from the Fig. 5.1, a subset of the outliers corresponds to objects at high redshift ($zSpec \geq 4.9$). The reason why this happens is the very low number of objects with high redshift ($zSpec \geq 3.2$) that we have in our data-set (28 objects out of $16,446$).
As we specified in Section 2.2, empirical methods do not perform well out of their PS boundaries and clearly these high redshift objects are outliers not for physical reasons, but because there were not enough "similar" points on which to train the model.

Using the zPhotMed as true estimate of the photometric redshift, our model classified as outliers 155 objects, on average 50 less than we would have obtained using anyone of the ten zPhot estimates.

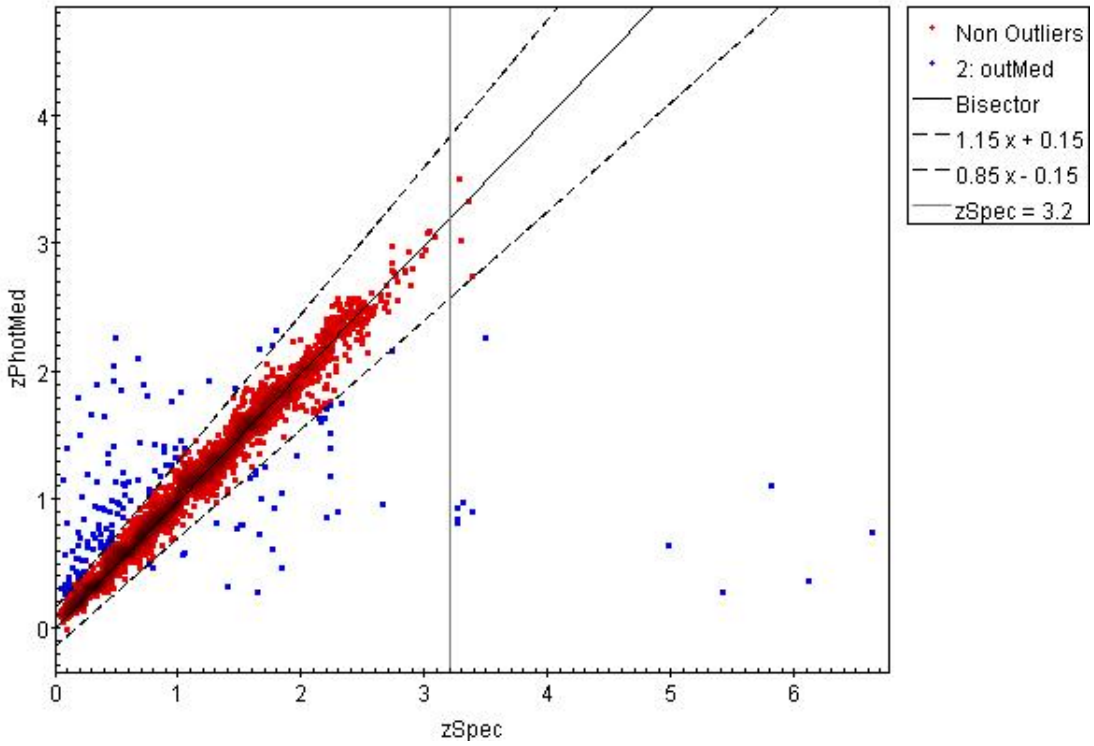

Figure 5.2: zSpec vs zPhotMed scatter plot of the Average experiment. The points in blue are the Outliers; as you can see the performance of the model has increased, decreasing the number of outliers.

5.2 Standard Deviation experiment

Having noticed an improvement in precision using the average of the ten photometric redshift estimations as the true estimate, we decided to study the extent of the variation of the zPhot estimations around the average and, to do so, we computed the standard deviation for every object in the test set.

Plotting in an histogram the standard deviations, we find that this standard deviation could be used to discriminate outliers.
As it can be seen from figures 5.3 and 5.4, the outliers have, in general, a higher standard deviation when compared to non outliers. We found that, by discarding from the test set all the objects with standard deviation σ higher than 0.08, (the vertical black line in figures 5.3 and 5.4), we remove

the 85.37% of outliers (642 out of 752), loosing only 6.13% of the non-outliers (357 out of 5828).
After eliminating from the test-set all objects with $\sigma > 0.08$ we used this new test-set to compute the statistical estimators; what we found was a considerable boost in performance and precision of the model as it can be seen in 4.8).

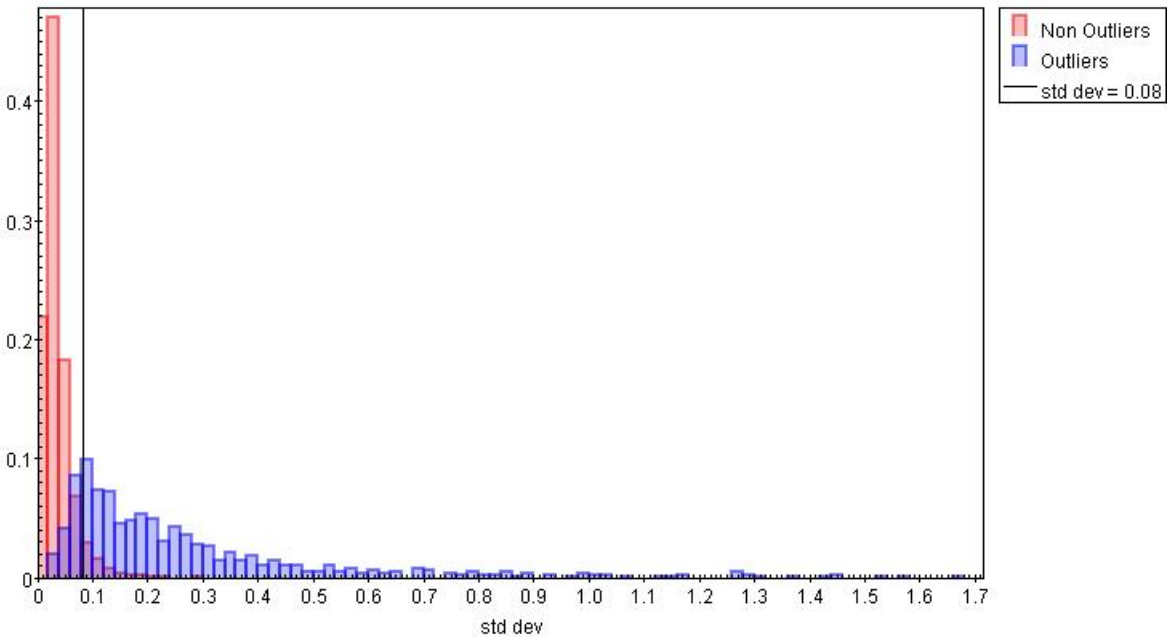

Figure 5.3: Full histogram of the standard deviation

After the cut in standard deviation, the number of objects labeled as outliers in all ten experiments dropped from 26 to just 5, while the number of outliers in the *Average* experiment dropped from 155 to 18.

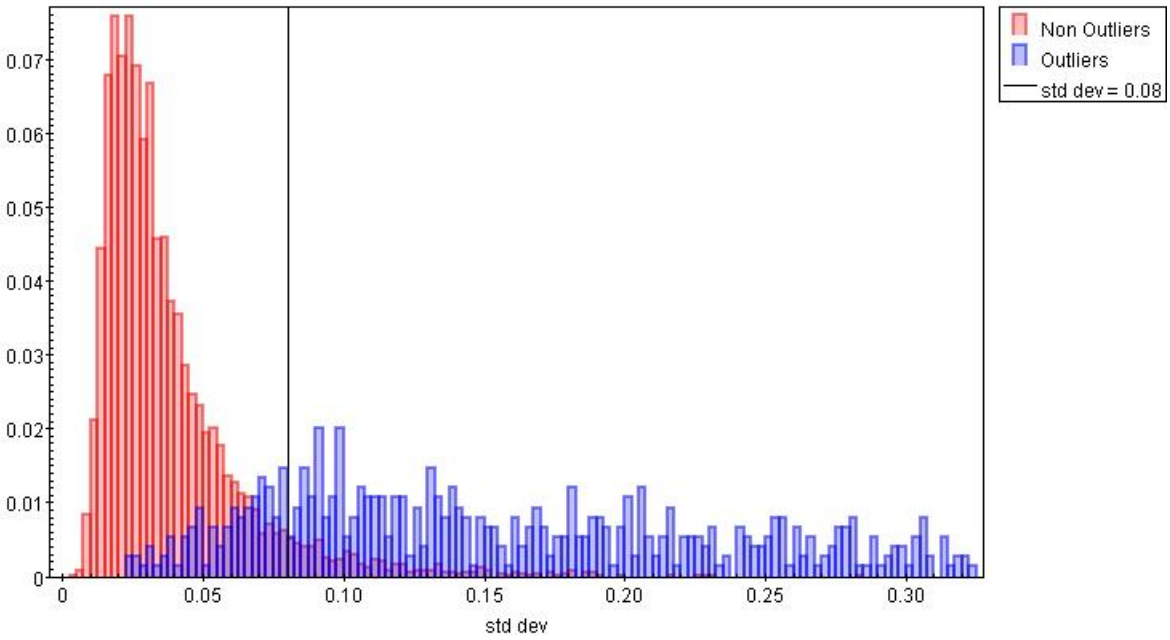

Figure 5.4: PRevious histogram zoomed in the cut region

5.3 n-dimensional distance

In our effort to characterize outliers, we decided to calculate the distances of the test set objects from the train set objects in terms of their photometry. Firstly we normalized between 0 and 1 all the features and then we used them to calculate the $15-dimensional$ Euclidean distances of each test-set object from each train-set object.

$$d = \sqrt{\sum_{i=1}^{15}(y_i - x_i)^2}$$

where x_i is a i-th feature of an object in the test-set and y_i is the corresponding feature of an object in the train-set.
We discovered that, on average, the outliers are more distant from the train-set objects than the non outliers with an average distance of 0.7129 against 0.6479 for the non outliers.
We then checked if our cut on the standard deviation of the test-set objects (see Standard Deviation experiment) would result in a variation in the average distance of the outliers. What we found is that outliers with $\sigma < 0.08$

are, on average, less distant from the train-set point if compared with the ones with $\sigma > 0.08$, having an average distance of 0.6720 against 0.7129.
We repeated the same calculation for the *Average* experiment finding an average distance of 0.7389 for the outliers and 0.6533 for the non outliers. After the standard deviation cut the distances became 0.6677 for the outliers and 0.6533 for the non outliers, thus confirming that the objects on which the model has a worst prediction are those who also have higher distances from the train set in the $15 - th$ dimensional feature space.

5.4 Photometric Quality

To confirm out hypothesis that the model does not perform well on sources with bad photometry, we used the previously unused photometry flags:

- DEBLENDED_ NOPEAK
- science_primary = 0
- Ph_qual

For a detail description of this flags see the Appendix A.
We found a higher percentage of objects with "bad photometry" among outliers than among non outliers.

	Outliers	**Non-Outliers**
DEBLENDED_NOPEAK	40%	32.11%
Ph_qual	46.54%	33.49%
Science_Primary	11%	4.95%

Table 5.1: Percentage of objects with active photometric flags between outliers and non-outliers

	Outliers	Non-Outliers
DEBLENDED_NOPEAK	40%	31.33%
Ph_qual	45.51%	34.73%
Science_Primary	5.8%	5.04%

Table 5.2: Percentage of objects with active photometric flags between outliers and non-outliers in the *Average* experiment

	Outliers	Non-Outliers
DEBLENDED_NOPEAK	61.1%	27.27%
Ph_qual	47.7%	34.01%
Science_Primary	11.1%	4.26%

Table 5.3: Percentage of objects with active photometric flags between outliers and non-outliers in the *Average* experiment after the standard deviation cut

As you can see from Table 5.3, all objects that are still classified as outliers after the standard deviation cut $\sigma \leq 0.08$ present some occurrences of bad photometry flags.

5.5 Blazars and Lensed Quasars search

In order to check whether the outlier/non outlier nature of a given object is or not correlated with some intrinsic peculiarities of the object itself, we cross-checked our data-set with a Blazar catalogue kindly provided by Raffaele D'Abrusco (an unpublished catalogue of 15248 confirmed Blazars; here after RD-BL-catalogue) and a lensed quasar catalogue compiled by Oguri et al. 2012, [12], containing 26 confirmed lensed quasars.

5.5.1 Blazars

Cross-matching our test set with the Raffaele D'Abrusco Blazars catalogue, we found 45 Blazars. Using the *zPhotMed* as true estimate of the photometric redshift (see the Section 5.1) and the normalized distance δz_{norm} to distinguish outliers from non outliers, we found that 5 Blazars out of the 45

were labeled by our model as outliers. Of this 5, 2 have zSpec values higher than 3.2 and therefore are labeled as outliers for the same problem explained at the end of Section 5.1 and not for physical reasons. When we applied the standard deviation cut $\sigma < 0.08$ we retained 34 of the 45 Blazars and all of them were non outliers.
Then, to understand why the model performed relatively "well" on Blazars, we cross-matched our train set with the Blazar catalogue finding 57 confirmed Blazars.
By cross-matching the 5 Blazars classified as outliers with the 26 outliers that we found in every experiment, we found, apart from the 2 objects with high redshifts, 2 objects in both catalogues.

5.5.2 Lensed Quasars

By cross-matching our test set with the Oguri lensed quasar database, we found just one object out of the 26. The SDSS object ID is J1251+2935 and it has been classified as an outlier by our model five out of ten times and not considered an outlier in our *Average* experiment with a δz_{norm} of 0.13995. The object falls inside our standard deviation cut with a standard deviation of 0.07131 and has an average distance from the train set of 0.66089. It seems to have a good photometry, but seems to have a borderline behaviour, being "almost" an outlier.

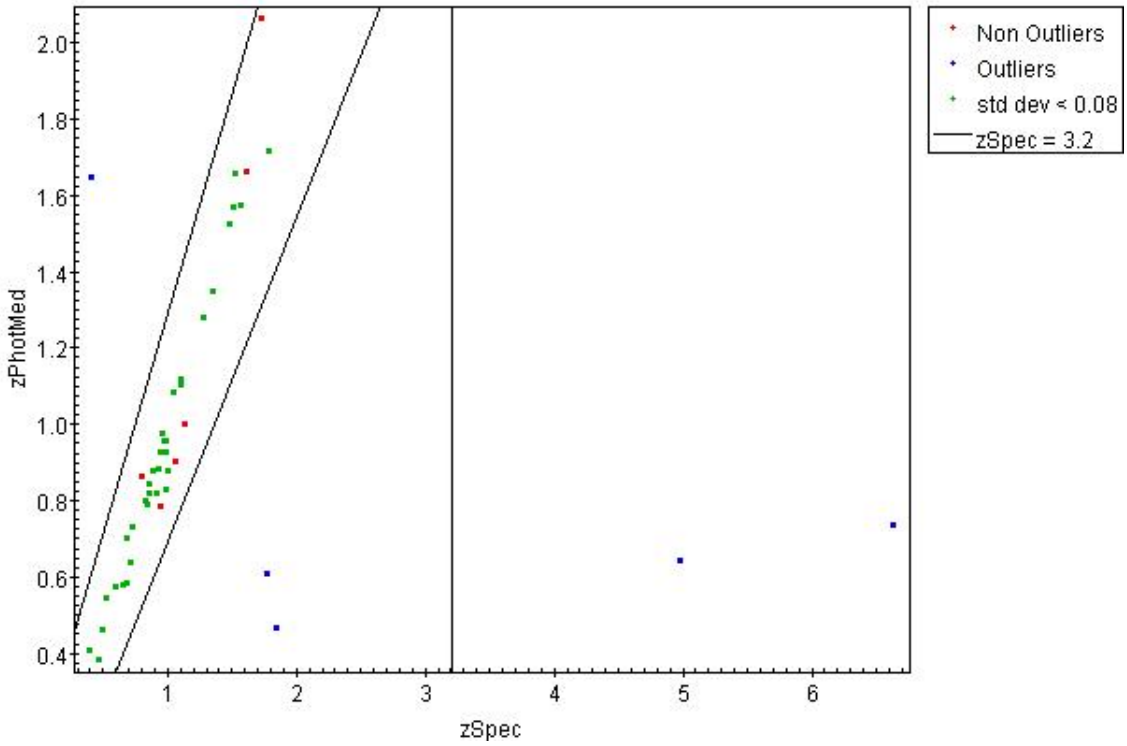

Figure 5.5: zSpec vs zPhotMed scatter plot of the 45 Blazars. The green ones are the 34 retained after the standard deviation cut

5.6 Direct observation of the outliers

Of the 5 outliers that are always classified as such by our model, even after the standard deviation cut, 4 presented a spectroscopic redshift higher than 3.2 and a signal to noise ration in the W4 band less then 2 or, in other words, two very good explanations for why they are classified as outliers by our model; these are the SDSS identifiers of the four objects:
SDSS J084256.69+281340.6
SDSS J161143.00+281543.6
SDSS J024918.81-002752.0
SDSS J145542.93+010726.7
They do not seem to have any further anomaly apart from those already listed by us.

The fifth object (SDSS J150638.20+034702.6) does not present, on direct observation, any peculiar characteristic, nor from the physical point of view nor from the photometric quality point of view. It's fifteen-dimensional distance is consistent with the non outliers average.
For now, we mark this object as a catastrophic outliers and we leave further investigations to be done in future works.

5.7 Conclusions

In this work we analysed the factors influencing the performance of MLPQNA on the photometric redshift evaluation through the analysis of those objects for which the model could not derive a good prediction: the catastrophic outliers.
We found that the outliers are mostly objects that present some problems within their photometry, especially low signal to noise ratios and/or the absence of a single peak (poor deblending or poor detection)in one of the bands. A small percentage of the outliers has relatively good photometry, but turn out to be intrinsically peculiar (blazars, lensed quasars).

We found that the distances of test objects from train objects in the $15-D$ parameter space seem to have a good predicting potential in understanding if a point is a potential outlier or not. This could mean that a distribution of the objects based on their relative distances instead of a random shuffle could improve the network performance.

Also we found, playing with the k-fold value and the max number of iterations, that higher values of both, boosted the performance of the model (we use $k = 10$ and $MaxIt = 8000$). In future applications one could therefore try to repeat the training of the model first with a higher number of iterations ($15,000$ could be a good starting point) and, then, to see if an increase in k could improve the model performance.

It has also been shown that another way to increase performances is by performing several experiments in order to evaluate an average output ($zPhotMed$). While, on the one hand this comes at the price of a large increase in computing time, on the other, it allowed us to perform a study of the standard deviation and to identify a strategy to flag out possible outliers also in asbsence of spectroscopic information.

Appendix A

Features and Flags

Here we present a list of all the features extracted from our SQL queries and all the flags used in this work for the data pruning process

SDSS

S1 **specObjID**: unique spectroscopic object ID.

S2 **ra**: object right ascension expressed in arcsecs

S3 **dec**: object declination expressed in arcsecs

S4 **z**: object spectroscopic redshift (also named in this work zSpec)

S5 **zErr**: spectroscopic redshift error

S6 **class**: spectroscopic class (GALAXY, QSO or STAR)

S7 **subClass**: spectroscopic subclass

S8 **zWarning**: a bitwise flag that can assume values between 0 and 8

1. no known problem
2. sky fiber
3. too little wavelength coverage (WCOVERAGE < 0.18)
4. chi-squared of best fit is too close to that of second best (< 0.01 in reduced chi-squared)
5. synthetic spectrum is negative (only set for stars and QSOs)
6. fraction of points more than 5 sigma away from best model is too large (> 0.05). That usually indicates a high signal-to-noise spectrum or broad emission lines in a galaxy
7. chi-squared minimum at edge of the redshift fitting range (Z_ERR set to -1)
8. a QSO line exhibits negative emission, triggered only in QSO spectra, if C_IV, C_III, Mg_II, H_beta, or H_alpha has $LINEAREA + 3 \times LINEAREA\_ERR < 0$

S9 **sciencePrimary**: best version of the spectrum at this location (it can assume two values: 1 if it is the best spectrum at the location, 0 if it is the opposite)

S10 **objID**: unique SDSS photometric object identifier

S11 **type**: type classification of an object, this variable can assume values between 0 and 9:

1. UNKNOWN
2. COSMIC RAY
3. DEFECT (the object is produced by a defect in the telescope or processing pipeline)
4. GALAXY
5. GHOST (Object created by reflected or refracted light)
6. KNOWNOBJ (Object is listed in a catalogue that is not the SDSS)
7. STAR
8. TRAIL (satellite or asteroid trail)
9. SKY (no object in this arcsec area)
10. NOTATYPE

S12 **flags**: Photo object attribute flags, in the CAS they are 64 flags combined in a single 64-bit integer. For a precise description of all 64 flags you can visit the SDSS DR9 schema browser website ad this url `http://skyserver.sdss.org/dr9/en/help/browser/enum.asp?n=PhotoFlags`. The only flag that we insert here is DEBLEND_NOPEAK that we used in the Photometric Quality chapter to characterize outliers.

- **DEBLEND_NOPEAK**: There was no detected peak within this child in at least one band.

S13 **petroMag**: For galaxy photometry, measuring flux is more difficult than for stars, because galaxies do not all have the same radial surface brightness profile, and have no sharp edges. In order to avoid biases, we wish to measure a constant fraction of the total light, independent of the position and distance of the object. To satisfy these requirements, the SDSS has adopted a modified form of the Petrosian (1976) system, measuring galaxy fluxes within a circular aperture whose radius is defined by the shape of the azimuthally averaged light profile.
We define the "Petrosian ratio" R_P at a radius r from the center of an object, to be the ratio of the local surface brightness in an annulus at r to the mean surface brightness within r, as described by Blanton et

al. 2001a, [3], and Yasuda et al. 2001:

$$R_P(r) = \frac{\int_{0.8r}^{1.25r} dr' 2\pi r' I(r') / [\pi(1.25^2 - 0.8^2) r^2]}{\int_0^r dr' 2\pi r' I(r') / (\pi r^2)}$$

where I(r) is the azimuthally averaged surface brightness profile.
The Petrosian radius r_P is defined as the radius at which $R_P(r_P)$ equals some specified value $R_{P,lim}$, set to 0.2 in our case. The Petrosian flux in any band is then defined as the flux within a certain number N_P (equal to 2.0 in our case) of r Petrosian radii:

$$F_P = \int_0^{N_P r_P} 2\pi r' dr' I(r')$$

In the SDSS five-band photometry, the aperture in all bands is set by the profile of the galaxy in the r band alone. This procedure ensures that the color measured by comparing the Petrosian flux F_P in different bands is measured through a consistent aperture.
The aperture $2r_P$ is large enough to contain nearly all of the flux for typical galaxy profiles, but small enough that the sky noise in F_P is small. Thus, even substantial errors in r_P cause only small errors in the Petrosian flux (typical statistical errors near the spectroscopic flux limit of $r \sim 17.7$ are ¡ 5%), although these errors are correlated.

S13.1 **petroMagErr**: error on the petroMag measure

S14 **psfMag**: For isolated stars, which are well-described by the point spread function (PSF), the optimal measure of the total flux is determined by fitting a PSF model to the object. In practice, we do this by sync-shifting the image of a star so that it is exactly centred on a pixel, and then fitting a Gaussian model of the PSF to it. This fit is carried out on the local PSF KL model at each position as well; the difference between the two is then a local aperture correction, which gives a corrected PSF magnitude. Finally, we use bright stars to determine a further aperture correction to a radius of $7.4''$ as a function of

seeing, and apply this to each frame based on its seeing. This involved procedure is necessary to take into account the full variation of the PSF across the field, including the low signal-to-noise ratio wings. Empirically, this reduces the seeing-dependence of the photometry to below 0.02 mag for seeing as poor as $2''$. The resulting magnitude is stored in the quantity psfMag. The flag PSF_FLUX_INTERP warns that the PSF photometry might be suspect. The flag BAD_COUNTS_ERROR warns that, because of interpolated pixels, the error may be underestimated.

S14.1 **psfMagErr**:error on the psfMag measure

S15 **ModelMag**: there are two model magnitudes associated with each catalogue object; devMag associated to a pure de Vancouleurs profile and expMag associated to a pure exponential profile. These two magnitudes are calculated from independent model in each band. ModelMag uses the better of this two fits in the r-band as a matched aperture to calculate the flux in all bands.For extended objects, modelMag usually provides the best available SDSS colors.
A pure de Vaucouleurs profile: $I(r) = I_0 \exp[-7.67(r/r_e)^{1/4}]$
A pure exponential profile: $I(r) = I_0 \exp(-1.68r/r_e)$

S15.1 **ModelMagErr**: error on the ModelMag measure

S16 **cModelMag**: (composite model magnitudes)the code takes the best exponential fit and best de Vancouleurs fits in each band and asks for the linear combination of the two that best fits the image

S16.1 **cModelMagErr**: error on the cModelMag measure

GALEX

G1 **ObjID**: GALEX specific object identification number

G2 **ra**: object rect ascension expressed in arcscs

G3 **dec**: object declination expressed in arcsecs

G4 **fuv_mag**: FUV calibrated magnitude

G5 **fuv_magerr**: error on the fuv_mag measure

G6 **FUV_MAG_ISO**: isophotal magnitude

G7 **FUV_MAGERR_ISO**: error on the isophotal magnitude measure

G8 **FUV_MAG_AUTO**: Kron-like elliptical aperture magnitude

G9 **FUV_MAGERR_AUTO**: error on the Kron-like elliptical aperture magnitude measure

G10 **FUV_MAG_APER_1**: magnitude aperture of 2000 px

G11 **FUV_MAG_APER_2**: magnitude aperture of 3000 px

G12 **FUV_MAG_APER_3**: magnitude aperture of 5000 px

G13 - G15 **FUV_MAGERR_APER_1**: errors on the 3 apertures magnitudes

G16 - G27 **nuv**: all the flags previously listed but for the NUV band

G28 **fuv_artifact**: see nuv_artifact

G29 **nuv_artifac**: this flag indicates if an artifact is present in the image or was removed from it

G30 **fuv_ambg**: fuv ambiguity flag

G31 **nuv_ambg**: nuv ambiguity flag

G32 **fuv_maskpix**: number of masked pixels near source

G33 **nuv_maskpix**: number of masked pixels near source

G34 **fuv_nc_r1**: neighbours count out to radius $R1 = 5.0''$

G35 **nuv_nc_r1**: neighbours count out to radius $R1 = 5.0''$

UKIDSS

X = y, j, h, k

U1 **sourceID**: unique UKIDSS database object ID

U2 **XHallMag**: total point source X mag

U2.1 **XHallMagErr**: error on the XHallMag measure

U3 **XPetroMag**: extended source X mag

U3.1 **XPetroMagErr**: error on the XPetroMag measure

U4 **XPsfMag**: point source X profile-fitted magnitude

U4.1 **XPsfMagErr**: error on the XPsfMag measure

U5 **XAperMag3**: default point/extended source X aperture corrected magnitude (2.0 arcsec aperture diameter)

U5.1 **XAperMag3Err**: error on the AperMag3 measure

U6 **XAperMag4**: default point/extended source X aperture corrected magnitude (2.8 arcsec aperture diameter)

U6.1 **XAperMag4Err**: error on the AperMag4 measure

U7 **XAperMag6**: default point/extended source X aperture corrected magnitude (5.7 arcsec aperture diameter)

U7.1 **XAperMag6Err**: error on the AperMag6 measure

U8 **XErrBits**: count of the number of zero confidence pixels in the default (2 arcsec) aperture

U9 **XppErrBits**: Post-processing error quality bit flags assigned (NB: from UKIDSS DR2 release onwards) in the WSA curation procedure for survey data

WISE

X = 1, 2, 3, 4

W1 **wXmpro**: WX magnitude measured with profile-fitting photometry, or the magnitude of the 95% confidence brightness upper limit, if the WX flux measurement has SNR ¡ 2. This column is *null* if the source is nominally detected in WX, but no useful brightness estimate could be made.

W2 **wXsigmpro**: WX profile-fit photometric measurement uncertainty in mag units. This column is *null* if the WX profile-fit magnitude is a 95% confidence upper limit or if the source is not measurable.

W3 **wXsnr**: WX profile-fit measurement signal-to-noise ratio. This value is the ratio of the flux (wXflux) to flux uncertainty (wXsigflux)in the WX profile-fit photometry measurement. This column is null if wXflux is negative, or if wXflux or wXsigflux are null.

W4 **na**: active deblending flag. Indicates if a single detection was split into multiple sources in the process of profile-fitting

0 : the source is not actively deblended

1 : the source is actively deblended

W5 - W8 **WXsat**: saturated pixel fraction in the image.The fraction of all pixels within the profile-fitting area in the stack of single-exposure images used to characterize this source that are flagged as saturated. A value larger than 0.0 indicates one or more pixels of saturation.

W9 **cc_flags**: contamination and confusion flag. Four character string, one character per band [W1/W2/W3/W4], that indicates that the photometry and/or position measurements of a source may be contaminated or biased due to proximity to an image artefact. The type of artefact that may contaminate the measurements is denoted by the following codes. Lower-case letters correspond to instances in which the source detection in a band is believed to be real but the measurement may be contaminated by the artefact. Upper-case letters are instances in

which the source detection in a band may be a spurious detection of an artefact

D,d : diffraction spike. Source may be a spurious detection of (D) or contaminated by (d) a diffraction spike from a nearby bright star on the same image

P,p : persistence . Source may be a spurious detection of (P) or contaminated by (p) a short-term latent image left by a bright source

H,h : halo. Source may be a spurious detection of (H) or contaminated by (h) the scattered light halo surrounding a nearby bright source

O,o : optical ghost. Source may be a spurious detection of (O) or contaminated by (o) an optical ghost image caused by a nearby bright source

0 : source is unaffected by known artefacts
A source extraction may be affected by more than one type of artefact or condition. In this event, the cc_flags value in each band is set in the following priority order: D,P,H,O,d,p,h,o,0.

W10 **ext_flg**: extended source flag. This is an integer flag, the value of which indicates whether or not the morphology of a source is consistent with the WISE point spread function in any band, or whether the source is associated with or superimposed on a previously known extended object from the 2MASS Extended Source Catalog (XSC). The values of the ext_flg indicate the following conditions:

0 : the source shape is consistent with a point-source and the source is not associated with or superimposed on a 2MASS XSC source

1 : the profile-fit photometry goodness-of-fit, wXrchi2, is > 3.0 in one or more bands

2 : the source falls within the extrapolated isophotal footprint of a 2MASS XSC source

3 : the profile-fit photometry goodness-of-fit, wXrchi2, is > 3.0 in one or more bands, and The source falls within the extrapolated isophotal footprint of a 2MASS XSC source

4 : the source position falls within $5''$ of a 2MASS XSC source

5 : the profile-fit photometry goodness-of-fit, wXrchi2, is > 3.0 in one or more bands, and the source position falls within $5''$ of a 2MASS XSC source

W11 **ph_equal**: photometric quality flag. Four character flag, one character per band [W1/W2/W3/W4], that provides a shorthand summary of the quality of the profile-fit photometry measurement in each band, as derived from the measurement signal-to-noise ratio.

A : source is detected in the band with a flux-signal-to-noise ratio $WXsnr \geq 10$

B : source is detected in the band with a flux-signal-to-noise ratio $3 \leq WXsnr \leq 10$

C : source is detected in the band with a flux-signal-to-noise ratio $2 \leq WXsnr \leq 3$

U : source is detected in the band with a flux-signal-to-noise ratio $WXsnr \leq 2$

X : a profile-fit measurement was not possible at this location in this band. The value of wXmpro and wXsigmpro will be "null" in this band

Z : a profile-fit source flux measurement was made at this location, but the flux uncertainty could not be measured. The value of wXsigmpro will be "null" in this band. The value of wXmpro will be "null" if the measured flux, wXflux, is negative, but will not be "null" if the flux is positive.

Bibliography

[1] Antonucci, R, *Unified models for active galactic nuclei and quasar*, Annual review of astronomy and astrophysics, Vol. 31, 1993.

[2] Bishop, CM, *Pattern Recognition and Machine Learning*, 2006, Springer.

[3] Blanton, MR, Dalcanton, J, Eisenstein, D, et al., *The Luminosity Function of Galaxies in SDSS Commissioning DataBased on observations obtained with the Sloan Digital Sky Survey*, The Astronomical Journal, Vol. 121, n. 5, 2001.

[4] Ilbert, O, Arnouts, S, McCracken, HJ, et al., *Accurate photometric redshifts for the CFHT legacy survey calibrated using the VIMOS VLT deep survey*, Astronomy & Astrophysics, Vol. 457, n. 3, 2006.

[5] Bolzonella, M, Miralles, JM, Pello, R *Photometric redshifts based on standard SED fitting procedures*, Astronomy & Astrophysics, Vol. 363, 2000.

[6] Brescia, M, Cavuoti, S, D'Abrusco, R, et al., *Photometric redshifts for quasars in multi-band surveys*, The Astrophysical Journal, Vol. 772 n.2, 2013.

[7] Brescia, M, Cavuoti, S, Longo, G, et al., *DAMEWARE: A web cyberinfrastructure for astrophysical data mining*, Publications of the Astronomical Society of the Pacific, Vol. 126, n. 942, 2012.

[8] Keams, M, *A bound on the error of cross validation using the approximation and estimation rates, with consequences for the training-test split*, Advances in Neural Information Processing Systems 8: Proceedings of the 1995 Conference, MIT Press, Vol. 8, 1996.

[9] McCulloch, WS, and Pitts, W, *A logical calculus of the ideas immanent in nervous activity*, The bulletin of mathematical biophysics, Vol. 5, n. 4, 1943, Springer.

[10] Minsky, M, and Papert, S, *Perceptron*, MIT press, 1969.

[11] Mosteller, F, and Tukey, JW, *Data analysis, including statistics*, 1968.

[12] Oguri, M, Inada, N, Strauss, MA, et al., *The Sloan Digital Sky Survey Quasar Lens Search. VI. Constraints on Dark Energy and the Evolution of Massive Galaxies*, The Astronomical Journal, Vol. 143, n. 5, 2012, IOP Publishing.

[13] Ahn, CP, Alexandroff, R, Prieto, CA, et al., *The ninth data release of the Sloan Digital Sky Survey: first spectroscopic data from the SDSS-III Baryon Oscillation Spectroscopic Survey*, The Astrophysical Journal Supplement Series, Vol. 203, n. 2, 2012.

[14] Martin, DC, Fanson, J, Schiminovich, D, et al., *The Galaxy Evolution Explorer: A space ultraviolet survey mission*, The Astrophysical Journal Letters, Vol. 619, n. 1, 2005, IOP Publishing.

[15] Lawrence, A, Warren, SJ, Almaini, O, et al., *The UKIRT infrared deep sky survey (UKIDSS)*, Monthly Notices of the Royal Astronomical Society, Vol. 379, n. 4, 2007, Oxford University Press.

[16] Wright, EL, Eisenhardt, PRM, Mainzer, AK, et al., *The Wide-field Infrared Survey Explorer (WISE): mission description and initial on-orbit performance*, The Astronomical Journal, Vol. 140, n. 6, 2010, IOP Publishing.

[17] Rosenblatt, F *The perceptron: a probabilistic model for information storage and organization in the brain*, Psychological review, Vol. 65, n. 5, 1958, American Psychological Association.

[18] Chauvin, Y and Rumelhart, DE, *Backpropagation: theory, architectures, and applications*, Psychology Press, 1995.

[19] Shanno, DF, *Recent advances in numerical techniques for large-scale optimization*, MIT Press, Cambridge, MA, 1990.

[20] Seyfert, CK, *Nuclear Emission in Spiral Nebulae*, The Astrophysical Journal, Vol. 97, 1943.

[21] Urry, CM and Padovani, Paolo, *Unified schemes for radio-loud active galactic nuclei*, Publications of the Astronomical Society of the Pacific, 1995, JSTOR.

[22] Werbos, Paul, *Beyond regression: New tools for prediction and analysis in the behavioural sciences*, 1974.