Università degli Studi di Napoli
"Federico II"
Facoltà di Scienze MM. FF. NN.

Tesi di Laurea Magistrale in Fisica

Anno Accademico 2012/2013

PHOTORAPTOR
PHOTOmetric Research APplication TO Redshifts and application to the SDSS-DR9 galaxies

Relatori:

Ch.mo Prof. Giuseppe Longo

Dott. Massimo Brescia

Candidato:

Virgilio De Stefano

Matricola N94/147

Contents

List of Figures

List of Tables

Introduction

Astronomy has a long history of acquiring, systematizing and interpreting large quantities of data. Starting from the earliest sky maps through the first major photographic sky surveys of the 20-th century, the acquisition is continuing today at an ever increasing rate. Thanks to the advances in telescope, detector and computer technology, today astronomers can map the universe systematically, and in a panchromatic manner. This will enable new science items, from statistical studies of our Galaxy and of the large-scale structure in the universe, to the discoveries of rare or even completely new types of astronomical objects and phenomena.

Indeed, astronomers can now explore all regions of the electromagnetic spectrum, from gamma rays up to radio wavelengths. Besides, computational advances have enabled detailed physical simulations similars to the largest observational datasets in terms of complexity. In order to investigate our cosmos, we need to assimilate all of this data, each presenting its own physical view of the Universe and requiring its own technology.

Astronomical data and its subsequent analysis can be broadly classified into five domains.

- **Imaging data** is the fundamental constituent of astronomical observations, capturing a two-dimensional spatial picture of the Universe within a narrow wavelength region at a particular instant of time. Astronomical images can be acquired directly, i.e. with imaging arrays such as CCDs, or synthesized from interferometric observations, as is done in radio astronomy. Since different physical processes emit radiation at different wavelengths, most astronomical images are obtained through specific filters, depending on the primary purpose of the observations and the type of recording device.

- **Catalogs** are generated by processing the imaging data. Each detected source can have a large number of measured parameters, for example coordinates, flux quantities and morphological information.

- **Spectroscopy, Polarization** and other measurements provide detailed physical quantification of the systems, including information on the distance (redshift), chemical composition (abundances of heavier elements compared to hydrogen) and measurements of the physical fields (electromagnetic or gravitational) in the source.
- **Studying the time domain** provides important information about the nature of the Universe because it allows to identify moving objects (near-Earth objects or comets), variable sources (e.g. pulsating stars) or transient objects (supernovae and gamma ray bursts). They require multiple epoch observations of fields (which is possible in the overlap regions of surveys) or a dedicated synoptic survey. In either case, the data volume and thus the difficulty in handling and analyzing them increases significantly.
- **Numerical Simulations** are theoretical tools which must be compared with observational data. Examples include simulations of the formation and evolution of large scale structure in the Universe, star formation in our Galaxy, supernova explosions, etc. Many of the physical processes that are involved in these studies are very complex, so this tools use both direct analytic solutions and numerical analysis.

Most of data are obtained in the form of images; the sensor output is processed removing instrumental signatures and performing calibrations. These first order data are then stored in local archives as raw data. Then they are processed to construct the catalogue related to all the object observed in the running. These informations will be stored in archives and freely accessible online.

In their motion the objects tends to modify their spectral features; studying the difference between the colors of the spectrum over time, astronomers can study theri evolution. This difference is usually known as **redshift**, a shift in the lines of the spectrum of an astronomical object towards a longer wavelength (the red end of an optical spectrum): this is usually due to the Doppler effect caused by the object's movement away from the viewer.

Usually the redshift is measured spectroscopically: emission or absorption lines are identified and their wavelengths are measured. The measured wavelengths are then compared with the rest wavelengths to determine the redshift. In spectroscopy, the light from the galaxy is separated into narrow wavelength bins with few angstroms width, so each bin receives only a small fraction of the total light from the galaxy. To achieve a sufficiently high signal-to-noise ratio in each bin, very long integration times are required. For photometry, the bins are much larger requiring a short exposure time to reach the same signal-to-noise ratio. In addition, imaging detectors generally cover a greater area of the sky

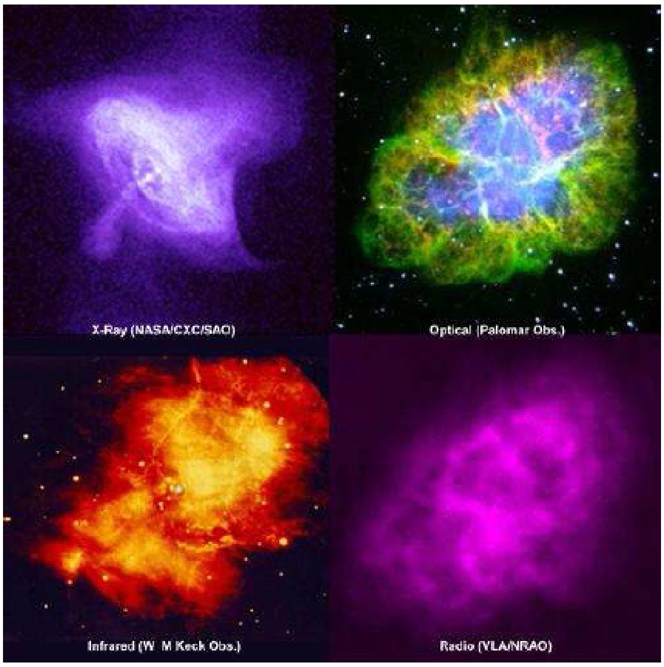


Figure 1: A multiwavelength view of the Crab nebula, a supernova remnant that was first sighted by Chinese astronomers in 1054 AD. Image Credit: NASA/CXC/SAO.

than multi-object spectrographs Connolly et al. (1997).

As result, photometric redshifts can be measured much faster and in larger quantities than their spectroscopic counterparts.
Photometric redshift was a viable technique in the 1960s, but it was largely replaced in the 1970s and 1980s by spectroscopic redshifts. This technique recently have experienced a burst of interest because deep multicolour photometric surveys have been carried out, with a large number of objects inaccessible to spectroscopic observations.
Photometric redshifts were originally determined by calculating the expected observed data from a known emission spectrum at a range of redshifts. In recent years, Bayesian statistical methods and artificial neural networks have been used to estimate redshifts from photometric data.

In the last decades data volumes from multiple sky surveys have grown up to terabytes, and will grow up to tens (or hundreds) of petabytes in the next decade. This exponential growth of data both enables and challenges effective astronomical research, requiring new approaches. So advance in information technology have been profoundly involved in the last science researches, also to store and analyze the huge amount of simulated data necessary to the development of new

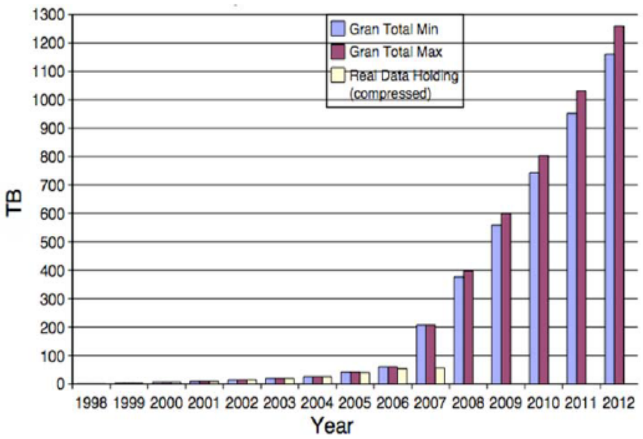


Figure 2: The Data Gap: Data growth in the ESO case, credit of ESO.

models.
According to the Moore's Law, computing power doubles every 18 months, corresponding to a factor of 100 in ten years. By comparison, data volumes appear to double every year, giving a factor 1000 in ten years: this makes very difficult for us to access and analyze our data collections. In astronomy in particular, advances in three technology areas (telescopes, detectors and computation) have continued unabated, leading to more and more data.

The trend in Fig. 2 shows how much a typical astronomical archive has increases in size over the last thirty years; such exponential growth is not matched by an equivalent increase in the number of data analyst and already now data analysis requirements have largely exceeded the power of dedicated human brains.

In 2009 Tony Hey analysed the problem of data in the book "The Fourth Paradigm" Hey et al. (2009) saying that data analysis needs to be considered the fourth independent methodological pillar of modern science after experiment, theory and simulation. Indeed, data taken for a specific purpose can be re-used, allowing the possibility of developing new sciences. As for examples, the time variability of phenomena or the comparison of phenomena in different energy bands (multi-wavelength astronomy, see Fig. 1).

To make the best use of all available data, new generations of astronomers will need to know more about data fusion, virtual working environments, web 2.0 technologies, machine learning and data mining, disciplines belonging to the emerging field of **Astroinformatics** Borne (2009): a new discipline between traditional astronomy, applied mathematics, computer science and Informatic and Communication Technology (ICT).

The topic of this thesis is to calculate the photometric redshift for an huge dataset extracted from the SDSS Data Release 9. The evaluation has been done using an empirical method based on artificial neural networks: the MLPQNA, originally created for the integration in DAMEWARE tools (DAta Mining & Exploration Web Application REsource) Cavuoti et al. (2012b).
In addition I have realized a Java desktop application that calculates the redshift for all the objects in an user's dataset and that allows to load, edit and display data in most common file format.

The application of the method presented in this work, has produced a refereed article, currently under revision by the editor Brescia et al. (2013b). Furthermore, there is also another manuscript in preparation, whose topic is the PhotoRApToR application [De Stefano et al. 2013, in preparation].

The present work is structured as follow: in Chapter 1, I describe the importance of Photometric Redshift estimation and the ways to obtain this information; in Chapter 2, I show the procedure followed to extract my own dataset from SDSS DR9 archive; in Chapter 3, I give a short overview of data mining and the problem of the Massive Data Sets. I also describe MLPQNA: the machine learning method used in my Java application; in Chapter 4, I present the Java Desktop Application (PhotoRApToR 1.0) that I realized. The last Chapter describes the application of PhotoRApToR to a real science case and shows the results. Finally in section 5.3 I show some conclusions and discuss some future developments.

Chapter 1

Photometric Redshift

According to the different excitation levels, each element in the periodic table emits photons only at certain wavelengths. This is reported as either emission or absorption lines in the spectrum of the astronomical objects; measuring the position of these spectral lines, we can determine which elements are present in the object itself or are interposed along the line of sight.

But with this analysis, the astronomers note that, for almost all extragalactic objects, the observed spectral lines are shifted to longer (redder) wavelengths: this phenomenon is known as "redshift" and, in general, also if the radiation is not within the visible spectrum, "redder" means an increase in wavelength, that corresponds to a lower frequency and photon energy.

There are two distinct causes for the spectral shift of the light emitted (or absorbed) by a galaxy: the kinematical Doppler effect of special relativity (SR) and the redshift caused by the expansion of the universe, governed by general relativity (GR). These two effects cannot be distinguished from one another by observing the spectrum of the galaxy or other light source.

The **Doppler shift** of SR is due to the relative velocity between the source and the observer and can be negative (blueshift) or positive (redshift), depending on whether the galaxy moves radially toward or away from us. It consists in the variation of frequency ν (or wavelength $\lambda = c/\nu$, where c is the light speed) of an electromagnetic wave (see Fig. 1.1). The first Doppler redshift was described by French physicist Hippolyte Fizeau in 1848, who pointed to the shift in spectral lines seen in stars as being due to the Doppler effect.

The redshift z is defined as

$$z \equiv \frac{\Delta\lambda}{\lambda} = \frac{\lambda_{obs} - \lambda_{rest}}{\lambda_{rest}} \tag{1.1}$$

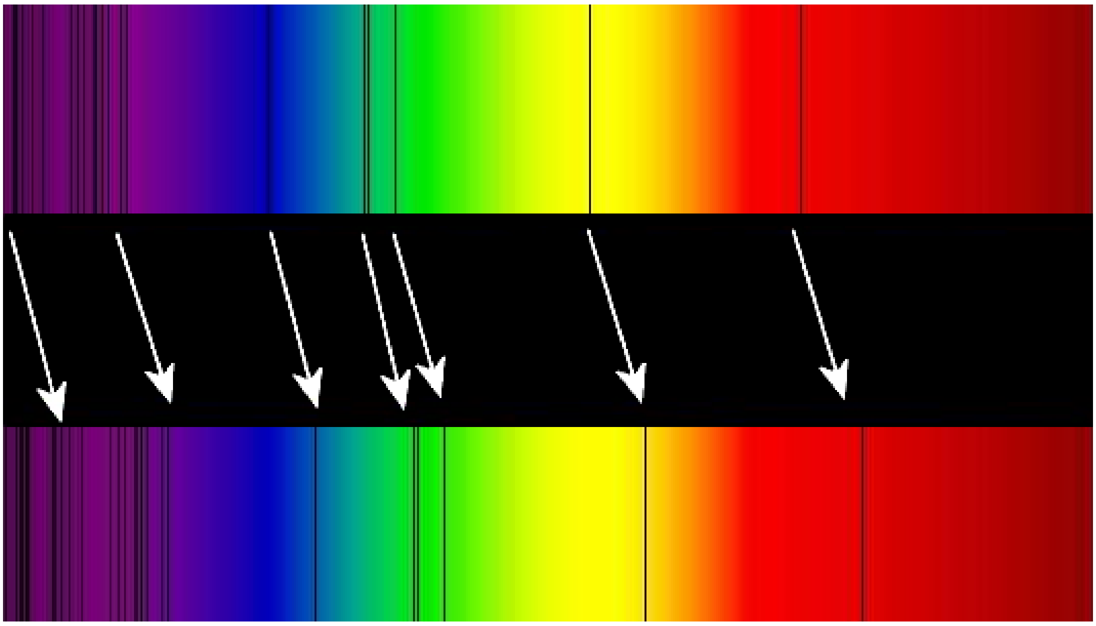


Figure 1.1: The dark absorption lines of a star at rest (up) get shifted towards red if the star is moving away from Earth (bottom). Credit: Wikipedia

where λ_{obs} is the observed wavelength and λ_{rest} is the emitted/absorbed wavelength; for small velocities ($v \ll c$):

$$z = \frac{v}{c} \tag{1.2}$$

where v is the radial velocity between the source and the obsrver and c is the speed of light. At larger distances the velocity increases and the theory of special relativity must be taken into account.

Assume the observer and the source are moving away from each other with a relative velocity v. Considering the problem in the reference frame of the source, let us suppose that one wavefront arrives at the observer. The next wavefront is then at a distance $\lambda_{rest} = c/\nu_{rest}$, away from him ($\nu_{rest}$ is the frequency of the wave the source emitted). Since the wavefront moves with velocity c, and the observer escapes with velocity v, the time between crest arrivals at the observer is

$$t = \frac{\lambda}{c-v} = \frac{c}{(c-v)\nu_{rest}} = \frac{1}{(1-\beta)\nu_{rest}}, \tag{1.3}$$

where $\beta = v/c$, is the velocity of the observer in terms of the speed of light. Due to the relativistic time dilation, the observer will measure this time to be

$$\begin{aligned} t_{obs} &= \frac{t}{\gamma}, \text{ where} \\ \gamma &= \frac{1}{\sqrt{1-\beta^2}} \end{aligned}$$

is the *Lorentz factor*. The corresponding observed frequency is

$$\nu_{obs} = \frac{1}{t_{obs}} = \gamma(1-\beta)\nu_{rest} = \sqrt{\frac{1-\beta}{1+\beta}}\nu_{rest}. \tag{1.4}$$

So it is possible to calculate the ratio

$$\begin{aligned} \frac{\nu_{rest}}{\nu_{obs}} &= \sqrt{\frac{1+\beta}{1-\beta}} \\ \frac{\nu_{rest}}{\nu_{obs}} &= \frac{\lambda_{obs}}{\lambda_{rest}} = \sqrt{\frac{1+\beta}{1-\beta}} \end{aligned}$$

So the standard special relativistic expression for the Doppler shift is

$$\begin{aligned} 1+z = \frac{\lambda_{obs}}{\lambda_{rest}} &= \sqrt{\frac{1+\beta}{1-\beta}} = \\ &= \sqrt{\frac{1+\frac{v}{c}}{1-\frac{v}{c}}} \end{aligned} \tag{1.5}$$

and finally

$$z = \sqrt{\frac{c+v}{c-v}} - 1 \tag{1.6}$$

The Eq. 1.5 can be inverted to give the relative velocity as a function of z:

$$\frac{v(z)}{c} = \frac{2z+z^2}{2+2z+z^2} \tag{1.7}$$

By expanding Eq. 1.7 in a Taylor series around $z = 0$, we obtain

$$\frac{v}{c} = z - \frac{z^2}{2} + \frac{z^4}{4}... \tag{1.8}$$

justifying the approximation $v \approx cz$ for small z.

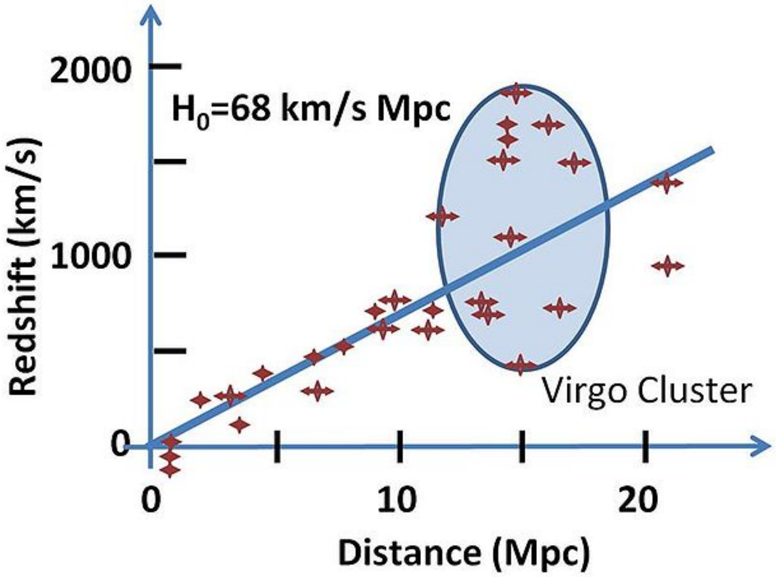


Figure 1.2: Edwin Hubble found a correlation between distance to a galaxy (horizontal axis) and how quickly it's moving away from Earth (vertical axis). The movement of galaxies in a nearby cluster adds some noise to this plot. Credit: William C. Keel (via Wikipedia)

In the early part of the twentieth century, Slipher, Hubble and others made the first measurements of the "red" and "blue" shifts of galaxies beyond the Milky Way. They interpreted the redshifts and blueshifts as solely due to the Doppler effect, but later first Lundmark and then Hubble (see Fig. 1.2) discovered a rough correlation between the redshifts and the distance of galaxies.

Theorists almost immediately realized that these observations could be explained by a different mechanism causing redshifts (the **Cosmological redshift**) and constructed a cosmological model starting from Hubble's correlation law between redshifts and distances.

Outside the nearby Universe redshifts or apparent radial velocities are dominated by the cosmological expansion. This expansion is properly described as the stretching of the metric.

In the standard mathematical description of cosmology, the *Friedmann-Lemaitre-Robertson- Walker model*, distances are defined in terms of the *Robertson-Walker metric*, which is the most general mathematical description for a uniform, homogeneous space that is expanding or contracting.

The RobertsonWalker line element is given by

$$ds^2 = c^2dt^2 - R^2(t)\left[\frac{dr^2}{1-kr^2} + r^2(d\theta^2 + \sin^2\theta d\phi^2)\right] \tag{1.9}$$

for spherical coordinate r, θ, ϕ and time coordinate ct. k is the curvature constant, being -1, 1, or 0 for an open, closed or flat space geometry and $R(t)$ is the *scale factor*, a function of time which represents the relative expansion of the universe: it increases as the universe expands in a manner that depends upon the cosmological model selected. Its meaning is that all measured distances $D(t)$ between co-moving points increase proportionally to R

$$d(t) = R(t)d_0$$

where $d(t)$ is the proper distance at epoch t, d_0 is the distance at the reference time t_0. Thus, by definition, $R(t_0) = 1$.
Considering an electromagnetic wave moving toward the observer along the radius r (so $\theta = 0$ and $\phi = 0$), in this case $ds^2 = 0$ and Eq. 1.9 becomes

$$c^2dt^2 = R^2(t)\frac{dr^2}{1-kr^2} \tag{1.10}$$

and

$$\frac{cdt}{R(t)} = -\frac{1}{\sqrt{1-kr^2}}dr \tag{1.11}$$

because the wave moves toward the origin of coordinate system. Now consider two waves separated in time by Δt, integrating the Eq. 1.11

$$\begin{array}{ll} wave1 & \displaystyle\int_{t_1}^{t_0} \frac{cdt}{R(t)} = -\int_0^r \frac{1}{\sqrt{1-kr^2}}dr \\ wave2 & \displaystyle\int_{t_1+\Delta t_1}^{t_0+\Delta t_0} \frac{cdt}{R(t)} = -\int_0^r \frac{1}{\sqrt{1-kr^2}}dr \end{array}$$

where t_1 and t_0 are the times of emission from the source and arrival to the observer.
Subtracting one from the other

$$\int_{t_1+\Delta t_1}^{t_0+\Delta t_0} \frac{cdt}{R(t)} - \int_{t_1}^{t_0} \frac{cdt}{R(t)} = 0 \tag{1.12}$$

The first integral becomes

$$\int_{t_1+\Delta t_1}^{t_0+\Delta t_0} = \int_{t_1}^{t_0} + \int_{t_0}^{t_0+\Delta t_0} - \int_{t_1}^{t_1+\Delta t_1} \tag{1.13}$$

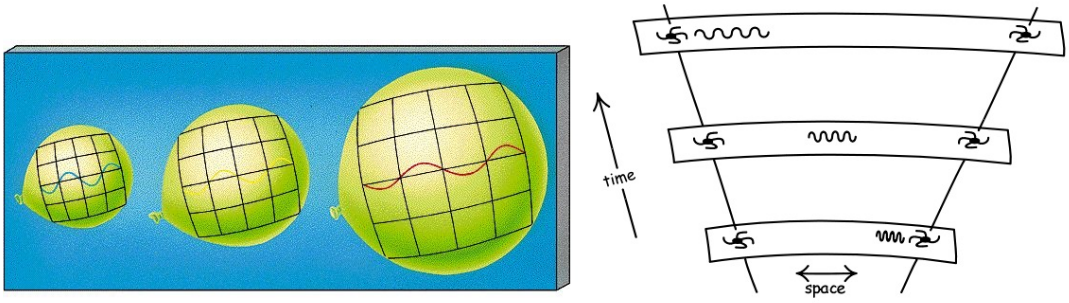


Figure 1.3: Cosmological Redshift. A distant galaxy emits light towards us. The light waves with their crests are carried by space towards us. For a distant galaxy, it can take a very long time for the light to reach us. During that time, the cosmic expansion of space proceeds. The effect is that the waves of the light signal get stretched with space. So the wavelength of the light increases and its frequency decreases. It becomes red shifted.

So we get

$$\begin{aligned}
\int_{t_1}^{t_1+\Delta t_1} \frac{cdt}{R(t)} &= \int_{t_0}^{t_0+\Delta t_0} \frac{cdt}{R(t)} \\
\int_{0}^{\Delta t_1} \frac{cdt}{R(t)} &= \int_{0}^{\Delta t_0} \frac{cdt}{R(t)} \\
\frac{\Delta t_1}{R(t_1)} &= \frac{\Delta t_0}{R(t_0)} \\
\frac{\Delta t_1}{\Delta t_0} &= \frac{R(t_1)}{R(t_0)} \qquad (1.14)
\end{aligned}$$

Remembering that $\lambda = c\Delta t$ the redshift relation (Eq. 1.1) can be written as

$$z = \frac{\lambda_0 - \lambda_1}{\lambda_1} = \frac{\Delta t_0 - \Delta t_1}{\Delta t_1} \qquad (1.15)$$

and finally,

$$z = \frac{R(t_0)}{R(t_1)} - 1 \qquad (1.16)$$

Therefore, redshift is related to the expansion factor of the Universe. If we measure a redshift of $z = 2$, the Universe is 3x bigger now than it was when that photon was emitted because of the variation of the scale factor.

Although cosmological redshift at first appears to be a similar effect to the more familiar Doppler shift, there is a distinction. In Doppler shift, the wavelength of the emitted radiation depends on the motion of the object at the instant the photons are emitted. If the object is travelling towards us, the wavelength is shifted

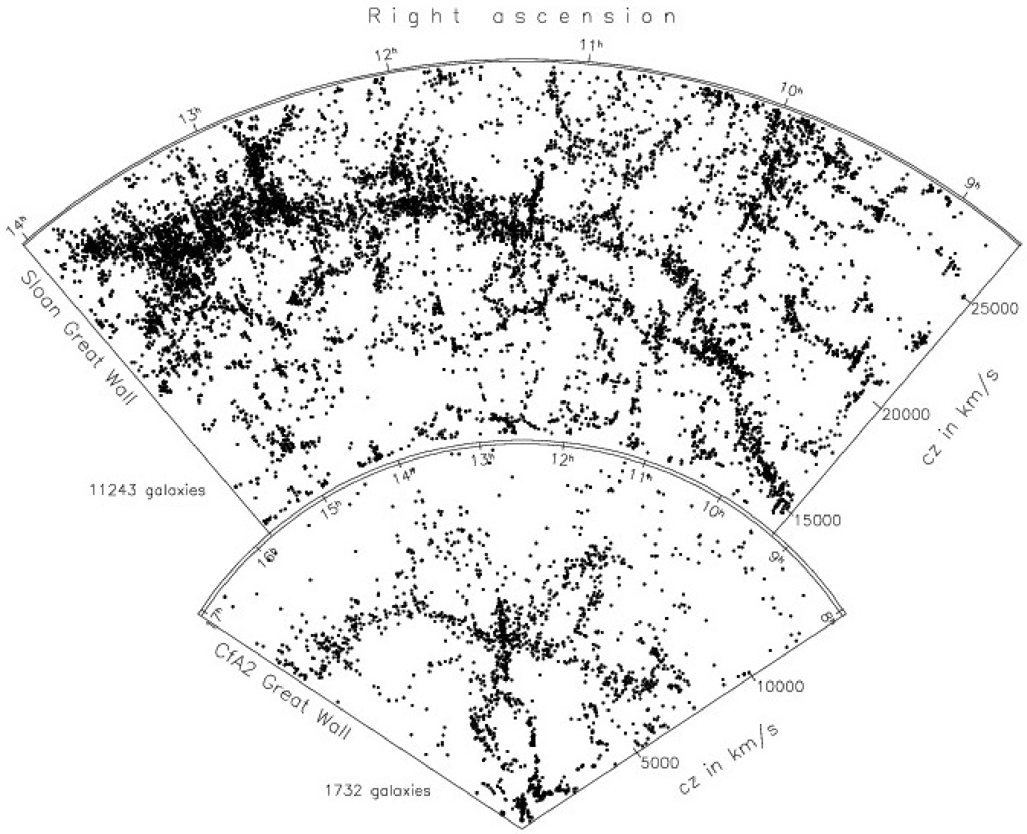


Figure 1.4: The same slice of the universe from CfA2 redshift survey (down) and the SDSS data (up). The famous Great Wall and the SLOAN Great Wall are clearly visible. The "wall" is more than 200 million light-years away, and stretches across roughly 600 million light-years.

towards the blue, while if the object is travelling away from us, the wavelength is shifted towards the red. Cosmological redshift results from the expansion of space itself and not from the motion of any individual body, so the wavelength at which the radiation is originally emitted is lengthened as it travels through (expanding) space (Fig. 1.3).

For example, in an extragalactic binary system it is theoretically possible to measure both a Doppler shift and a cosmological redshift. The Doppler shift would be determined by the motions of the individual stars in the binary system, in their approaching or receding at the time the photons were emitted. The cosmological redshift would be determined by how far away the system was when the photons were emitted. Increasing the distance to the system, the emitted photons have travelled longer through expanding space and higher is the measured cosmological redshift.

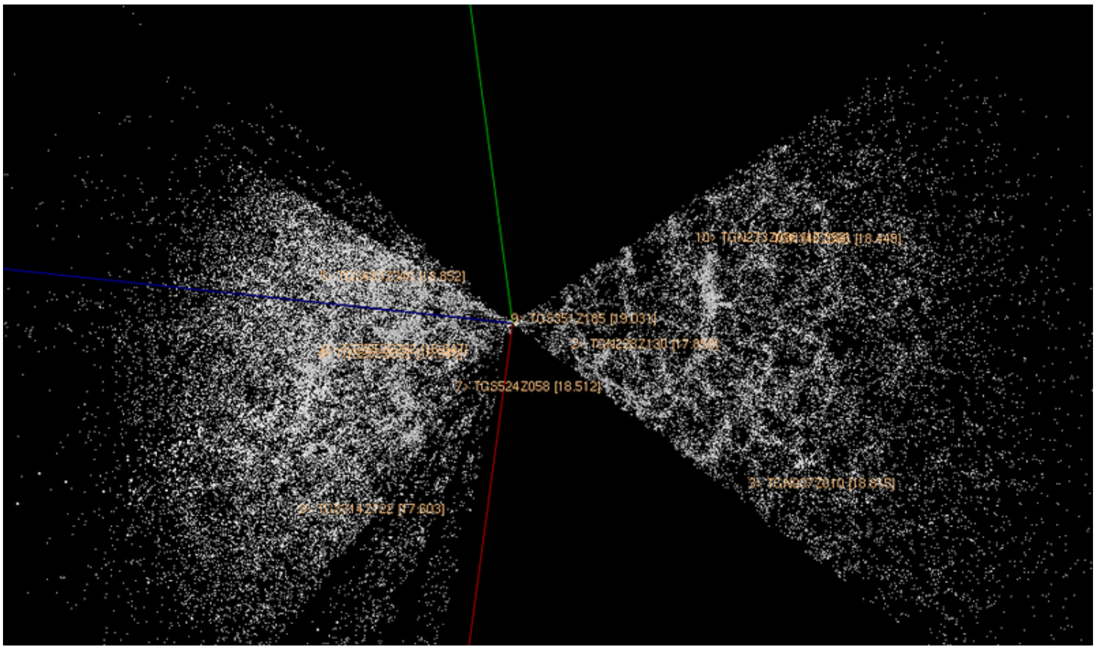


Figure 1.5: Rendering of the 2dF Galaxy Redshift Survey data.

With advent of automated telescopes and improvements in spectroscopes, a number of experiments have been made to map the universe in redshift space: by combining redshift with angular position data has obtained a 3D distribution of matter. These observations are used to measure the properties of the large-scale structure of the universe. The Great Wall (Fig. 1.4), a vast supercluster of galaxies over 500 million light-years wide, provides a dramatic example of a large-scale structure that redshift surveys can detect. The first redshift survey was the CfA Redshift Survey, started in 1977 with the initial data collection completed in 1982. More recently, the 2dF Galaxy Redshift Survey (Fig. 1.5) determined the large-scale structure of one section of the Universe, measuring redshifts for over 220000 galaxies; data collection was completed in 2002 and the final data set was released 30 June 2003. The Sloan Digital Sky Survey (SDSS) is still ongoing and aims to measure the redshifts of around 3 million objects. SDSS has recorded redshifts for galaxies up to $z = 0.8$, and has been involved in the detection of quasars beyond $z = 6$.

1.1 The need for photo-z

When only photometric observations are available, the problem becomes more complicated due to the lack of spectral features. In this case, because of the

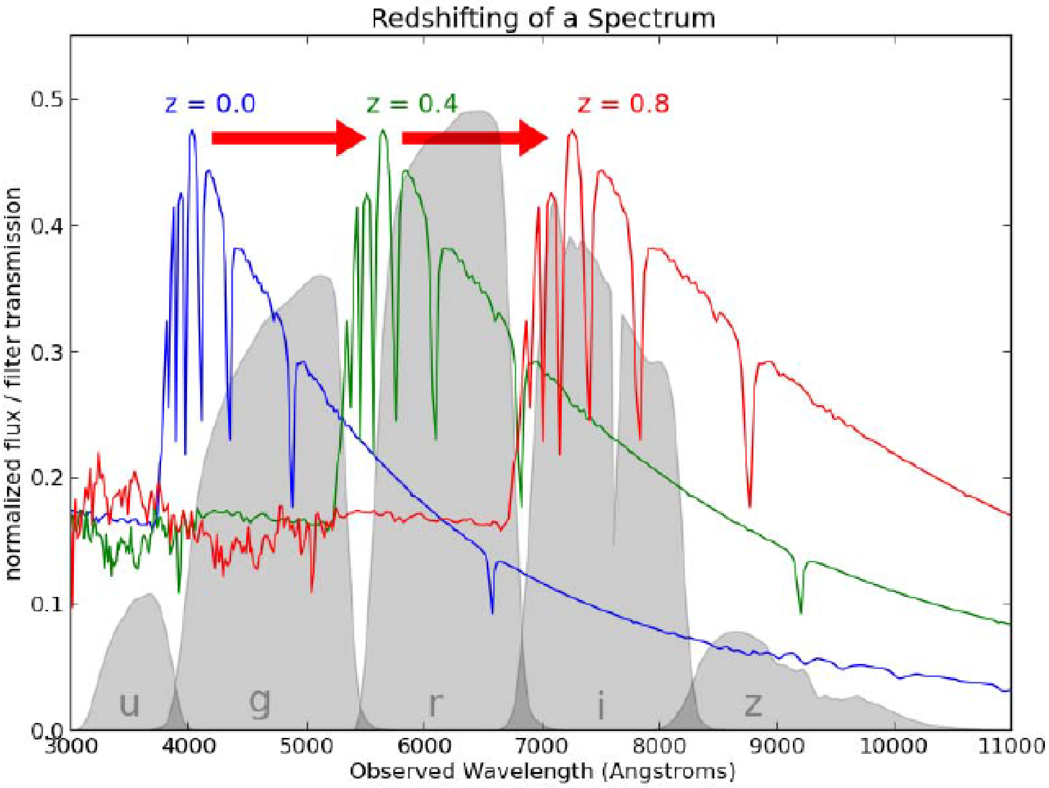


Figure 1.6: The spectrum of the star Vega (α-Lyr) at three different redshifts. The SDSS ugriz filters are shown in gray for reference. At redshift z = 0, the spectrum is bright in the u and g filters, but dim in the i and z filters. At redshift z = 0:8, the opposite is the case. This suggests the possibility of determining redshift from photometry alone. The situation is complicated by the fact that each individual source has unique spectral characteristics, but nevertheless, these photometric redshifts are often used in astronomical applications.
Credit of: Pedregosa et al. (2011)

spectrum shift, an identical source at different redshifts will have a different color through each pair of filters as illustrated in Fig. 1.6.
Technical advances in the instrumentation, combined with the development of 10 m class telescopes, guarantee a large increase in the number of detected galaxies, bright and faint, for which spectroscopic redshifts will be obtained in the near future. In spite of the progress in the numbers of available spectra, the $I \approx 24$ limit is likely to stand for awhile yet: from this value in magnitude then even the best instruments available produce spectra that are susceptible to line misidentification, even when carefully analyzed by expert observers. This means that most of the galaxies detected in very deep exposures are in practice inaccessible to spectroscopic analysis. The best example is the Hubble Deep Field North (HDF-N; Williams et al. (1996)): after several years of intensive efforts by the

astronomical community, the spectroscopic sample only comprises $\approx 20\%$ of the $I < 27$ galaxies detected in that field. Very few areas of the sky will be investigated in a similar way in the near future, so this value in magnitude is considered as a limit for any spectroscopic survey. In contrast, accurate photometric redshifts were quickly obtained for most of the HDF-N galaxies (notably by Sawicki et al. (1997); see also Lanzetta et al. (1996); Gwyn & Hartwick (1996)).

One of the fundamental issues in oservational astronomy is the estimation of redshifts for celestial objects, with the advent of modern multiband digital sky surveys, photometric redshifts (photo-z) have become crucial because provides redshift estimates for objects fainter than the spectroscopic limit and turn out to be much more efficient in terms of the number of objects per telescope time with respect to spectroscopic ones (spec-z).

The photometric redshift estimation is a long time problem (Baum (1962); Puschell et al. (1982); Koo (1985); Loh & Spillar (1986); Connolly et al. (1995)). It started as a rarely used technique for special kinds of objects and now is a tool widely used for a moltitude of observational programmes. From their beginning, photometric redshifts have been seen as an efficient way to study the statistical properties of galaxies and their evolution. They are essentially a technique for inverting a set of observable parameters (e.g. colors) into estimates of the physical properties of galaxies (e.g. redshift, type and luminosity).

Redshift for fainter objects are now accessible by photometry thanks to the improving telescope technology. This makes photo-z extremely attractive for observing programmes depending on redshifts especially with the advent of modern panchromatic digital surveys. For instance, they are essential in constraining dark matter and dark energy through weak gravitational lensing, for the identification of galaxy clusters and groups (e.g. Capozzi et al. (2009)), for type Ia supernovae, and for the study of the mass function of galaxy clusters (Albrecht et al. (2006); Peacock et al. (2006); Keiichi et al. (2012)).

There are many aspects which influence the performances of photo-z's. The observing strategy sets the theoretical limit for the accuracy: choice of the filters and the distribution of the available observing time over the different filters to reach certain depths can have a great impact on photo-z. Both the accurate photometric calibration and the removal of effects of the different point-spread-function (PSF) in the different bands are fundamental.

The need for fast and reliable methods of photo-z evaluation is important for improving ongoing and planned surveys; in fact, future large-field public imaging projects, such as KiDS (Kilo-Degree Survey), DES (Dark Energy Survey), LSST (Large Synoptic Survey Telescope) and Euclid (Euclid Red Book (2011)) require extremely accurate photo-z to obtain accurate measurements that do not compromise the survey's scientific goals.

1.2 Methods

Although originally the photo-z idea (Baum (1962)) was received with skepticism, the next generation telescopes plan to perform photometric observations completely based on these estimation techniques for most of their key science projects including cosmology and large-scale structure.

The evaluation of photo-z is possible thanks to the complex correlation between the fluxes, as measured in broad band photometry, the morphological types of the 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 challenge is to constrain physical properties of sources with some observables in a data set denoted by $Q : \{y_q\}$, starting from a training set, $T : \{x, \xi\}$, since model spectra would never be perfectly suitable for all desired parameters. In general, let be x a set of observables in the training set T that contains information about the physical properties ξ , and let y denote the observables of the query set Q. The model M can predict the observables x and y for a given parameter basing on the density $p(x, y \mid \theta, M)$ and has a prior on its parameters $p(\theta \mid M)$[1].

The goal is to derive the probability density function (PDF) of the physical properties ξ for a given query point q with y_q observations using the model M. This function, $p(\xi \mid y_q, M)$, is the solution of the generalized photometric inversion problem.

The first step is to establish the connection between the observables. It can be done formally by calculating the probability density of x for the query point q, according to the following formula

$$p(x \mid y_q, M) = \frac{p(x, y_q \mid M)}{p(y_q \mid M)} \tag{1.17}$$

where

$$\begin{aligned} p(x, y_q \mid M) &= \int d\theta \quad p(\theta \mid M) p(x, y_q \mid \theta, M) \\ p(y_q \mid M) &= \int dx \quad p(x, y_q \mid M). \end{aligned}$$

Traditionally the relation between the observable and the desired physical parameters assumes the properties of interest as a function of the observables. Some of the existing methods utilize explicit functions such as a polynomial or piecewise

[1] $p(\theta \mid M) = \frac{p(\theta, M)}{p(M)}$

linear, while others use mappings such as a decision tree or an artificial neural net. They are assuming a fitting function

$$\xi = \hat{\xi}(x), \tag{1.18}$$

tuned to reproduce the elements of the training set as well as possible. Since of the presence of degenerancies in most data sets, the Eq. 1.18 cannot guarantee that the same x observables always correspond to the same ξ properties. Clearly, the Eq. 1.18 is an unnecessary restriction over the general relation of x and ξ denoted by $p(\xi \mid x)$. Using Dirac's δ symbol, the traditional model prescript that

$$p(\xi \mid x) = \delta(|\xi - \hat{\xi}(x)|) \tag{1.19}$$

A better way is not to restrict the distribution arbitrarily to an unknown surface by observing that

$$p(\xi \mid x) = \frac{p(\xi, x)}{p(x)} \tag{1.20}$$

where both the densities in the ratio can be estimated from the training set. Using the Eq. 1.20, one can compute the final PDF of interest as the integral over the possible observables in the training set

$$p(\xi \mid y_q, M) = \int dx \quad p(\xi \mid x) \quad p(x \mid y_q, M) \tag{1.21}$$

Photometric redshifts and other such properties are often used in statistical studies thanks to their availability for a large number of sources, even though they provide relatively loose constraints on individual objects. The full PDFs of the sources are best suited to derive the ensemble properties of entire catalogs or even specific subsamples. The distribution of the properties over a set of measurements Q is given by the average

$$p(\xi \mid Q, M) = \langle p(\xi \mid y_q, M) \rangle \tag{1.22}$$

Essentially all currently existing implementations can be categorized into two classes of methods: theoretical and empirical.

Theoretical methods use templates, such as libraries of either observed galaxy spectra or model spectral energy distributions (SEDs). These templates can be shifted to any redshift and then convolved with the transmission curves of the filters used in the photometric survey to create the template set for the redshift estimators (e.g. Koo (1999), Massarotti et al. (2001a), Massarotti et al. (2001b), Csabai et al. (2003)). Photometric redshifts can be obtained by comparing observed galaxy fluxes at the ith photometric band, f_i^{obs}, with a library of reference fluxes, $f_i^{templ}(z, T)$ depending on redshift z on a set of parameters T, that accounts for galaxy morphological type, age, metallicity, dust reddening etc. For each galaxy a χ^2 confidence test provides the values of z and T that minimize flux residuals between observations and reference templates.

However, for datasets in which accurate and multiband photometry for a large number of objects are complemented by spectroscopic redshifts, and for a statistically significant subsample of the same objects, the *empirical methods* offer greater accuracy. These methods use the subsample of the photometric survey with spectroscopically-measured redshifts as a training set to constrain the parameters of a fit mapping the photometric data as redshift estimators.

The variety of methods and approaches and their application to different types of datasets, as well as the adoption of different and often not comparable statistical indicators, make it difficult to evaluate and compare performances. Blind tests of photo-z have been performed in Hogg et al. (1998) on spectroscopic data from the Keck telescope on the Hubble Deep Field (HDF), in Hildebrandt et al. (2008) on spectroscopic data from the VIMOS VLT Deep Survey (VVDS) and the FORS Deep Field (FDF; Noll et al. (2004), and in Abdalla et al. (2008)) on the sample of luminous red galaxies from the SDSS-DR6.

A significant advance in comparing different methods has been introduced by Hildebrandt and collaborators (Hildebrandt et al. (2010)), with the so-called PHAT (PHoto-z Accuracy Testing) contest, which adopts a black-box approach typical of benchmarking. They performed a homogeneous comparison of the performances, concentrating the analysis on the photo-z methods themselves that will be analyzed in next paragraphs.

1.2.1 Theoretical Methods

Template based techniques are free from the limitation of a training set and can be applied over a wide range of redshifts and intrinsic colors. They rely, however, on having a set of galaxy templates that accurately map the true distribution of galaxy spectral energy distributions (and their evolution with redshift) and on the assumption that the photometric calibration of the data is free from systematics. This approach simply compares the expected colors of a galaxy (derived from template spectral energy distributions) with those observed for an individual galaxy.

The standard scenario for template fitting is to take a small number of spectral templates T (e.g. E, Sbc, Scd and Irr galaxies) and choose the best fit by with a likelihood method for redshift, type and luminosity. Variations on this approach have been developed in the last few decades, as example one that use a continuous distribution of spectral templates enabling the error function in redshift and type to be well defined.

A representative set of spectrophotometrically calibrated spectral templates is not easy to obtain. Firstly, because it is complex to calibrate them spectrophotometrically over the full spectral range, secondly because of the redshift of a galaxy, we need spectra over a wavelength range that is wider than the range of actual optical filters (3000 - 12000 Å). Such spectra cannot therefore be measured by a single spectrograph. The alternative to empirical templates is to use the outputs of spectral synthesis models that provides the spectral energy distributions (SED) for the different objects.

As a first step, the SED fitting algorithms convert the galaxy observed magnitudes for each i-th photometric band into incoming apparent flux, $f_i^{obs}(\lambda)$. This is equivalent to reconstructing the SED of target galaxies at very low spectral resolution by sampling their luminosity at the effective wavelength of the available photometric bands.

For each object, the sampled flux has to be compared with the reference spectral libraries of template galaxies, $f_i^{templ}(z, T)$, computing the χ^2 of the fitting residuals such as

$$\chi^2 = \sum_{i=1}^{N} \frac{[f_i^{obs} - s f_i^{templ}(z,T)]^2}{\sigma_i^2} \tag{1.23}$$

where, N is the number of photometric bands, i is the number of band and s is the scale factor chosen in such a way as to minimize the χ^2 for each template:

$$s = \frac{\sum_{i=1}^{N} f_i^{obs} f_i^{templ}(z,T)/\sigma_i^2}{\sum_{i=1}^{N} f_i^{templ}(z,T)/\sigma_i^2}.$$

The selection of the reference flux library is very impotant in SED fitting methods. In particular, one has to decide whether it is more appropriate to use empirical or synthetic galaxy templates. The major advantage of the first choice is that the observed SEDs for a suitable set of local galaxies, spanning the whole range of Hubble morphological types, gives by definition a physically consistent picture of the real galaxies, at least the nearby ones. On the other hand, there is no obvious argument supporting a straightforward extension of the current galaxy properties to highredshift objects. At earlier cosmological epochs, evolution could play a substantial role in changing both morphological and spectrophotometric properties of distant galaxies (Buzzoni (1998)). To take into account evolution, synthesis models should be preferred; of course, some differences exist among current theoretical codes, especially in the SED predictions for the UV range (Charlot et al. (1996)).

In Fig. 1.7 are respectively shown the results of the template fitting technique using one of the most frequently used SED contained in Coleman, Wu & Weedman CWW (1980) on the left and a set of SEDs from the spectral synthesis

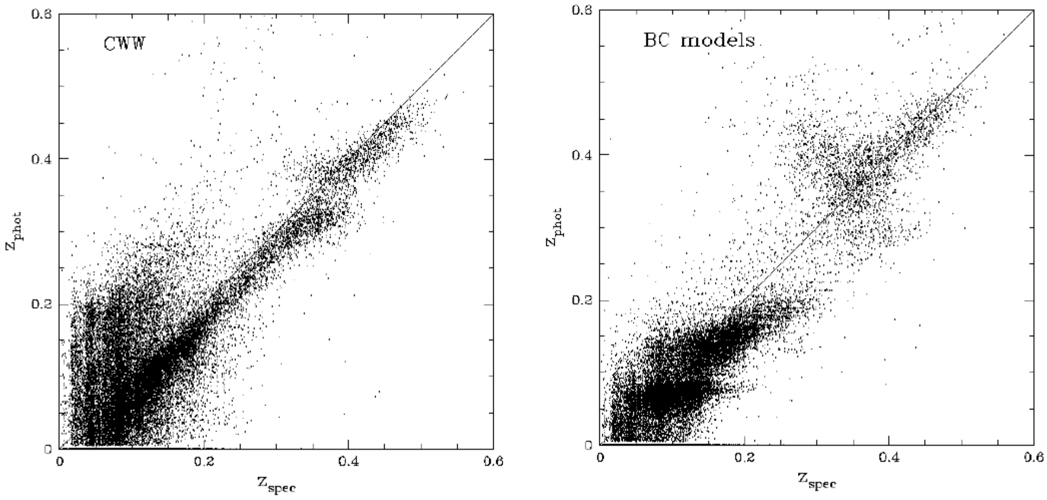


Figure 1.7: On the left: Photometric redshift estimation using the CWW spectral energy distributions. On the right: Photometric redshift estimation using the Bruzual and Charlot spectral energy distribution.

models of Bruzual & Charlot Bruzual & Charlot (1993) on the right. The dispersion about this relation is 0.062 and 0.051 for the CWW and BC templates respectively Csabai et al. (2003). The CWW templates produce a photometric redshift relation where the majority of galaxies have a systematically lower redshift than that given by the spectroscopic data (by approximately 0.03 in redshift) and there is a broad tail of galaxies of low z, for which the photometric redshifts are systematically overestimated. For the BC templates the galaxy redshifts tend to be systematically underestimated, with a greater dispersion for redshifts $z \geqslant 0.3$).

An improvement over standard template methods is the introduction of magnitude priors within a Bayesian framework Benitez (2000).

BPZ

Bayesian Photo-z, BPZ Benitez (2000), introduced the use of Bayesian inference and priors to photometric redshift estimation. The code uses a prior $P(z; T \mid m_0)$ which gives the likelihood that given an apparent magnitude m_0 for a galaxy with redshift z and SED type T. For each galaxy, this information is combined (in a Bayesian manner) with the likelihood $P(C \mid z; T)$ of observing the galaxy colours C for each redshift and SED pair, yielding the final $P(z; T \mid C; m_0)$.

EAZY

Easy and Accurate Redshifts from Yale, EAZY Brammer et al. (2008), is a template-fitting code designed to produce un-biased photometric redshift estimates for deep multi-wavelength surveys that lack representative calibration samples with spectroscopic redshifts.
EAZY uses a unique template set derived using the non-negative matrix factorisation algorithm (Sha et al. (2007); Blanton & Roweis (2007)) trained on synthetic photometry from the semi-analytic light-cone produced by De Lucia & Blaizot (2007). These templates can be considered the principal component spectra of all galaxies at $0 < z < 4$ in the light-cone, allowing for subtle differences between local and high-redshift galaxy samples. EAZY is able to reproduce complex star-formation histories by fitting non-negative linear combinations of the templates, wich include emission lines following the prescription of Ilbert et al. (2009).

Hyperz

Hyperz Bolzonella et al. (2000) is a publicly available code based on SED templates fitting using a standard χ^2 minimisation method. The codes uses the observed fluxes of an object in a set of given filters and compares them with the theoretical fluxes of galaxies in the same filters obtained from template spectra, either synthetic or empirical, taking into account the observational uncertainties but also the possible observational hidden effects such as reddening. It computes not only a best-fit solution which minimises the differences, therefore a most probable photometric redshift, but also a full probability function depending on the redshift.
Hyperz uses a given set of templates, filters, reddening laws and Lyman forest modelling and can be easily adapted to use any kind of parameters that would fit the needs of the user.

Le Phare

The public code Le Phare, PHotometric Analysis for Redshift Estimate (Arnouts et al. (2008); Ilbert et al. (2006)) is primarily dedicated to estimate photo-z, but it can also be used to estimate other physical parameters like stellar masses and infrared luminosities. Le Phare is based on a standard template fitting procedure; the templates are redshifted and integrated through the instrumental transmission curves. The opacity of the intergalactic medium (IGM) is taken into account and internal extinction could be added as a free parameter to each galaxy. The photo-zs are obtained by comparing the modelled fluxes and the observed fluxes with a χ^2 merit function.

LRT

Low-Resolution Spectral Templates, LRT (Assef et al. (2008), Assef et al. (2010)), is a set of subroutines intended to estimate $K-$corrections[2] and photometric redshifts on the basis of empirical low resolution SED templates (hence LRT) for galaxies and AGNs. Every galaxy is represented by a non-negative linear combination of three empirically determined SED templates that resemble an elliptical, an Sbc spiral and an Im irregular galaxy.

ZEBRA

Zurich Extragalactic Bayesian Redshift Analyzer, ZEBRA Feldmann et al. (2006), is an open source photometric redshift code based on a SED template-fitting approach. Built on top of a traditional Maximum Likelihood approach it introduces and combines several novel methods that help to improve the accuracy of photometric redshift estimates for galaxies and AGNs (Oesch et al. (2010); Luo et al. (2010), for some recent applications). ZEBRA is able to detect and correct photometric offsets in the input catalogue and can use spectroscopic redshifts on a small fraction of the photometric sample to iteratively correct the original set of input templates. This correction allow the decrease of the bias, the scatter, and the number of outliers in the redshift estimation. When run in Bayesian mode, ZEBRA computes the prior in redshift-template space in a self-consistent manner from the input catalogues and the redshift-template likelihood functions. This prior is consequently used to derive the posterior probability distribution of each input object.

1.2.2 Empirical Methods

Empirical approach can be tipically applied to galaxies with colors that lie within the range of colors and redshifts found within the training set. One of the first

[2] The $K-$correction can be defined as the correction needed to transform the observed magnitude through bandpass b of an object at redshift z to the magnitude we would measure for an object with the same SED and the same apparent bolometric magnitude but located at redshift z_0.

$$m_b(z) = m_b(z_0) + K_b$$

and the $K-$correction K_b is defined

$$K_b = -2.5log\left[\frac{(1+z)}{(1+z_0)}\frac{\int_0^\infty \frac{R_b(\lambda)}{\lambda} f[(1+z)\nu]d\lambda}{\int_0^\infty \frac{R_b(\lambda)}{\lambda} f[(1+z_0)\nu]d\lambda}\right]$$

where $f(\nu)$ is the rest frame SED and R_b is the filter bandpass response per photon of wavelength λ. Usually $z_0 = 0$ so that magnitude is corrected to the rest frame.

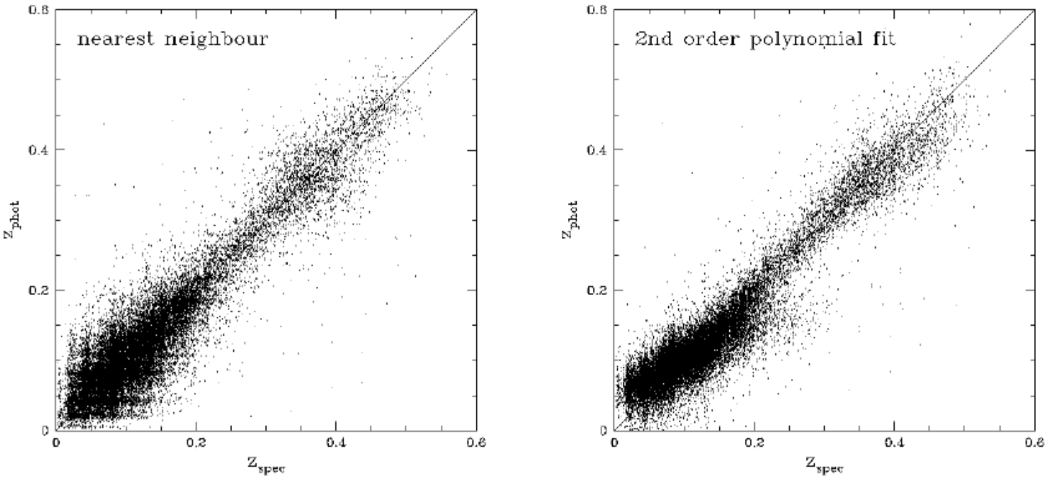


Figure 1.8: The photometric redshift estimations with the simple empirical methods.

successful empirical methods is based on fitting the relation between the spectroscopic redshift of a galaxy and its colors or magnitudes Connolly et al. (1995). The fitting function is typically a 2nd or 3rd order polynomial. Fig. 1.8 shows the photometric vs. the spectroscopic redshifts using data from Early Data Release of the SDSS (EDR main galaxy and Luminous Red Galaxy spectroscopic samples Csabai et al. (2003). As the size of the training set is large (more than 30,000 objects) when compared to the number of fitted parameters (21), we can expect that this fit will work likewise as long as the data are selected over the same color and redshift range as the training set. One of the most important uncertainty within this technique comes from the fact that the fitting function is just an approximation of the more complex relation between colors and redshift of a galaxy. We would, therefore, expect the fitting function to accurately follow the redshift-color relation over a narrow range of redshift otherwise one can use separate functions in different redshift (Brunner et al. (1999)) or color ranges.

A second empirical estimator is the *"nearest neighbour"* method. In the training it finds the galaxy within the set with the smallest distance in the color (or magnitude) space (weighted by the errors) and its redshift is assigned to the test galaxy. In the ideal case the training set contains sufficient galaxies to find a close neighbour for each unknown object. In Fig. 1.9 we see that redshift estimation error increases with the distance from the nearest neighbour in color space. Obviously, a larger dataset correspond to a better accurancy, as long as that all galaxy types are represented in the training set. By larger training sets mean that the search time increases so it has needed to use an efficient multidimensional search technique (e.g. kd-trees) instead of a standard linear

search. The comparison between the estimated and spectroscopic redshifts for the nearest-neighbour technique is given in Fig. 1.8. The dispersion about this relation is $\sigma_z = 0.033$.
A natural limitation of the nearest neighbour technique is that a large number of training galaxies alone is not enough to cover the range of the colors of the unknown objects in a more or less uniform way. To resolve this problem one can search for more than one nearest neighbour and apply an interpolation or a fitting function, solving also a second problem, namely that because of the finite number of objects in the training set, the photometric redshifts will have discrete values making them problematic to use in some statistical studies. First results from the PHAT contest, presented in Hildebrandt et al. (2010), described the pro's and con's of many other different empirical methods.

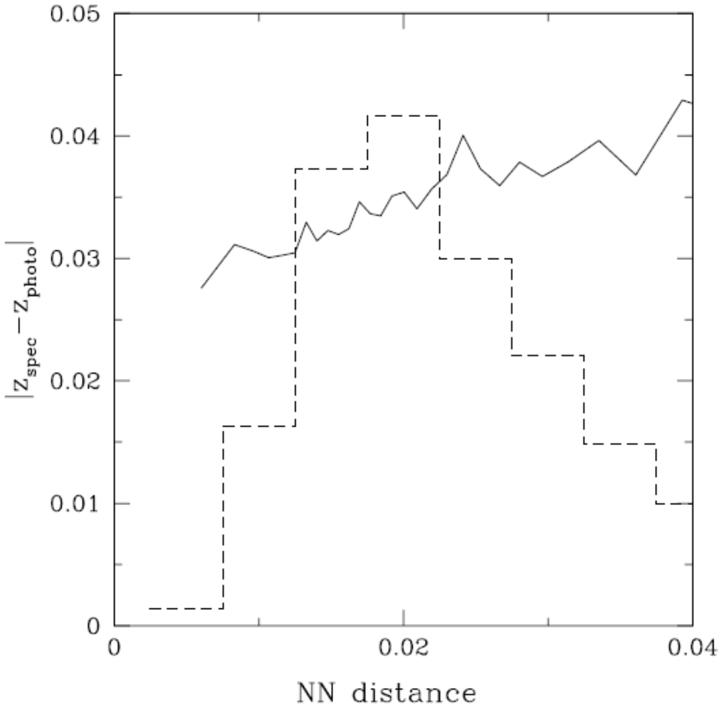


Figure 1.9: The dependence of redshift average estimation error on the color space distance from the nearest reference object (solid line). As expected, smaller distances result smaller error. The dashed line is for the histogram of number of objects with a given nearest neighbour distance. One can see, that for most of the objects the nearest neighbour is not close enough to get the best estimation.

ANNz

ANNz Collister & Lahav (2004) is an empirical photo-z code based on Artificial Neural Networks, made up of several layers, each consisting of a number of nodes. The first layer receives the galaxy magnitudes as inputs, while the last layer outputs the estimated photometric redshift. The layers contains the nodes wich are inter-connected and carries a *"weight"* which is a free parameter in the parametrisation. When a network is trained the weights of all node connections are determined by minimising a cost function E. Then the algorithm is applied on the sample for redshift estimation. Neural networks have been used e.g. for estimation of photo-z for the SDSS (Collister et al. (2007); Abdalla et al. (2008)), as well as forecasts of photometric redshifts for future surveys like the Dark Energy Survey Banerji et al. (2008) and Euclid Abdalla et al. (2008).

Boosted Decision Tree (BDT) algorithm

The Boosted Decision Tree (BDT) algorithm Gerdes et al. (2010) combines an ensemble of weak classifiers into a single powerful classifier. The spectroscopic training set is first divided into redshift bins whose width is approximately half the expected photo-z resolution of the algorithm for the given sample. For each bin, a set of trees is trained labeling as "*signal*" those galaxies whose redshift falls within the bin in question, and "*background*" those that fall more than 2σ away from the signal bin, where σ is the iteratively-determined photo-z resolution. As training variables are used the observed magnitudes in each band. The result is a tree containing nodes with predominantly signal and predominantly background galaxies. The process of *boosting* iteratively repeats this process, giving higher weight to galaxies that were initially misclassified. The method produces a photo-z probability for each galaxy as a function of redshift. This method provides an estimate of the best photo-z, of the error and a reconstruction of the full redshift PDF. In Gerdes et al. (2010) it was shown that the BDT algorithm improves upon the default photo-z in the SDSS spectroscopic sample and that the PDFs yield a more accurate reconstruction of the redshift distribution.

Purger (Nearest-Neighbour Fit)

This empirical method compares the observed colours to the reference set. The estimation method first searches the colour space for the k nearest neighbours of every object in the estimation set (i.e. the galaxies for which we want to estimate redshift) and then estimates the redshift by fitting a local low order polynomial to these points. An improved version of this code is using a k-d tree index for fast nearest neighbour search Csabai et al. (2007). It was used to calculate photometric redshifts for the SDSS Data Release 7 Abazajian et al.

(2009). The advantage of this method versus a template-based method might be the better estimation accuracy, but it cannot extrapolate so the completeness of the reference set is crucial.

Random Forest for Photometric Redshifts

The method by Carliles et al. (2010) is based on Random Forests which are an empirical, non-parametric regression technique. A Random Forest builds an ensemble average of randomised regression tree redshift estimates. Bootstrap samples are extracted from the training set and each regression tree is trained on its own bootstrap sample. Given a new test object, each regression tree produces its own redshift estimate; all the estimates are averaged to yield the final Random Forest redshift estimate. This technique also results in Gaussian errors and this behaviour has a strong theoretical statistical explanation. For a new galaxy one can hypothesise the existence of a distribution which reflects the similarity of the new galaxy to any given point in the event space. The Random Forest approximates this distribution per object and the process results in easily computable per-object error parameter estimates.

MLPQNA

Given the huge dataset collected for this thesis (see Ch.§2) the better method to evaluate photo-z is the empirical one. In this thesis I used a machine learning technique called MLPQNA. It is a Multi Layer Perceptron Bishop (2006) implemented with a learning rule based on the Quasi Newton Algorithm (QNA).
A Multi-Layer Perceptron may be represented by an input layer, with a number of perceptrons equal to the number of input variables, an output layer, with as many neurons as the output variables; the network may have an arbitrary number of hidden layers (in most cases one) which in turn may have an arbitrary number of perceptrons. In a fully connected feed-forward network each node of a layer is connected to all nodes in the adjacent layers representing an adaptive weight calculated on the strength of the synaptic connection between neurons. In order to find the model that best fits the data, one has to provide the network with a set of examples.
In general Quasi Newton Algorithms (QNA) are variable metric methods used to find local maxima and minima of functions Davidon (1968) and, in the case of MLP's they can be used to find the stationary (i.e. the zero gradient) point of the learning function. QNA are based on Newton's method to find the stationary point of a function, where the gradient is 0. The Hessian matrix of second derivatives of the function to be minimized does not need to be computed, but is updated by analyzing successive gradient vectors instead.

The workflow to determine phot-z is the follow: extraction of the Knowledge Base (KB), i.e. the training set, determination of the optimal model parameter setup and application of the tuned model to the whole dataset.

A more accurate characterization of functionalities of the MLPQNA has been described in Ch.§3.

Chapter 2

Experimental Data

The first systematic survey of all that is visible by the naked eye was performed by Hipparchus in the 2nd century BC: he drew up a catalog including about 850 stars. After more than a thousand years, in 1968, using the telescope Galileo began a revolution for astronomical observations: for the first time the craters of the Moon and the Jupiter's moons of Jupiter were observed.

In 1917, a new telescope was built on Mount Wilson in California; it was the largest ever built in the world and it unveiled an entirely new picture of the universe. Thanks to this giant telescope astronomers discovered that many of the nebulae were other galaxies like our own Milky Way.
Although spacecraft missions have revolutionized our understanding of the solar system, ground-based telescopes continue to play a very important role in making new discoveries. During the 1970s, NASA constructed ground-based telescopes to support its planetary missions; it funded the construction of the 2.7 m McDonald telescope, the University of Hawaii 2.2 m telescope and the 3.0 m NASA Infrared Telescope Facility (IRTF) to provide mission support.

Cosmology had seen the demise of the perfect cosmological principle in 1929, with Edwin Hubble's discovery that universe is expanding and therefore changes over time, but the Universe is yet considered homogeneous and isotropic. To demonstrate the validity of this basic assumption it is necessary to be able to find some volumes of the Universe that are representative of the whole. We know that telescopes are designed to collect and focus starlight onto a detector, while ground-based observers have to contend with limitations imposed by physics, the atmosphere and technology: for this purpose it is clear that very large galaxy surveys are crucical.

The advent of the new class of 10 m ground-based telescopes is having a strong impact on the study of galaxy evolution. Sky surveys and so-called *deep* *fields* represent different strategies for studying extraterrestrial objects. In particular **sky surveys** include projects performing photometric and/or spectral observations of a significant fraction of the sky. The effective depth of surveys is $z \sim 0.1$ or several hundred megaparsecs (Mpc). Surveys are often restricted to one band of the electromagnetic spectrum due to instrumental limitations, although multiwavelength surveys can be made by using multiple detectors, sensitive to different bandwidths.

Digital sky surveys are essentially changing the field of research of astronomical data because of the sheer quantity of data being generated over multiple wavelengths and the homogeneity of the data within each survey. The federation of different surveys would further improve the efficacy of future ground- and space-based targeted experiments and also open up entirely new avenues for research.

Photographic plates have long endured as efficient mechanisms for recording surveys; indeed, they have useful lifetimes and offer superb information storage capacity, but unfortunately they are not directly computer-accessible and must be digitized before being put to a modern scientific use. Their supremacy in a digital world, however, is being challenged by new technologies. Indeed, many photographic surveys have been performed, for example from the Palomar Schmidt telescope in California and the UK Schmidt telescope in New South Wales (Australia), but their data become most useful when the plates are digitized and catalogued.

Traditional ground-based observatories have been saved data, mainly as emergency backups for the users, for a significant time, accumulating impressive quantities of highly valuable and heterogeneous data. Unfortunately the lack of adequate funding have limited the efforts to properly archive this wealth of information and make it easily available to the broad astronomical community.

There is a large number of experiments and surveys; as it is impossible to describe here all of them, I will report only an example of valuable and useful sky surveys that fills archives with their images.

DPOSS - The **D**igitized **P**alomar **O**bservatory **S**ky **S**urvey is a digital survey of the entire Northern Sky in three visible-light bands, formally indicated by g, r and i (bluegreen, red and nearinfrared respectively).

It is based on the photographic sky atlas, POSSII, the second Palomar Observatory Sky Survey, which was completed at the Palomar 48inch Oschin SchmidtTelescope Reid et al. (1991). It consist of a set of three photographic plates, one for filter, each covering 36 square degrees. It takes at each of 894 pointings spaced by 5 degree covering the Northern sky, al-

though many of these were repeated exposures, due to various artifacts such as the aircraft trails, plate defects, etc. The plates were then digitized at the Space Telescope Science Institute (STScI) and scanned producing about 1 Gb per plate or about 3 Tb of pixel data in total.
These scans were processed independently at STScI, to construct a new guide star catalog for the HST and at Caltech, for the DPOSS project. Catalogs of all the detected objects on each plate were generated, down to the flux limit of the plates which roughly corresponds to the equivalent blue limiting magnitude of approximately 22.

2MASS - The **Two M**icron **A**ll **S**ky **S**urvey project is designed to close the gap between our current technical capability and the knowledge of the near-infrared sky. In addition to provide a context for the interpretation of results obtained at infrared and other wavelengths, 2MASS is giving direct answers to immediate questions on the large-scale structure of the Milky Way and the Local Universe. The optimal use of the next generation of infrared space missions, such as HST/NICMOS, the Space Infrared Telescope Facility (SIRTF) and the Next Generation Space Telescope (NGST), as well as powerful ground-based facilities, such as Keck I, Keck II and Gemini, require a new census with vastly improved sensitivity and astrometric accuracy greater than previously available.
To achieve these goals, 2MASS has uniformly scanned the entire sky in three near-infrared bands to detect and characterize point sources brighter than about 1 mJy, with signal-to-noise ratio (SNR) greater than 10, using a pixel size of 2.0".
2MASS used two highly-automated 1.3 m telescopes, one at Mt. Hopkins (Arizona) and one at CTIO (Chile). Each telescope was equipped with a three-channel camera, each one consisting of a 256×256 array of HgCdTe detectors, capable of observing the sky simultaneously at J (1.25 microns), H (1.65 microns) and K_s (2.17 microns). The northern 2MASS facility began routine operations in 1997 June and the southern facility in 1998 March. Survey operations were complete for both hemispheres on 2001 February 15.

In the next section the focus will be on the **S**loan **D**igital **S**ky **S**urvey (SDSS), in particular on its Ninth Data Release (DR9), because data used during this thesis were extracted from DR9 catalogues.

2.1 SDSS

The **S**loan **D**igital **S**ky **S**urvey started as a large astronomical collaboration focused on constructing the first CCD photometric survey of the North Galactic hemisphere (10000 square degrees, $\sim 1/4$ of the entire sky). So for the SDSS is the most ambitious and important survey in the history of astronomy.
It uses a dedicated wide-field 2.5 m telescope Gunn et al. (2006) at Apache Point Observatory (APO) in the Sacramento Mountains (Southern New Mexico). It was originally instrumented with a wide-field imaging camera with an effective area of 1.5 deg^2 Gunn et al. (1998) and a pair of double spectrographs fed by 640 fibers Smee et al. (2013). The initial survey York et al. (2000) carried out imaging in five broad bands (*ugriz*) Fukugita et al. (1996) to a depth of $r \sim 22.5$ over 11,663 deg^2 of high-latitude sky and spectroscopy of 1.6 million galaxy, quasar and stellar targets over 9380 deg^2. The spectra were calibrated and redshifts and classifications determined Bolton et al. (2012). The data have been released publicly in a series of roughly annual data releases as the project went through two funding phases, named SDSS-I (2000-2005) and SDSS-II (2005-2008). Over eight years of operations it obtained deep, multi-color images covering more than a quarter of the sky and created 3-dimensional maps containing more than 930,000 galaxies and more than 120,000 quasars.
In 2008, the SDSS entered a new phase with four components, designated SDSS-III Eisenstein et al. (2011), currently operating: a total of 1,231,051,050 imaging data are catalogued (469,053,874 object, removing all duplicates and overlaps) and the last Data Release (DR10) includes a total of 1,507,954 BOSS spectra, comprising 927,844 galaxy spectra, 182,009 quasar spectra and 159,327 stellar spectra.

- The Baryon Oscillation Spectroscopic Survey (**BOSS**) Dawson et al. (2013) has substituted the spectrographs to improve the rate of work and increase the number of fibers to 1000 Smee et al. (2013). BOSS enlarged the imaging footprint of SDSS to 14,555 deg^2, it is obtaining spectra of galaxies and quasars with the primary goal of measuring the oscillation signature in the clustering of matter as a cosmic measure to constrain cosmological models.

- The Sloan Extension for Galactic Understanding and Exploration 2 (**SEGUE-2**), an expansion of a similar project carried out in SDSS-II Yanny et al. (2009), used the SDSS spectrographs to obtain spectra of about 119,000 stars, mostly at high Galactic latitudes.

- The Apache Point Observatory Galactic Evolution Experiment (**APOGEE**) uses a 300-fiber spectrograph to observe bright ($H < 13.8$

mag) stars in the H band at high resolution ($R \sim 22,500$) for accurate radial velocities and detailed elemental abundance determinations.

- The Multi-Object APO Radial Velocity Exoplanet Large-area Survey (**MARVELS**), which finished its data-taking in 2012, used a 60-fiber interferometric spectrograph to measure high-precision radial velocities of stars in a search for planets and brown dwarfs.

SDSS-III began observations in July 2008 and gave the Eighth Data Release (DR8) Aihara et al. (2011) in January 2011. included all data from the SEGUE-2 survey, as well as ~ 2500 deg^2 of new imaging data in the Southern Galactic Cap as part of BOSS.
The Ninth Data Release (DR9) Ahn et al. (2012) included the first spectroscopic data from the BOSS survey: over 800,000 spectra selected from 3275 deg^2 of sky. The Tenth Data Release (DR10) given to the public on 31 July 2013, offers the latest data from the Sloan Digital Sky Survey. DR10 includes almost 680,000 new BOSS spectra, covering an additional 3100 deg^2 of sky. It also includes the first public release of APOGEE spectra, with almost 180,000 spectra of more than 57,000 stars in a wide range of Galactic environments, in addition to all imaging and spectra from prior SDSS data releases.
DR11 will be an internal release only; for a public release would occur only six months before the final public data release for SDSS-III, DR12, which will be released in December 2014 and will contain all of data taken during the six years of the project. All data released are publicly available on the SDSS-III website[1].

For the analysis described in this work the catalogues contained in DR9, released in August 2012, were used.

2.1.1 The Ninth Data Release

DR9 presents the release of the first 1.5 years of data from the SDSS-III BOSS spectroscopic survey. BOSS began survey-quality observations on the night of 2009 December 5. DR9 contains all processed data until the telescope shutdown occurred in 2011 July[2], included the spectroscopic data from SDSS-I/II and SEGUE2; the details of the data included in DR9 are summarized in Table 2.1.1. The imaging data and catalogs are the same present in DR8, with an improved astrometric solution that correct an error affecting objects at high declinations Aihara et al. (2011b).

[1] http://www.sdss3.org/

[2] The SDSS telescope pauses science operations during the monsoon in July/August in the southwestern United States. This time is used for telescope maintenance and engineering work.

Imaging[a]		
	Total	Unique[b]
Area Imaged	31,637 deg^2	14,555 deg^2
Cataloged Objects	1,231,051,050	469,053,874
New BOSS Spectroscopy		
	Total	Unique[b]
Spectroscopic Footprint effective area	...	3275 deg^2
Plates	831	819
Specta Observed	829,073	763,425
Galaxies	535,995	493,845
CMASS galaxies	336,695	309,307
LOWZ galaxies	110,427	102,890
ALL Quasars	102,100	93,003
Main Quasars	85,977	79,570
Main Quasars, $2.15 < z < 3.5$	59,783	55,047
Ancillary program spectra	32,381	28,968
Stars	90,897	82,645
Standard stars	16,905	14,915
Sky spectra	78,573	75,850
All spectroscopyy from SDSS-I/II/III		
Total number of spectra	2,674,200	
Total number of useful spectra[c]	2,598,033	
Galaxies	1,457,002	
Quasars	228, 468	
Stars	668,054	
Sky	181,619	
Unclassified[d]	62,890	

Table 2.1.1: Contents of DR9. **a)** These numbers are unchanged since DR8. **b)** Removing all duplicates and overlaps. **c)** Spectra on good or marginal plates. Spectrum refers to a combined set of sub-exposures that define a completed plate. Duplicates are from plates that were observed more than once, or are objects that were observed on overlapping plates. **d)** Non-sky spectra for which the automated redshift/classification pipeline Bolton et al. (2012) gave unreliable results, as indicated by the ZWARNING flag.

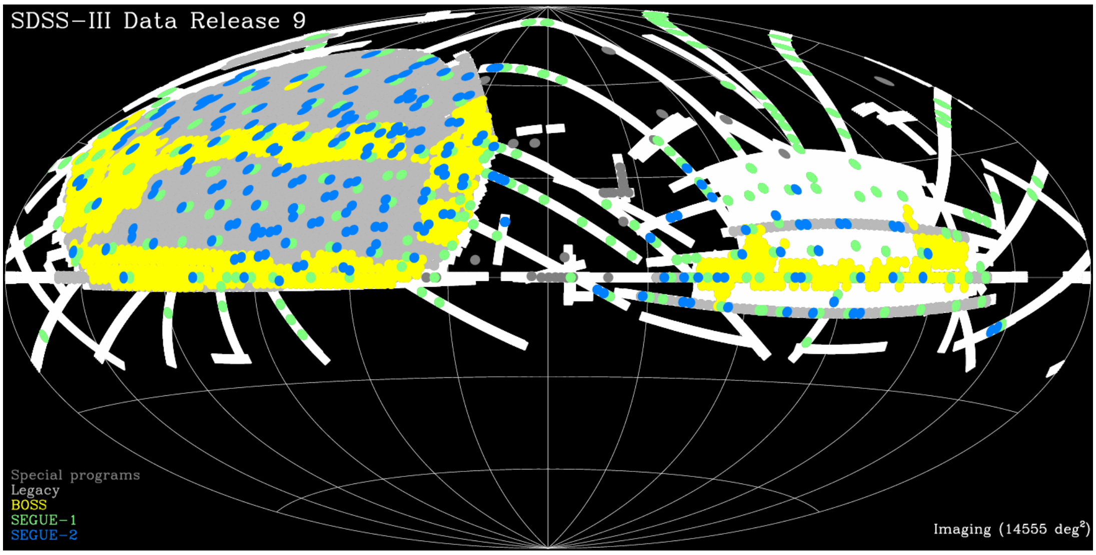


Figure 2.1: The sky coverage of DR9.

The SDSS is targeted on different items.
BOSS aims to measure spectra and redshifts for a sample of 1.5 million galaxies extending to z = 0.8 over 10,000 deg^2. In addition, 150,000 quasars with $z > 2.15$ will be observed to measure the clustering of the Lyman-α forest and to determine the baryon oscillation scale at $z \sim 2.5$, the era preceding the domination of the dark energy in the expansion of the universe.
BOSS aims also to measure large-scale clustering of galaxies at higher redshifts and at lower luminosities: so it samples the density field at higher space density and can target significantly fainter galaxies.
The samples of galaxies and quasars needed to carry out this program are significantly fainter than those targeted in SDSS-I and SDSS-II Eisenstein et al. (2001); Strauss et al. (2002); Richards et al. (2002) and have an higher density on the sky. So, the SDSS spectrographs and the related infrastructure were extensively rebuilt to increase observing efficiency, as described in detail in Smee et al. (2013).
In BOSS data we can select four categories: galaxies, quasars Ross et al. (2012), ancillary targets and standards and calibrations Dawson et al. (2013). The SDSS-III Data Release 9 presents the first data from the BOSS survey, with $\sim 102,000$ new quasar spectra, $\sim 91,000$ new stellar spectra and $\sim 536,000$ new galaxy spectra. We will focusing on the objects selected with the data queries carried out in this thesis.
The SDSS-I/II Legacy survey have labeled the galaxies into two categories: a magnitude-limited sample of galaxies in the r band Strauss et al. (2002), with a

median redshift of $z \sim 0.10$ and a sample of fainter galaxies limited in magnitude and color designed to select the most luminous red galaxies (LRG) at each redshift Eisenstein et al. (2001). The LRG sample is approximately volume-limited to $z \sim 0.38$ and includes galaxies to $z \sim 0.55$.

The galaxy target selection algorithm uses the DR8 imaging catalog: it selects two categories of objects using the different colors of the galaxy population evolution with redshift Maraston et al. (2009). The LOWZ subsample, containing about a quarter of all galaxies in BOSS, targets galaxies with $0.15 < z < 0.4$ with colors similar to LRGs, but with lower luminosity. The constant-mass (CMASS sample), contains three times more galaxies than LOWZ and is designed to select galaxies with $0.4 < z < 0.8$. The rest-frame color distribution of this sample is significantly broader than that of the LRG sample, thus CMASS contains a nearly complete sample of massive galaxies above the magnitude limit of the survey.

The galaxies contained in the BOSS sample have a magnitude fainter than the SDSS-I/II LRG sample and thus the ratio S/N of the spectra tends to be lower, despite the higher throughput of the spectrographs. Nevertheless, in DR9 the vast majority of the galaxy targets are confirmed galaxies with confidently measured redshifts: 95.4% of all CMASS and 99.2% of LOWZ. The 4.6% of unsuccessful galaxy redshifts for CMASS targets are mostly erroneously targeted red stars.

The BOSS spectrographs include 1000 fibers in each plate, in comparison with the 640 fibers per plate in SDSS-I/II. In addition, the spectral coverage has been increased from 3800-9200 Å to 3600-10,400 Å. The median resolution of the BOSS spectra remains $R = \lambda/\Delta\lambda \approx 2000$ as in SDSS-I/II, with a similar wavelength dependence Smee et al. (2013); in particular the resolution goes from $R \approx 1500$ at 3700 Å, to $R \approx 2500$ at 9500 Å.

The diameter of the spectroscopic fibers in BOSS has been decreased in size from 3” to 2”; this improves the S/N ratio for point-like objects and smaller galaxies (due to decreased sky background relative to the source signal), but the smaller fiber size affects the spectrophotometry and is more subject to differential chromatic aberration and seeing effects. As in SDSS-I/II, the spectrophotometry is bound to the PSF photometry of stars on each plate. In SDSS-I/II, the Root Mean Square (RMS) scatter between the PSF photometry and synthesized photometry from the calibrated spectra was of order 4% Adelman-McCarthy et al. (2008); with BOSS, it is closer to 6% Dawson et al. (2013). The photometric catalog released in DR8 and DR9 provides the 2” photometry (termed FIBER2MAG) for each object to complement 3” photometry (FIBERMAG).

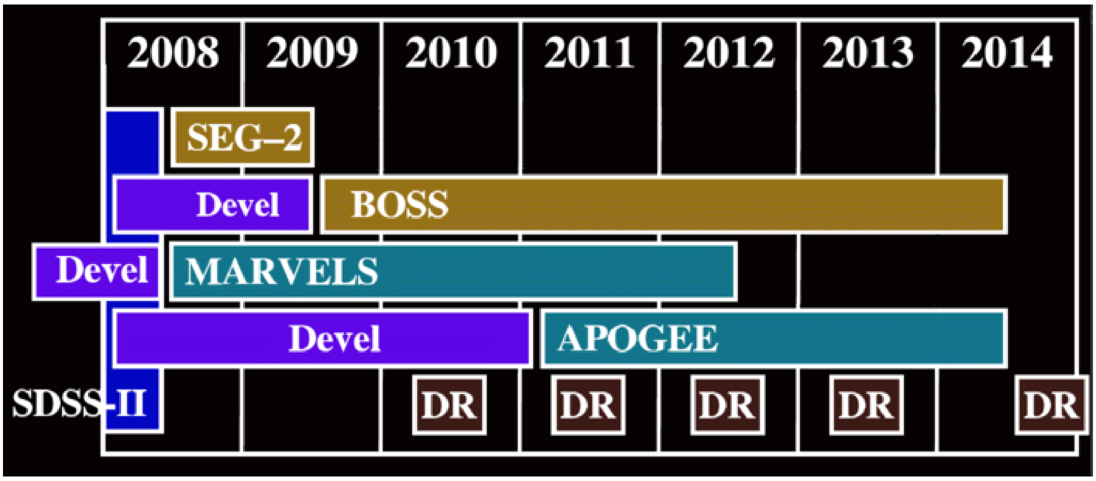


Figure 2.2: Final spectra for SEGUE-2 were released with DR8 in 2011 January. The MARVELS and BOSS spectroscopic surveys began in 2008 and 2009 and APOGEE began in 2011. SDSS-III will be busy taking data through the summer of 2014. The BOSS and SEGUE-2 programs require "dark" time when the Moon is less than 60% illuminated, or below the horizon. The APOGEE and MARVELS programs are executed during the remaining "bright" time.

2.1.2 Data Extraction

All DR9 data are available through data access tools reported on its website.

Data Release 9 includes images, spectra and catalog data. The data are stored both in the Science Archive Server (SAS) and in the Catalog Archive Server (CAS). A number of different interfaces are available, each designed to accomplish a specific task; as example:

1. color images of sky regions in JPEG format, based on the g, r and i images Lupton et al. (2004) can be viewed in a web browser with the SkyServer Navigate tool;

2. FITS images can be downloaded through the SAS;

3. complete catalog information (astrometry, photometry etc.) of any imaging object can be viewed through the SkyServer Explore tool;

4. FITS files of the spectra can be downloaded through the SAS.

Catalog data contain quantities measured from the images and spectra such as magnitudes, redshifts and object classifications. These are available either from the CAS database or as binary tables in FITS file format. As the aim of this thesis is to show the application of a photometric redshift evaluation method on real data objects, it has been necessary to create a dataset from the entire

catalog.
Different catalog search tools are available through the SkyServer interface to the CAS, each of which returns catalog data for objects that match supplied criteria. For more advanced queries, a powerful and flexible catalog search website called “CasJobs” allows users to create the own data sets. The DR9 web site also contains data access tutorials, a glossary of SDSS terms and a detailed documentation about the algorithms used to process the imaging and spectroscopic data and select spectroscopic targets. Imaging and spectroscopic data from all prior data releases are also available through DR9 data access tools.

CASJOBS Query Construction

CasJobs is an online workspace for large scientific catalogs, designed to emulate and enhance local free-form query access in a web environment.

This application includes, as example:

- synchronous and asynchronous query execution, in the form of ‘quick’ and ‘long’ jobs;
- a query ‘History’ that records the queries and their status;
- a server-side, personalized user database, called *MyDB*, enabling persistant table/function/procedure creation;
- data sharing between users, via the ‘Groups’ mechanism;
- data download, via MyDB table extraction, in various formats;
- multiple interface options, including a browser client as well as a java-based command line tool.

In our specific case, the data needed for photo-z evaluation were the magnitudes of all galaxies in DR9 archives. So, 38 queries were executed to cover all the DR9 map (see Fig.2.1) with the following stracture.

```
SELECT objID, ra, dec, psfMag_u, psfMag_g, psfMag_r, psfMag_i,
    psfMag_z, psfMagErr_u, psfMagErr_g, psfMagErr_r, psfMagErr_i,
    psfMagErr_z, extinction_u, extinction_g, extinction_r, extinction_i,
    extinction_z
FROM Galaxy
WHERE (dec between * and *)
```

AND clean = 1

AND psfMag_r < 23.6

The **SELECT** command allows to choose which parameters will compose the final dataset. In this case is possible to recognize:

objID, the SDSS object's identifier;

ra, dec, the right ascension and declination in J2000 coordinates;

psfMag, the magnitudes in the different bands (ugriz);

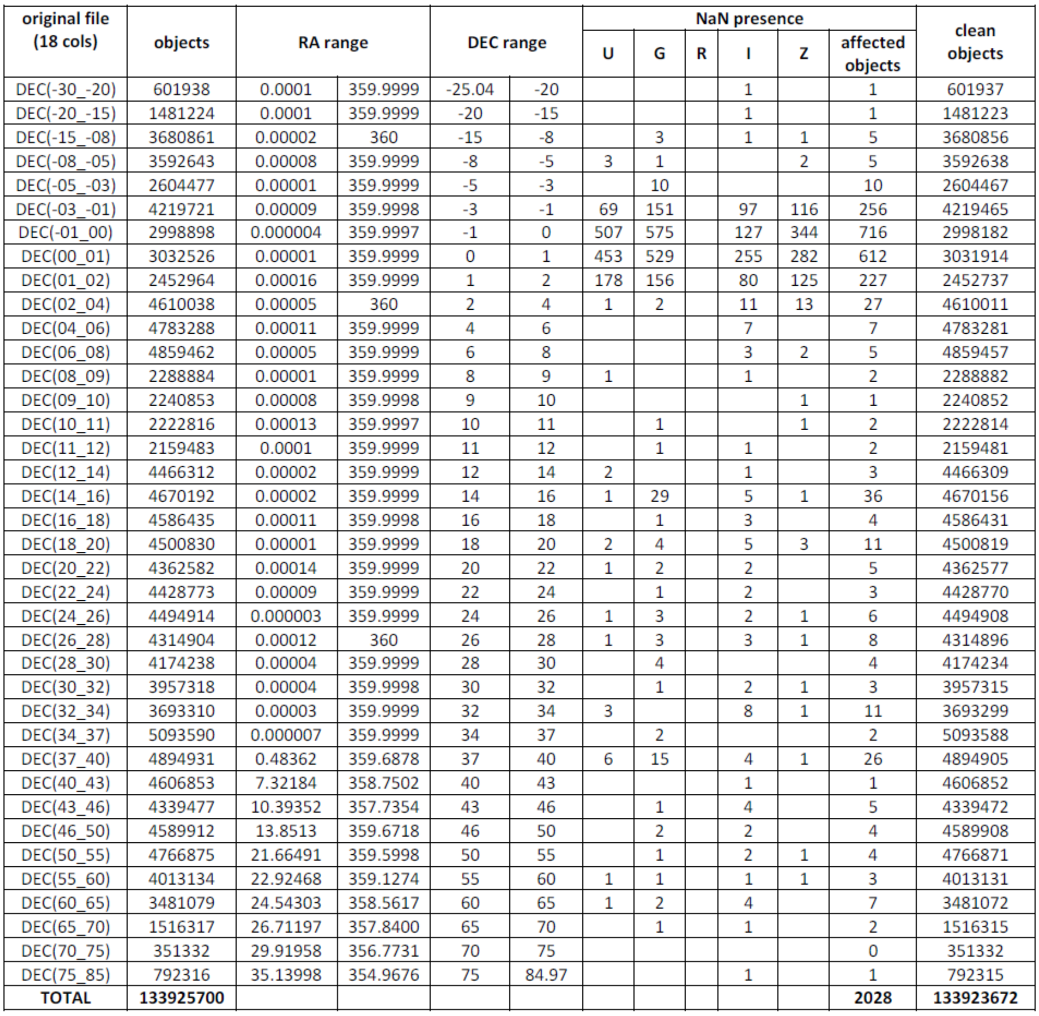
psfMagErr, the magnitude errors;

extinction, the extinction corrections in magnitudes at the position of each object.

All that parameters are selected from a specific table in the SDSS archive with the **FROM** command. **Galaxy** is a view that contains the photometric parameters (no redshifts or spectroscopic parameters) measured for resolved primary objects. In fact this view is derived from **PhotoPrimary** that contains only one primary object associated with each physical object on the sky; upon subsequent observations secondary objects are generated, but are excluded from PhotoPrimary view of the table. On the other hand **PhotoObjAll** contains all photo objects (Star, Galaxy, Sky and Unknown).
To limit the dimension of the outputs, for every query were selected little declination ranges to cover all the DR9 Sky Coverage Map. The **WHERE** command allow to take a selection using one parameter's value, in this case, the declination range of values. In addition there are two other constrains **clean = 1** limits searches to return only objects that have clean photometry, ensuring to have a good sample and **psfMag_r**< 23.6 puts a magnitude limit in R band.

2.1.3 Resulting Dataset

All the object resulting by the 38 queries are 133925700 within a declination range [-30°, +85°] and a magnitude R < 23.6.
A filter is applied to clean the dataset from the NaN objects, that represents the 0.0015% of the total number of available objects.
The resulting table informations are reported in Fig.2.3.



original file (18 cols)	objects	RA range		DEC range		NaN presence						clean objects
						U	G	R	I	Z	affected objects	
DEC(-30_-20)	601938	0.0001	359.9999	-25.04	-20				1		1	601937
DEC(-20_-15)	1481224	0.0001	359.9999	-20	-15				1		1	1481223
DEC(-15_-08)	3680861	0.00002	360	-15	-8		3		1	1	5	3680856
DEC(-08_-05)	3592643	0.00008	359.9999	-8	-5	3	1			2	5	3592638
DEC(-05_-03)	2604477	0.00001	359.9999	-5	-3		10				10	2604467
DEC(-03_-01)	4219721	0.00009	359.9998	-3	-1	69	151		97	116	256	4219465
DEC(-01_00)	2998898	0.000004	359.9997	-1	0	507	575		127	344	716	2998182
DEC(00_01)	3032526	0.00001	359.9999	0	1	453	529		255	282	612	3031914
DEC(01_02)	2452964	0.00016	359.9999	1	2	178	156		80	125	227	2452737
DEC(02_04)	4610038	0.00005	360	2	4	1	2		11	13	27	4610011
DEC(04_06)	4783288	0.00011	359.9999	4	6				7		7	4783281
DEC(06_08)	4859462	0.00005	359.9999	6	8				3	2	5	4859457
DEC(08_09)	2288884	0.00001	359.9999	8	9	1			1		2	2288882
DEC(09_10)	2240853	0.00008	359.9998	9	10					1	1	2240852
DEC(10_11)	2222816	0.00013	359.9997	10	11		1			1	2	2222814
DEC(11_12)	2159483	0.0001	359.9999	11	12		1		1		2	2159481
DEC(12_14)	4466312	0.00002	359.9999	12	14	2			1		3	4466309
DEC(14_16)	4670192	0.00002	359.9999	14	16	1	29		5	1	36	4670156
DEC(16_18)	4586435	0.00011	359.9998	16	18		1		3		4	4586431
DEC(18_20)	4500830	0.00001	359.9999	18	20	2	4		5	3	11	4500819
DEC(20_22)	4362582	0.00014	359.9999	20	22	1	2		2		5	4362577
DEC(22_24)	4428773	0.00009	359.9999	22	24		1		2		3	4428770
DEC(24_26)	4494914	0.000003	359.9999	24	26	1	3		2	1	6	4494908
DEC(26_28)	4314904	0.00012	360	26	28	1	3		3	1	8	4314896
DEC(28_30)	4174238	0.00004	359.9999	28	30		4				4	4174234
DEC(30_32)	3957318	0.00004	359.9998	30	32		1		2	1	3	3957315
DEC(32_34)	3693310	0.00003	359.9999	32	34	3			8	1	11	3693299
DEC(34_37)	5093590	0.000007	359.9999	34	37		2				2	5093588
DEC(37_40)	4894931	0.48362	359.6878	37	40	6	15		4	1	26	4894905
DEC(40_43)	4606853	7.32184	358.7502	40	43				1		1	4606852
DEC(43_46)	4339477	10.39352	357.7354	43	46		1		4		5	4339472
DEC(46_50)	4589912	13.8513	359.6718	46	50		2		2		4	4589908
DEC(50_55)	4766875	21.66491	359.5998	50	55		1		2	1	4	4766871
DEC(55_60)	4013134	22.92468	359.1274	55	60	1	1		1	1	3	4013131
DEC(60_65)	3481079	24.54303	358.5617	60	65	1	2		4		7	3481072
DEC(65_70)	1516317	26.71197	357.8400	65	70		1		1		2	1516315
DEC(70_75)	351332	29.91958	356.7731	70	75						0	351332
DEC(75_85)	792316	35.13998	354.9676	75	84.97				1		1	792315
TOTAL	**133925700**										**2028**	**133923672**

Figure 2.3: Information on original downloaded files. The last column reports the total number of objects containing various combinations of the detected NaN among their bands. The RA and DEC ranges are reported for the cleaned objects (i.e. after NaN removal).

Chapter 3

Data Mining in Astronomy

Scientific data mining is fundamental in AstroInformatic.
With the advent of synoptic sky surveys, which cover large areas of the sky there has been a great increase in the amount of available data. It is not just the data abundance that is fueling this ongoing revolution, but also Internet-enabled data access, and data re-use: in most cases, researchers who obtain the data can only extract a small fraction of the science that is enabled by it. Since physical understanding comes from the confrontation of experiment and theory, and both are now expressed as ever larger and more complex data sets, science is truly becoming data-driven in the ways that are both quantitatively and qualitatively different from the past.

The goal is to provide standards that describe all astronomical information resources worldwide and to enable standardized access to these collections. The astronomical community has responded to the complexity of this problem in the late 90s with the concept of a Virtual Observatory (VO): a web-based research environment for astronomy where there is a collection of interoperating data archives, information infrastructures and specific tools by which data can be analyzed. VO also supposed to facilitate the transition from old data poverty regime to the overwhelming data abundance; a number of national VOs are now active and are now federated through the International Virtual Observatory Alliance[1] (IVOA); in Italy, the VO is currently embodied as Italian Virtual Observatory[2].

The various astronomical catalogs, databases and observation logs have a large variety in schema, metadata, information content and knowledge representation. In most cases, the data are high-dimensional, thus requiring efficient and effective approaches (algorithms and data structures) for managing, mining, visualizing and analyzing high-dimension data sets.
The implementation of the VO framework over the past decade was focused on

[1] http://ivoa.net

[2] http://vobs.astro.it/

Name	Description
Simple Cone Search (SCS)	Retrieve all objects within a circular region on the sky
Simple Image Access (SIA)	Retrieve all images of objects within a region on the sky
Simple Spectral Access (SSA)	Retrieve all spectra of objects within a region on the sky
Simple Line Access (SLA)	Retrieve spectral line data
Simulations (SIMDAL)	Retrieve simulation data
Table Access (TAP)	Retrieve tabular data

Table 3.0.1: Different types of data access protocol defined by the International Virtual Observatory Alliance (IVOA).

the production of the necessary data infrastructure, interoperability, protocols and even a few useful data federation and analysis services. Although much still remains to be done, data discovery and access in astronomy have never been easier and the established infrastructure can at least in principle expand to the next generation of sky surveys and space missions

Even before the VO astronomers had already done very successful attempts toward standardization, for instance the fact that they adopted early universal standards for data exchange, such as the Flexible Image Transport System (FITS; Wells et al. 1981).

Within the VO a common set of data access protocols ensures that the same interface is employed across all data archives to perform the same type of data query (see Tab. 3.0.1 for a summary of those defined). Common data models define the shared elements across data and metadata collections and provide a framework for describing relationships between them, so different representations can interoperate in a transparent manner. When individual measurements of arbitrarily named quantities are reported, either as a group of parameters or in a table, their broader context within a standard data model can be established through the IVOA Utypes mechanism. These strings act as reference pointers to individual elements within a data model thus identifying the concept that the reported value represents.

Working with large amounts of data also requires proper infrastructure components: the VO provides a common interface **"VOSpace"** to the host of data storage solutions that are available, ranging in scale from a local filesystem on a laptop to a data farm in the cloud. It does not define how data is stored or transferred, only the control messages to gain access to data and manage data

flows, such as online analysis of large distributed data sets. Finally, the IVOA provides a Registry tool where descriptions of available data archives and services can be found, i.e. catalogs of white dwarfs or photometric redshift services.
The key to further progress is the availability of data exploration and analysis tools capable to operate on the Terascale data sets and beyond.

3.1 Machine Learning

"Knowledge Discovery in Databases" (KDD) or Data Mining (DM) regards the discovery of "models" for data. There are, however, many different methods which can be used to discover these underlying models: statistical pattern recognition, machine learning, summarization, etc.
Large data volumes tend to preclude direct human examination of all data and thus an automatization of these processes is needed, requiring use of Machine Learning (ML) techniques.

Machine learning is a scientific discipline concerned with the design and development of algorithms that allow computers to evolve behaviours based on empirical data. Usually what we can easily verify is not if a computer is able to learn, but mostly if it is able to give correct answers to specific questions.Indeed to verify that a machine gives correct answers to direct questions, used to train it, is only the preliminary step of its complete learning, because the crucial point is the machine behaviour in unpredicted situations, i.e. those never submitted to the machine during training. A "learner" can take advantage of examples (data) to capture characteristics of interest of their unknown underlying probability distribution. These data form the so called Knowledge Base (KB): a fairly large set of examples to be used for training and to test the performances. Hence the learner must possess some generalization capabilities in order to be able to produce useful outputs when it encountred new instances. So, if DM is the automatic (or semi-automatic) process of information discovery within massive data sets, ML has the following definition: *"a machine has learned if it is able to modify own behaviour in an autonomous way such that it can obtain the best performance in terms of answer to external stimuli"*Brescia (2012).

A large family of ML methods (the so called supervised ones) require the availability of relatively large and well characterized Knowledge Bases from which the ML methods can learn the underlying patterns and trends. In supervised ML we have a set of data points or observations for which we know the desired output, expressed in terms of categorical classes, numerical or logical variables or as generic observed description of any real problem. Finally, when the algorithm is able to correctly predict observations, we define it a classifier. Some classifiers are also capable of providing results with a regression giving a probability of

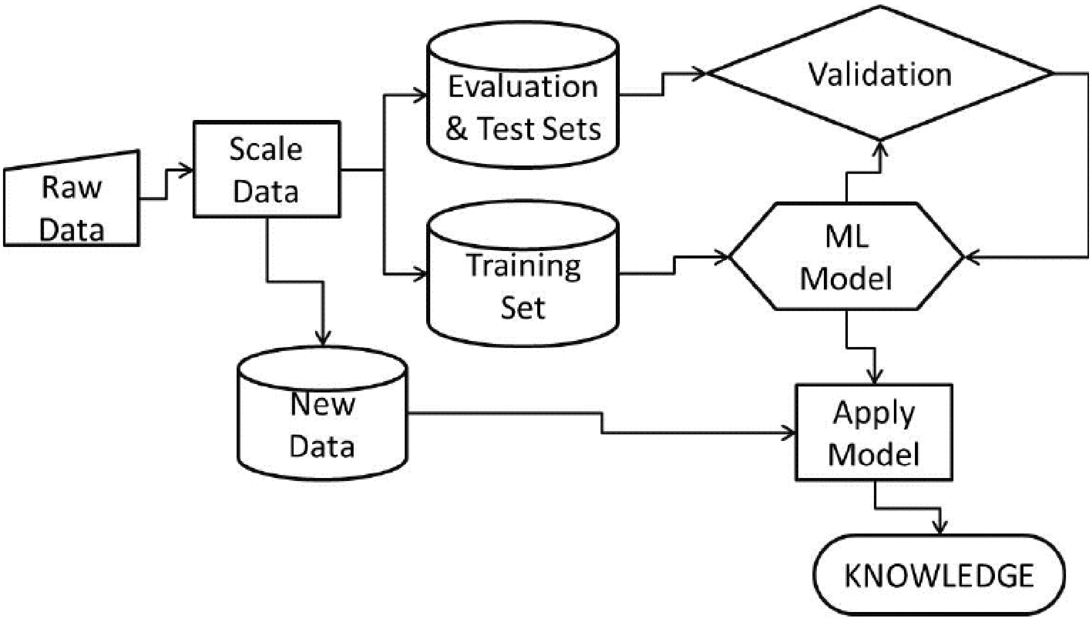


Figure 3.1: A workflow based on supervised learning models.

a data point belonging to class. We usually refer to such model behaviour as regression.

The supervised algorithm (see Fig. 3.1 can be described by the following steps:

1. **Pre-processing of data**: which includes scaling and preparation of data to built the input patterns.

2. **Creation of data sets for training and evaluation**: data are randomly splitted in a "training set" to learn their internal feature correlations and an "evaluation set" that is used to validate the already trained model in order to get an error rate (or other validation measures) that can help to identify the performance and accuracy of the classifier.

3. **Training of the model**: in this step the model is executed on the training data set. The output result consists of a model that (in the successful case) has learned how to predict the outcome when new unknown data are submitted.

4. **Validation**: to verify and measure the generalization capabilities of the model. The model is applied on new data, if the classification error of the validation set is higher than the training error, then we have to go back and adjust model parameters.

5. **Use**: if validation was successful, the model has correctly learned the underlying real problem. So far we can proceed to use the model to classify/predict new data.

For unsupervised algorithms there exist different problems. Indeed, instead of trying to predict a set of known classes, they are trying to identify the patterns inherent in the data that separate similar observations in one way or another. In other words, the main difference is that we are not providing a target variable like we did in supervised learning.
This marks a fundamental difference in how both types of algorithms operate. On one hand, we have supervised algorithms which try to minimize the error in classifying observations, while unsupervised learning algorithms don't have such gain, because there are no outcomes or target labels. Unsupervised algorithms try to create clusters of data that are inherently similar. In some cases we don't necessarily know what makes them similar, but the algorithms are capable of finding relationships between data points and group them in possible significant ways.
For unsupervised learning, the process follows these items:

1. **Pre-processing of data**, as with supervised learners;

2. **Execution of model training**, where the unsupervised algorithm is runned on the scaled data set to get groups of similar observations;

3. **Validation**, to verify if the data are clusterized in significant ways. This includes the calculation of a set of statistics on the resulting outcomes, as well as analysis based on domain knowledge.

Most ML algorithms used so far by the astronomers cannot deal well with missing data (i.e. no measurement was obtained for a given attribute) or with upper limits (a measurement was obtained, but there is no detection at some level of significance). While in many other fields this is only a minor problem, since the data are often redundant and can be cleaned of all records having incomplete or missing information, in astronomy all data recorded, including those with an incomplete information, are potentially scientifically interesting and cannot be ignored.

Examples of early uses of modern ML tools for analysis of massive astronomical data sets include automated classification of sources detected in sky surveys as stars (i.e., unresolved) vs. galaxies (resolved morphology), using Artificial Neural Nets (ANN, see Fig.3.2) or Decision Trees (DT) Weir et al. (1995). Brescia et al. (2012) have recently used several ML method for a different type of resolved/unresolved objects separation, namely the identification of globular clusters in external galaxies. Another set of ML applications is in classification

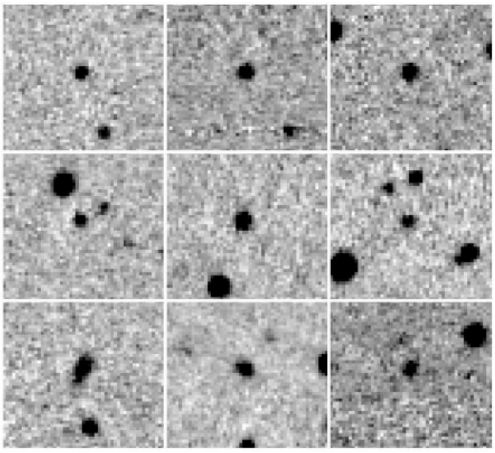


Figure 3.2: Examples of sources from the Palomar-Quest survey, classified using ANN techniques. Top row: sources classified as stars with a probability p* > 90%; bottom row: sources classified as galaxies, with p* < 10%; middle row: intermediate-classification sources with p*≈ 50%.

or selection of objects of a given type in some parameter space, for example colors that are the ratios of fluxes measured at different wavelengths. This is particularly well suited for the identification of quasars and other active galactic nuclei, which are morphologically indistinguishable from normal stars, but represent largely different physical phenomena.

This work of thesis is concentrated on another application of these methods: the estimation of the photometric redshift, that are derived from colors rather than from spectroscopy (much more costly in terms of the observing time). ANN have performed very well in this task (Tagliaferri et al. (2002); Firth et al. (2003); Hildebrandt et al. (2010); Cavuoti et al. (2012a)).

Classification is a procedure in which individual items are placed into groups based on quantitative information using the knowledge contained in a training set of previously labeled items (KB). Because of the supervised nature of the classification task, the system performance can be measured by means of a test set during the testing procedure, in which unseen data are given to the system to be labelled. Typical astrophysical problems which have been addressed with this functionality are the so called "star/galaxy" separation (which would be better called resolved-unresolved objects separation), morphological classification of galaxies, classification of stellar spectra, etc.

Regression is instead generally intended as the supervised search for a mapping from a domain in R^n to a domain in R^m, where $m < n$. Regression methods bring out relations between variables, especially whose relation is not surjective, i.e it has not one y for each given x. The most common astrophysical example of a regression problem is the evaluation of photometric redshifts of galaxies from a limited but statistically sufficient KB based on spectroscopic redshift samples.

Clustering is a division of data into groups of similar objects. Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification (Jain et al. (1999)). Clustering models are also referred to as unsupervised methods, since they do not require the use of an extensive KB. Clustering is often followed by a stage where a decision tree or a set of rules is inferred in order to allocate each instance to the cluster to which it belongs.

Dimensional reduction is the process of reducing the number of random variables under consideration and it can be divided into *feature selection* and *feature extraction*. Feature selection (Guyon & Elesseeff (2003)) approaches try to find a subset of the original variables by using filter (e.g. information gain) or wrapper (e.g. search guided by the accuracy) strategies. Feature extraction transforms the data in the high-dimensional space to a space of fewer dimensions (Guyon & Elesseeff (2006)).

As we have seen before, it is necessary to merge the capabilities of a file system to store and transmit bulk data from experiments, with logical organization of files into indexed data collections, allowing efficient query and analytical operations. It is also necessary to incorporate extensive metadata describing each experiment and the produced data.
The harder problem for the future is heterogeneity of platforms, data and applications, rather than simply the scale of the deployed resources. The goal should be to allow scientists to explore the data easily, with sufficient processing power for any desired algorithm to process it. But computing machines will not get much faster; they can be networked into clouds or grids of clusters and to perform tasks that were traditionally restricted to supercomputers at a fraction of the cost.

A first step in this direction is Service-Oriented Architectures (SOA) that supports reuse of both functionality and data in cross-organizational distributed computing settings Shadbolt et al. (2006). The fundamental characteristic of SOA infrastructures is the ability to locate and invoke a service across machine and organizational boundaries, both in a synchronous and an asynchronous manner. The implementation of a service can be achieved by wrapping legacy scientific application code and resource schedulers, which allows for a viable migration path Taylor (2007). The standards available for service design and their implementation support the rapid definition and execution of scientific workflows. With the advent of abstract machines, it is now possible to mix compilation and interpretation as well as integrate code written in different languages seamlessly into an application or service.

Most existing ML methods scale badly with both increasing number of records and of dimensionality so larger is the data set, more difficult is the analysis. So the training and validation of the methods are performed on these manageable data subsets, and the results are extended to the whole data set. This approach obviously may introduce biases difficult to control; typically, a lengthy fine tuning procedure is needed for such subset, which may require a lot of experiments to be performed in order to identify the optimal Data Mining method for a specific problem. DAMEWARE (DAta Mining & Exploration Web Application REsources; see Brescia et al. (2010)) resource was designed by taking all these issues into account.

Several Data Mining packages have been evaluated by Donalek et al. (2011), including Orange, Rapid Miner, Weka, VoStat and DAME.

In particular ***Data Mining and Exploration*** (DAME) web application[3] is a joint effort between the Astroinformatics groups at University Federico II, the Italian National Institute of Astrophysics and the California Institute of Technology. DAME offers a completely transparent architecture, a user-friendly interface and the possibility to seamlessly access a distributed computing infrastructure. It adopts VO standards in order to facilitate interoperability of data, although at the moment it is not yet fully VO compliant. This is partly due to the fact that new standards need to be defined for data analysis, DM methods and algorithm development. This implies a definition of standards in terms of an ontology and a well-defined taxonomy of functionalities to be applied to the astrophysical use cases. DAME offers asynchronous access to the infrastructure tools, thus allowing the running of jobs and processes outside the scope of any particular web application. The user, via a simple web browser, can access application resources and can keep track of his jobs by recovering related information (partial/complete results). Furthermore, DAME has been designed to run both on a server and on a distributed computing infrastructure (e.g. Grid or Cloud). A detailed technical description of the other components can be found in Brescia et al. (2010).

[3] `http://dame.dsf.unina.it` or `http://dame.caltech.edu/`

3.1.1 MLPQNA Method

In Ch.§4 the procedure used to evaluate photometric redshifts for a galaxy data set is described. The algorithm that I used for this evaluation has been realized for the integration in DAMEWARE tools (Cavuoti et al. (2012b)).

From a technical point of view, the MLPQNA method is a Multi Layer Perceptron (MLP; Bishop (2006)) implemented with a learning rule based on the Quasi Newton Algorithm (QNA); MLPQNA differs from more traditional MLPs implementations in the way the optimal solution of the regression problem is found.

The algorithm was involved in several experiments on astronomical datasets, both in regression (photometric redshift on galaxies and quasar) and classification (active galactic nuclei, globular clusters and transients) with remarkable results. According to Bishop (2006), feed forward neural networks in their various implementations provide a general framework for representing non linear functional mappings between a set of input variables (also called features) and a set of output variables (the targets).

The MLP architecture is one of the most typical feed-forward neural network model, in which the neurons are organized in layers, with proper own role. The term feed-forward is used to identify basic behaviour of such neural models, in which the impulse is propagated always in the same direction, from neuron input layer towards output layer, through one or more hidden layers, that represent the network brain, by combining weighted sum of weights associated to all neurons (except for the input layer). The input signal, simply propagated throughout the neurons of the input layer, is used to stimulate next hidden and output neuron layers. The output of each neuron is obtained by means of an activation function, applied to the weighted sum of its inputs.

What is different in such a neural network architecture is typically the learning algorithm used to train the network: the learning case approached with the MLP architecture is the supervised learning methods. In this case, the network must be firstly trained submitting the input patterns to the network as couples (input, desired known output). The feed-forward algorithm is then achieved and at the end of the input submission, the network output is compared with the corresponding desired output in order to quantify the learning quote. It is possible to perform the comparison in a batch way (after an entire input pattern set submission) or incremental (the comparison is done after each input pattern submission); also the metric used for the distance measure between desired and obtained outputs, can be chosen accordingly problem specific requirements (in the MLP-BP the Mean Square Error is used). After each comparison and until a desired error distance is unreached (typically the error tolerance is a precalculated value or a constant imposed by the user), the weights of hidden layers must be changed accordingly to a particular law or learning technique. After the training phase the network should be able not only to recognize correct output for each input already used as training set, but also to achieve a certain degree of generalization, to provide correct output for new inputs. The degree of generalization varies, as obvious, depending on how good has been the learning phase. This important feature is realized because the network does not associates a single input to the output, but it discovers the relationship present behind their association. After training, such a neural network can be seen as a black box able to perform a particular function (input-output correlation) whose analytical shape is a priori not known.

Bigger the training set, higher will be the network generalization capability. Despite of these considerations, it should been always taken into account that neural networks application field should be usually referred to problems where it is needed high flexibility more than high precision.

The training of a neural network can be seen as the search for the function which minimizes the errors of the predicted values with respect to the true values available for a small but significant subsample of objects in the same data set. This subset is also called training set or knowledge base.

The formal description of a feed-forward neural network with two computational layers is given in the Eq. 3.1 (where (1) and (2) are the first layer and the second one):

$$y_k = \sum_{j=0}^{M} w_{kj}^{(2)} g \left(\sum_{i=0}^{d} w_{ji}^{(1)} x_i \right) \tag{3.1}$$

Which can be better understood by using the graph shown in Fig. 3.3. The input layer (x_i) is made of a number of neurons (also known as perceptrons) equal to the number of input variables (d); the output layer, on the other hand, will have as many neurons as the output variables (k). In the general case, the network may have an arbitrary number of hidden layers each of one can be formed by an arbitrary number of neurons (M); in the depicted case there is just one hidden layer as in most real implementations. 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, w_{kj}, which can be regarded as the the strength of the synaptic connection between neurons k and j, while the response of each perceptron to the inputs is represented by a non-linear function g, referred to as the activation function.

All the above characteristics, the topology of the network and the weight matrix of its connections, define a specific implementation and are usually called **model**. The model, however, is only part of the story. In fact, in order to find the model that best fits the data in a specific problem, one has to provide the network with a set of examples. These data form the so called training set or Knowledge Base

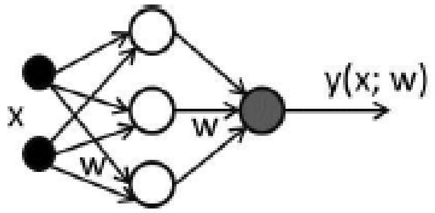


Figure 3.3: Scheme of a Multi Layer Perceptron general architecture for two input variables and one output value.

(KB) and through a learning rule are used by the network to find the optimal model.

The MLPQNA use for learning rule the Quasi Newton Algorithm (QNA), which differs from the Newton Algorithm in how the Hessian of the error function is computed. Newtonian models are variable metric methods used to find local maxima and minima of functions (Davidon (1968)) and, in the case of MLPs, they can be used to find the stationary point of the learning function.

The Quasi Newton Learning Rule

Most Newton methods use the Hessian of the function to find the stationary point of a quadratic form. It needs to be stressed, however, that the Hessian of a function is not always available and in many cases it is far too complex to be computed in an analytical way. More often it is easier to compute the function gradient which can be used to approximate the Hessian via N consequent gradient calculations. In order to understand the importance of QNA it is needed to start from the classical and quite common Gradient Descent Algorithm (GDA) used for Back Propagation Bishop (2006). In GDA, the direction of each updating step for the MLP weights is derived from the error descent gradient, while the length of the step is determined from the learning rate. This method is inaccurate and ineffective and therefore may get stuck in local minima. A more effective approach is to move towards the negative direction of the gradient (line search direction) not by a fixed step, but by moving towards the minimum of the function along that direction. This can be achieved by first deriving the descent gradient and then by analyzing it with the variation of the learning rate Brescia (2012). Let us suppose that at step t, the current weight vector is $w^{(t)}$, and let us consider a search direction $d^{(t)} = -\nabla E^{(t)}$ (the gradient of the error function). If we select the parameter λ in order to minimize $E(\lambda) = E(w^{(t)} + \lambda d^{(t)})$, the new weight vector can be expressed as:

$$w^{(t+1)} = w^{(t)} + \lambda d^{(t)} \tag{3.2}$$

and the problem of line search becomes a 1-dimensional minimization problem which can be solved in many different ways. Simple variants are:

- to move $E(\lambda)$ by varying λ by small intervals, then evaluate the error function at each new position and stop when the error begins to increase;
- to use the parabolic search for a minimum and compute the parabolic curve crossing pre-defined learning rate points.

The minimum d of the parabolic curve is a good approximation of the minimum of $E(\lambda)$ and it can be derived by means of the parabolic curve which crosses the fixed points with the lowest error values.

Another approach makes instead use of *trust region* based strategies which minimize the error function, by iteratively growing or contracting the region of the function by adjusting a quadratic model function which best approximates the error function.

All these approaches, however, rely on the assumption that the optimal search direction is given at each step by the negative gradient; this is not always true, but can also lead to serious wrong convergence. Indeed, if the minimization is done along the negative gradient direction, the subsequent search direction (the new gradient) will be orthogonal to the previous one: when the line search founds the minimum, we have:

$$\frac{\partial E}{\partial \lambda}(w^{(t)} + \lambda d^{(t)}) = 0 \tag{3.3}$$

and hence,

$$g^{(t+1)T} d^{(t)} = 0 \tag{3.4}$$

where $g \equiv \nabla E$. The iteration of the process therefore leads to oscillations of the error function which slow down the convergence process.

The method implemented here relies on selecting other directions so that the gradient component, parallel to the previous search direction, would remain unchanged at each step. Suppose that you have already minimized with respect to the direction $d^{(t)}$ starting from the point $w^{(t)}$ and reaching the point $w^{(t+1)}$; in this point Eq. 3.4 becomes

$$g(w^{(t+1)})^T d^{(t)} = 0 \tag{3.5}$$

Choosing $d^{(t+1)}$ to preserve the gradient component parallel to $d^{(t)}$ equal to zero, it is possible to build a sequence of directions d in such a way that each direction is conjugated to the previous one on the dimension $\mid w \mid$ of the search space (this is known as *conjugate gradients method*; Golub & Ye (1999)).

With a squared error function, the update weights algorithm is

$$w^{(t+1)} = w^{(t)} + \alpha^{(t)} d^{(t)} \tag{3.6}$$

where

$$\alpha^{(t)} = -\frac{d^{(t)T} g^{(t)}}{d^{(t)T} H d^{(t)}} \tag{3.7}$$

Furthermore, d can be obtained for the first time via the negative gradient and in the subsequent iterations, as a linear combination of the current gradient and of the previous search directions

$$d^{(t+1)} = -g^{(t+1)} + \beta^{(t)} d^{(t)} \tag{3.8}$$

where

$$\beta^{(t)} = \frac{g^{(t+1)T} H d^{(t)}}{d^{(t)T} H d^{(t)}} \tag{3.9}$$

This algorithm finds the minimum of a square error function at most in $\mid w \mid$ steps but with a high computational cost, since in order to determine the values of α and β, it makes use of that *hessian matrix H* which, as we already mentioned, is very demanding in terms of computing time: this puts serious constraints on the application of this family of methods to large data sets. Excellent approximations for the coefficients α and β can, however, be obtained from analytical expressions that do not use the Hessian matrix explicitly.

For instance, β can be calculated through any one of the following expressions (respectively Hestenes & Stiefel (1952); Fletcher & Reeves (1964); Polak & Ribiere (1969)):

$$\begin{aligned} Hestenes - Sitefel \quad &: \quad \beta^{(t)} = \frac{g^{(t+1)T}(g^{(t+1)} - g^{(t)})}{d^{(t)T}(g^{(t+1)} - g^{(t)})} \\ Fletcher - Reeves \quad &: \quad \beta^{(t)} = \frac{g^{(t+1)T} g^{(t+1)}}{g^{(t)T} g^{(t)}} \\ Polak - Ribiere \quad &: \quad \beta^{(t)} = \frac{g^{(t+1)T}(g^{(t+1)} - g^{(t)})}{g^{(t)T} g^{(t)}} \end{aligned} \tag{3.10}$$

These expressions are all equivalent if the error function is square-typed, otherwise they assume different values. Typically the *Polak-Ribiere* equation obtains better results because, if the algorithm is slow and subsequent gradients are quite alike between them, its equation produces values of β such that the search direction tends to assume the negative gradient direction.

Concerning the parameter α, its value can be obtained by using the *line search method* directly. The method of conjugate gradients reduces the number of steps to minimize the error up to a maximum of $\mid w \mid$, because there could be almost $\mid w \mid$ conjugate directions in a $\mid w \mid$-dimensional space. In practice however, the algorithm is slower because, during the learning process, the property *conjugate* of the search directions tends to deteriorate. To avoid the deterioration, to

restart the algorithm after $\mid w \mid$ steps is needed, by resetting the search direction with the negative gradient direction.
By using a local square approximation of the error function, we can obtain an expression for the minimum position. The gradient in every point w is given by

$$\nabla E = H \times (w - w^*) \tag{3.11}$$

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

$$w^* = w - H^{-1} \times \nabla E \tag{3.12}$$

The vector $(-H^{-1} \times \nabla E)$ is known as *Newton direction* and it is the base for a variety of optimization strategies, such as for instance the QNA, which instead of calculating the H matrix and then its inverse, uses a series of intermediate steps of lower computational cost to generate a sequence of matrices which are more and more accurate approximations of H^{-1}. From the Newton Eq. 3.12 we note that the weight vectors on steps t and $t+1$ are correlated to the correspondent gradients by the formula

$$w^{(t+1)} - w^{(t)} = -H^{(-1)}(g^{(t+1)} - g^{(t)}) \tag{3.13}$$

which is known as *Quasi Newton Condition*. The approximation G is therefore built in order to satisfy this condition as

$$G^{(t+1)} = G^{(t)} + \frac{pp^T}{p^T\nu} - \frac{(G^{(t)}\nu)\nu^T G^{(t)}}{\nu^T G^{(t)}\nu} + (\nu^T G^{(t)}\nu)uu^T \tag{3.14}$$

where

$$\begin{aligned} p &= w^{(t+1)} - w^{(t)}; \\ \nu &= g^{(t+1)} - g^{(t)}; \\ u &= \frac{p}{p^T\nu} - \frac{G^{(t)}\nu}{\nu^T G^{(t)}\nu} \end{aligned}$$

The above expression could carry the search out of the interval of validity for the squared approximation. The solution is hence to use the line search to found the minimum of function along the search direction. By using such system, the weight updating expression (Eq. 3.6) can be formulated as follows

$$w^{(t+1)} = w^{(t)} + \alpha^{(t)} G^{(T)} g^{(t)} \tag{3.15}$$

where α is obtained by the *line search*.

One of the main advantage of QNA, compared with conjugate gradients, is that the *line search* does not require the calculation of α with a high precision, because it is not a critical parameter. Unfortunately it requires a large amount of memory to calculate the matrix $G(\mid w \mid \times \mid w \mid)$, for large $\mid w \mid$. One way to reduce the required memory is to replace at each step the matrix G with a unitary matrix and to multiply by g (the current gradient), to obtain

$$d^{(t+1)} = -g^{(t)} + Ap + B\nu \tag{3.16}$$

Note that if the line search returns exact values, then the above equation produces mutually conjugate directions, where A and B are scalar values defined as:

$$\begin{aligned} A &= -(1 + \frac{\nu^T \nu}{p^T \nu}) \frac{p^T g^{(t+1)}}{p^T \nu} && (3.17) \\ B &= \frac{p^T g^{(t+1)}}{p^T \nu} && (3.18) \end{aligned}$$

Chapter 4

PhotoRApToR

In the previous chapters the relevance of photometric redshifts has been discussed. Different methods to evaluate photometric redshift from a dataset have been presented, but this thesis is focused on the MLPQNA empiric method, described in chapter 3.
Now it is important to describe the experimental part of this thesis: the creation of a specific desktop tool to evaluate photo-z using the MLPQNA method. Due to the necessity to evaluate redshift for huge sky survey datasets, it seemed important to provide the astronomical community with an instrument able to fill this gap.
The problem is that a great part of astronomical data is stored in private archives that are not accessible on line. So, in order to evaluate photo-z it is needed a desktop application that can be used by everyone on its own personal computer. The name choosen for this application was **PhotoRApToR**, i.e. **Photo**metric **R**esearch **Ap**plication **To** **R**edshift.
Java is an object-oriented computer programming language with an implemetation available for Mac OS X, Windows and Linux, so in order to favour a large diffusion of this tool, the code was developed in Java language using NetBeans IDE 7.3. By means of the Swing libraries also a Graphic User Interface (GUI) was realized for a more simple data analysis and the PhotoRApToR project has been completed with a Primer Wizard: a tutorial dialog for beginners.

In the next sections, after a brief description of Java language and programming instrument used to realize this project, I describe all the PhotoRApToR's functions and in Ch.§5 I shall present the results from the application of this tool on real data from SDSS (see Ch.§2) are shown.

4.1 Java language

Java was originally developed by James Gosling at Sun Microsystems (which has since merged into Oracle Corporation) and released in 1995 as a core component of Sun Microsystems' Java platform. James Gosling, Mike Sheridan, and Patrick Naughton initiated the Java language project in June 1991. Java was originally designed for interactive television, but it was too advanced for the digital cable television industry at the time. The language derives much of its syntax from C and C++, although some differences exist.

From the beginning, Java was designed to be a platform neutral language and this is one of the principal motivation to choose this language for an application created for a wide community.

Java was not originally intended to directly generate code that operate on a specific platform: applications are compiled to *bytecode*, the form of instructions that the Java Virtual Machine executes. The Java Virtual Machine is a program often implemented to run on an existing operating system, but can also be implemented to run directly on hardware; its specification gives the rules by which bytecodes must be interpreted.

Commonly used operating systems have a JVM implemented, so there are few platforms where Java programs cannot be executed.

Java is one of the most popular programming languages in use, particularly for client-server web applications. On November 13, 2006, Sun released much of Java as free and open source software under the terms of the GNU General Public License (GPL). On May 8, 2007, Sun finished the process, making all of Java's core code available under free software/open-source distribution terms, aside from a small portion of code to which Sun did not hold the copyright.

4.1.1 Java GUI toolkit

Java provided a mechanism knows as the Abstract Window Toolkit (**AWT**) that contained a set of classes that enabled the construction of GUI objects such as buttons, scroll bars or windows. In AWT, each component is rendered and controlled by a native peer component specific to the underlying windowing system.

In 1997 the Java Foundation Classes (JFC), a graphical framework for building portable Java-based graphical user interfaces, was created by Sun Microsystems and Netscape Communications Corporation to provide a wide set of graphical components. Wherever possible, compatibility was preserved between JFC components and Abstract Window Toolkit.

The "Java Foundation Classes" were later renamed "**Swing**", adding the capability for a pluggable *look and feel* of the widgets: the way in which visual components are rendered. This allowed Swing programs to maintain a platform-

independent code base and mimic the look of a native application. Now the Swing toolkit has totally replaced the AWT's widgets also drawing its own widgets.
Swing components are implemented as *lightweight*: application controls that are implemented in Java without a corresponding native peer. Lightweight controls do not have a peer entity in the operating system to manage the data, state and appearance of a control. In this way, Swing components do not paint themselves: an user interface manager directs paint requests to a delegate object. In addition, unlike native platform components, Swing components are engineered to provide object-oriented access to the control's data and state.

4.1.2 NetBeans IDE 7.3

NetBeans IDE is an Integrated Development Environment (IDE) to write, compile and debug software applications for Java platform and other environments. It includes many features, as text editor, visual design tools or source code management support. NetBeans IDE is written in Java and can run on Windows, OS X, Linux, Solaris and other platforms supporting a compatible JVM.
IDE is a software application to help computer programmers for software development: several modern IDEs use Intelli-sense coding features[1]. Many IDEs have various tools to simplify the construction of a GUI; sometimes a compiler and an interpreter are present, such as Net Beans and Eclipse.

NetBeans IDE 7.3 was released in February 2013. All the functions of the IDE are provided by modules; NetBeans contains all the modules needed for Java development, but new features, such as support for other programming languages, can be added by installing additional modules.

Two basis modules are the following:

1. **NetBeans Profiler** is a tool for the monitoring of Java applications: it helps developers to find memory leaks and optimize speed (it is integrated into the core IDE since version 6.0);

2. **GUI design-tool** (formerly Project Matisse) enables developers to prototype and design Swing GUIs by dragging and positioning GUI components. The GUI builder automatically takes care of the correct spacing and alignment.

[1] In computer programming means *intelligent code sense*, a programming environment that speeds up the process of coding applications by reducing common mistakes, usually through auto completion popups when typing, querying parameters of functions, ecc.

4.2 Primer Wizard

When the program is launched, in addition to the main program window, also a tutorial dialog is started.

1. The first dialog explains scientifical applications of the program and gives the possibility to skip the tutorial to the main program.

2. In the second dialog it is possible to open table data (selectable choices: ASCII, FitsTable, CSV, VoTable). During this operation, the Wizard verifies the correspondence between table extension and the allowed choices. After this check, columns names become visible.

3. In the third dialog it is possible to manipulate tables headers to select only needed columns by a checkbox. After this we can separate our data into two files (TRAIN and TEST) using the Split function.

4. In the fourth dialog the experiment setup begins. With a checkbox we can select between two options: Classification or Regression, while a drop-down menu allows to choose between TRAIN and TEST.

5. The fifth dialog is different depending on the chosen experiment and allows to insert the parameters to setup the experiment. Clicking Run button, the experiment starts and a popup window shows the running processes.

6. This is the final dialog. Here we can see the output table with its path directory. Clicking on the path link, we open the table in a different dialog, while, clicking on the PLOT button, in a different dialog we can see a scatter plot zphot/zspec.

4.3 GUI Description

The main window of the PhotoRApToR application (see Fig. 4.1) is divided in three parts.
The first one is the *Menu Bar* with a *Button Bar* below, the second one is the *Table List* on the left and the third one is the panel on the right with *Table Properties*, *Table Editor* and the *Split* panel below.

Beginnig from the *Menu Bar*, it is possible to decribe all the commands:

File is the menu from which to launch standard commands to open or save files. The following options are shown:

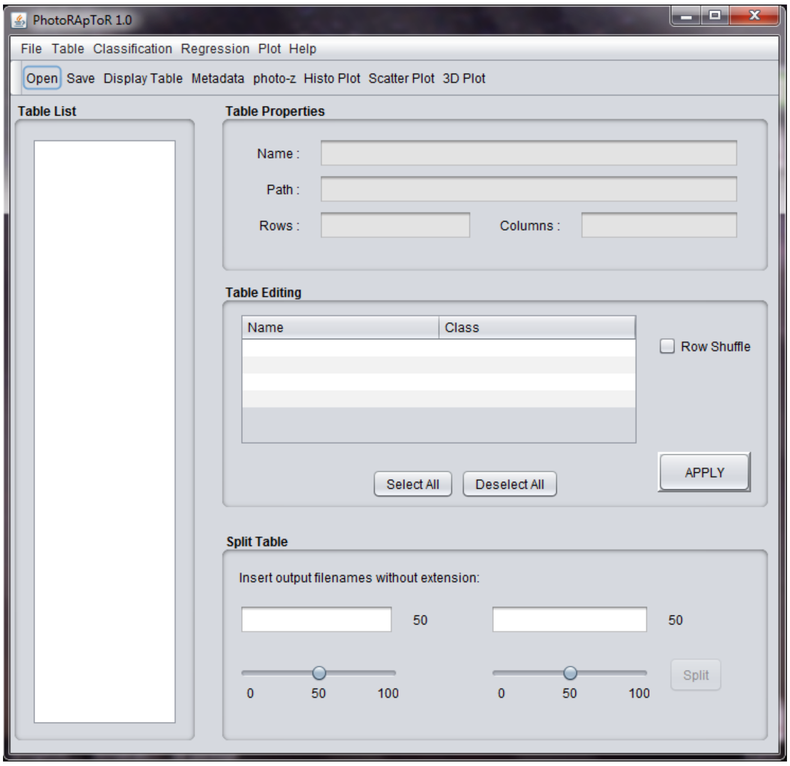

PhotoRApToR 1.0
File Table Classification Regression Plot Help
Open Save Display Table Metadata photo-z Histo Plot Scatter Plot 3D Plot
Table List
Table Properties
Name :
Path :
Rows :
Columns :
Table Editing
Name
Class
Row Shuffle
APPLY
Select All
Deselect All
Split Table
Insert output filenames without extension:
50
50
0
50
100
0
50
100
Split

Figure 4.1: PhotoRApToR main window.

Load Table : opens a new dialog where it is possible to select table format and file;

Discard Table : allows to erase a table item from the *Table List*;

Save Table : saves the selected table using a Browse Dialog;

Exit

Table is a menu containing the commands that allow to see and modify the table properties:

Table Data : opens the selected table in a new window;

Table Metadata : shows only the column's metadata for the selected table;

Row Shuffle : the selected table rows are shuffled and the new table is opened in a new window;

Not a Number opens a new window for managing the table data in order to remove the Not a Number elements (e.g. -9999) from the dataset;

Classification is a menu where are selectable different options that allow to run different experiments:

each one of the options **Train, Test** or **Run**, it opens a new window where it is possible to set the experiment parameters necessary to use MLPQNA algorithm;

Regression is a menu witn other experimental options:

the options **Train, Test** and **Run** are like those in the Classification menu;

Statistics: in a new window the user can select the Target column and the Output column to generate statistics;

Outliers: opens a new window where the user can generate a dataset without outliers by setting statistical parameters;

Plot is the menu that shows three different ways to generate data plots:

Histo Plot : opens a new window where to select which column of the table we want to plot;

Scatter Plot : it has some more parameters to set up, like for instance the type of line plot or the marker for the data points;

3D Plot : after the selection of three table columns and the setting of parameters, a cube plot is generated with also the possibility of rotating the viewing angle;

Help is the menu with manuals and program's credits

Help : opens a program description with the manual and the template use cases for beginners;

Open Wizard : starts the Primer Wizard window;

About : credits and collaboration's links are reported

The *Button Menu* below allows a fast access to the main functions of the application.

Open is a button that opens a dialog for the selection of table format and the browsing for the file (see Fig. 4.2);

Save opens a dialog to save the table;

Display Table shows the whole table dataset in a new window;

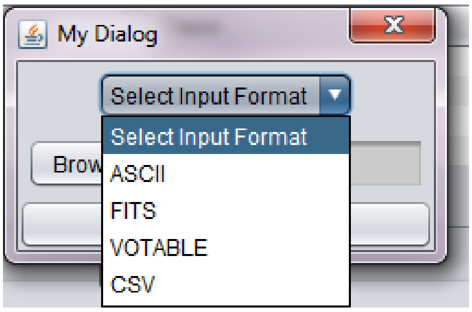


Figure 4.2: Load Table Dialog.

Metadata displays only the metadata of the columns;

photo-z launchs the main experiment that allows the evaluation of photometric redshift in a new window: sets MLPQNA parameters and generates an output table and a file containing the relevant statistics;

Histo Plot works like the menu item with the same name;

Scatter Plot works like the menu item with the same name;

3D Plot works like the menu item with the same name;

The first step is to open a table selecting the format file and browsing with the Load Dialog (Fig. 4.2). For this description a file named "**demonstration.fits**" was choosen: this file contains a little sample of the objects classified as galaxies in the SDSS Data Release 9 described in the Ch.§2.
Every time a new table is loaded, a new item with the table name is added to the *Table List*: a double click opens a new window showing the complete table. Selecting one item from the *Table List*, all the table properties are displayed inside the right panel: in particular the table name, its complete path and the number of columns and rows. Below, in the *Edit Table* panel, are displayed column metadata for the selected table and it is possible to select a subset of the table by choosing only the needed columns. After the selection, a table subset is created by clicking the **Create Subset** button and, if the selectable checkbox **Row Shuffle** is selected, the subset table is also shuffled by rows.

The last panel on the right of the main window is the *Split* panel. PhotoRApToR is an application of the MLPQNA method described in the Ch.§3 and before launching an experiment, it may be necessary to split the dataset in a TRAIN subset and in a TEST subset. This is a simple action made possible by the Split tool. When the table is selected in the *Table List*, we must choose

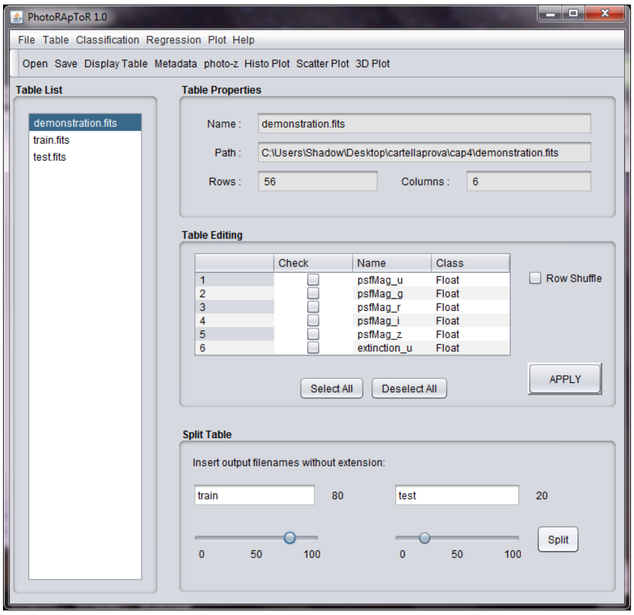

PhotoRApToR 1.0
File Table Classification Regression Plot Help
Open Save Display Table Metadata photo-z Histo Plot Scatter Plot 3D Plot
Select All
Deselect All
APPLY
Split
0 50 100
0 50 100

Table List

- demonstration.fits
- train.fits
- test.fits

Table Properties

Name : demonstration.fits

Path : C:\Users\Shadow\Desktop\cartellaprova\cap4\demonstration.fits

Rows : 56 Columns : 6

Table Editing

	Check	Name	Class
1	□	psfMag_u	Float
2	□	psfMag_g	Float
3	□	psfMag_r	Float
4	□	psfMag_i	Float
5	□	psfMag_z	Float
6	□	extinction_u	Float

□ Row Shuffle

Split Table

Insert output filenames without extension:

train 80

test 20

Figure 4.3: Use of the Split tool. After selecting the table to be split, two different names are choosen for the files and the sliders are dragged to select a different percentage. Clicking on Split button generate the two split datasets.

two different names for the split files (in this case "train" and "test") and two different percentages of the original dataset. Clicking on Split button the two split datasets are generated and added to the *Table List* (Fig. 4.3), ready for the next phase.

4.3.1 Experiment Parameters

As described before, PhotoRApToR allows to select different options for the MLPQNA parameters. The photo-z evaluation is a particular case of regression experiments. The complete description of photo-z evaluation on SDSS data is presented in Ch.§5, whereas in this section are described the **Regression** and **Classification** experiment window.
A click on **Regression**− >**Train** menu item opens a new window (Fig. 4.4) where it is necessary to set MLPQNA's input parameters.

- A drop-down menu allows to select the input file; (**this parameter is a field required**)

- if we had already done the training phase, it is possible to use the **trained weight file**;

- the **Number of input neurons** is the number of input dataset columns (except for the target column); (**this parameter is a field required**)

- the **Number of first hidden layer neurons** is the number of neurons of the first hidden layer of the network; (**this parameter is a field required**)

- the **Number of second hidden layer neurons**, as a suggestion this number should be selected smaller than the previous layer. By default the second hidden layer is empty (not used);

- **Max number of iteration** is one of the internal model parameters. It indicates the number of algorithm iterations and it is one of the stopping criteria. By default this value is set to 1500;

- **Hessian approximation cycles** indicates the number of restarts for each approximation step of the Hessian inverse matrix. By default this value is set to 20;

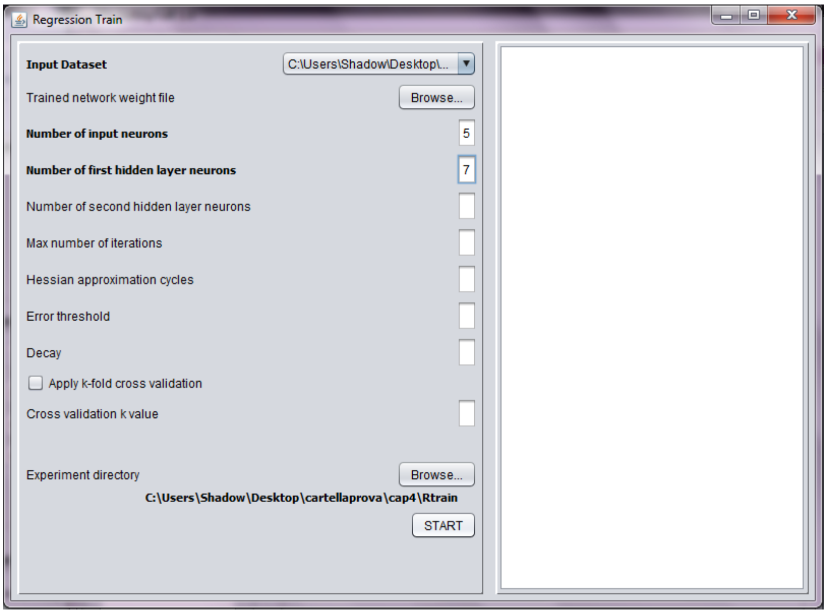
- **Error threshold** indicates the minimum weight error at each iteration step. Except for problems which are particularly difficult to solve, in which a value of 0.0001 should be used, a value of 0.01 is usually considered sufficient. By default this value is therefore set to 0.01;

- **Decay** indicates the weight regularization decay. If accurately chosen, this parameter leads to an important improvements of the generalization error of the trained neural network and implies an acceleration of training. By default the value is set to 0.001;

- **Cross validation** is based on an automatic procedure that splits in different subsets the training dataset, applying a k step cycle in which the training error is evaluated and its performances are validated. By default the k value is set to 10;

- finally **Experiment output directory** is the parent directory of the output for the experiments.

After the parameters setup, a click on *START* button launchs the MLPQNA regression experiment and the resulting output is displayed in the main panel on the left. After the experiment, also the statistics is generated with a specific algorithm and the result is presented in the text panel on the top of the panel. To complete the description of the experimental use of PhotoRApToR, by clicking on **Classification**− >**Train** it opens a new window (Fig. 4.5) for the MLPQNA's parameters setting.



Regression Train

Input Dataset C:\Users\Shadow\Desktop\...

Trained network weight file Browse...

Number of input neurons 5

Number of first hidden layer neurons 7

Number of second hidden layer neurons

Max number of iterations

Hessian approximation cycles

Error threshold

Decay

□ Apply k-fold cross validation

Cross validation k value

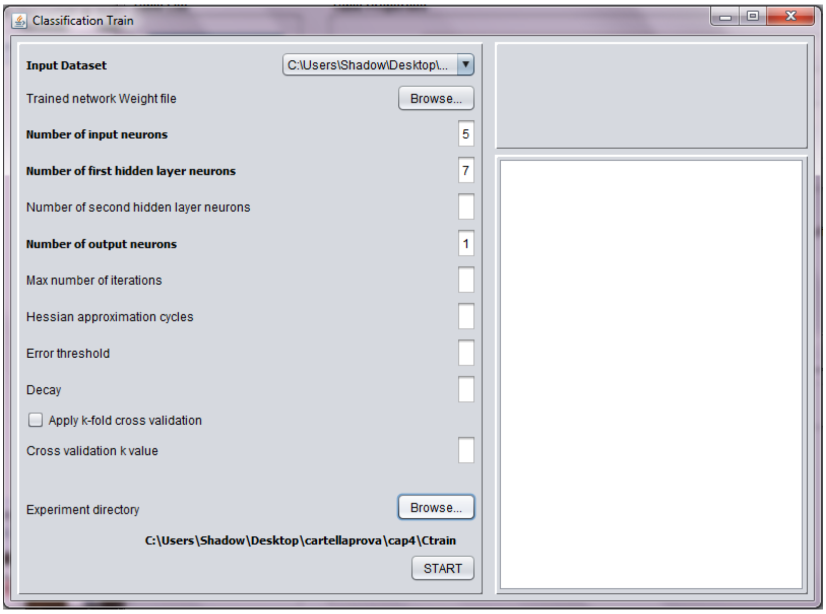
Experiment directory Browse...

C:\Users\Shadow\Desktop\cartellaprova\cap4\Rtrain

START

Figure 4.4: Regression Train window. On the left there are the fields for setting parameters. On the right there are two panels. When the experiment is complete, in the upper panel the regression train statistics is displayed. In the lower panel the final table with the photo-z column is reported.

- A drop-down menu allows to select the input file; (**this parameter is a field required**)
- if we had already done the training phase, it is possible to use the **trained weight file**;
- **Number of input neurons:** as above; (**this parameter is a field required**)
- **Number of first hidden layer neurons:** as above; (**this parameter is a field required**)
- **Number of second hidden layer neurons:** as above;
- the **Number of output neurons** is the number of neurons in the output layer of the network. It must correspond to the number of target columns in the input file; (**this parameter is a field required**)
- **Max number of iteration:** as above;
- **Hessian approximation cycles:** as above;



Input Dataset C:\Users\Shadow\Desktop\...

Trained network Weight file

Number of input neurons 5

Number of first hidden layer neurons 7

Number of second hidden layer neurons

Number of output neurons 1

Max number of iterations

Hessian approximation cycles

Error threshold

Decay

☐ Apply k-fold cross validation

Cross validation k value

Experiment directory

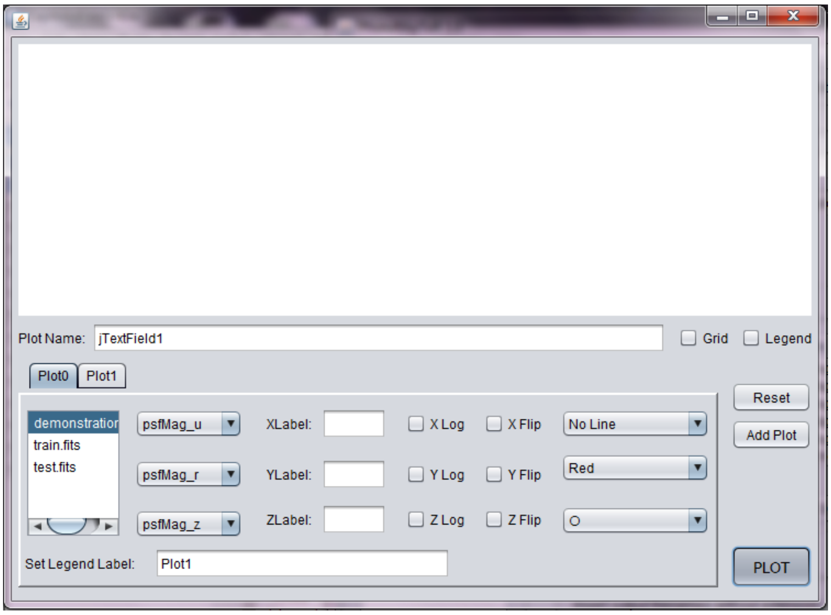
C:\Users\Shadow\Desktop\cartellaprova\cap4\Ctrain

Figure 4.5: Classification Train window. On the left there are the fields for setting parameters. On the right there are two panels. When the experiment is complete, in the upper panel the Confusion Matrix is displayed. In the lower panel the final table with the photo-z column is reported.

- **Error threshold:** as above;
- **Decay:** as above;
- **Cross validation:** as above;
- finally **Experiment output directory** is the parent directory of the output for the experiments.

A click on the *START* button launchs the MLPQNA classification experiment and the resulting output is displayed in the main panel on the left. The text panel above the *Confusion Matrix* is reported. By clicking on **Test** or **Run** options of *Regression* and *Classification* menu items, it opens a window similar to those described for the **Train** case.

After every regression experiment, a statistical report is automatically generated; clicking on the menu item **Experiment**$-$ >**Statistics** opens a new window where is possible to select a *Target* column and an *Output* column between which the algorithm generates statistical indicators.



Plot Name: jTextField1 □ Grid □ Legend

demonstration
train.fits
test.fits

psfMag_u XLabel: □ X Log □ X Flip No Line

psfMag_r YLabel: □ Y Log □ Y Flip Red

psfMag_z ZLabel: □ Z Log □ Z Flip O

Set Legend Label: Plot1

Figure 4.6: Plot window.

These indicators are:

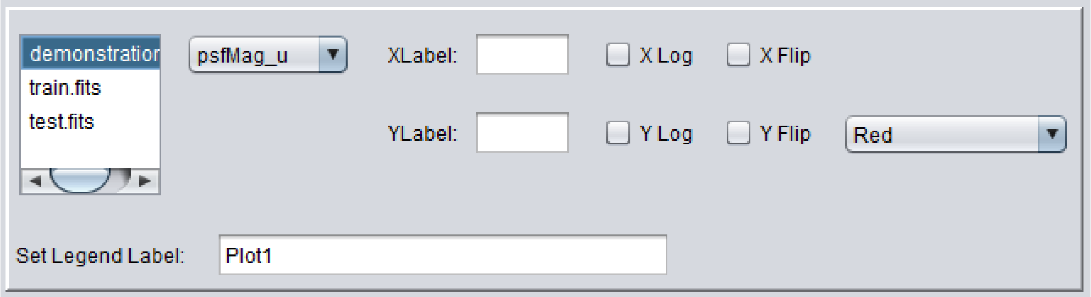
$$
\begin{aligned}
bias(x) &= \frac{\sum_{i=1}^{N} x_i}{N} \\
\sigma(x) &= \sqrt{\frac{\sum_{i=1}^{N}\left[x_i - \left(\frac{\sum_{i=1}^{N} x_i}{N}\right)\right]^2}{N}} \\
RMS(x) &= \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N}} \\
MAD(x) &= Median(\mid x \mid) \\
NMAD(x) &= 1.48 \times Median(\mid x \mid)
\end{aligned}
$$

where σ is the *Standard Deviation*, RMS is the *Root Mean Square*, MAD the *Median Absolute Deviation* and NMAD the normalized MAD.

4.3.2 Plot Selection

As described before, in the PhotoRApToR's menus are present also instruments to generate different types of plot. When one of the plot options (Histo Plot, Scatter Plot or 3D Plot) is clicked, a new window (Fig.4.6) opens where to set the plot parameters: in the upper panel will be displayed the plot, below there

are a text field where is possible to set the name of the plot and two checkboxes that allows to enable/disable a grid and a legend for the plot. At the right there are two buttons. The **Add Plot** button adds other tabs to the previous panel where it is possible to set the parameters of the graph with different colours in such a way to compare data from different tables.
By clicking on the **Plot** button, in the upper panel the plot is diplayed and is saved in JPEG file format.



demonstratior
train.fits
test.fits

psfMag_u

XLabel: □ X Log □ X Flip

YLabel: □ Y Log □ Y Flip

Red

Set Legend Label: Plot1

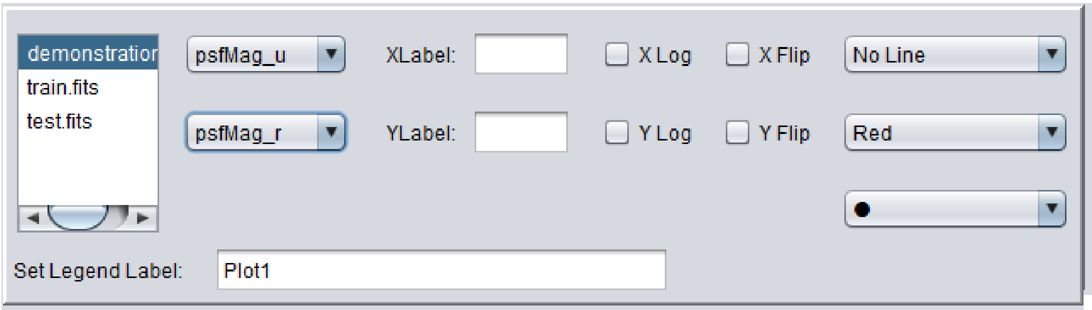
Figure 4.7: Histo Plot panel.

The bottom panel has different fields for every plot option. For the **Histo Plot** (Fig.4.7) the parameters are:

- a **Table List** that is the same of the main window;
- a drop-down menu to set the **XAxis** selecting a column of the table;
- two text fields where it is possible to change the labels for the axis x and y;
- two checkboxes for each axis, one to flip and another to set the axis in logarithmic scale;
- another drop-down menu allow to set the colour;
- below there is another text field where to change the label for the plot legend.

For the **Scatter Plot** option (Fig.4.8) the parameters are:

- a **Table List** that is the same of the main window;
- two drop-down menu to set the **XAxis** and **YAxis** selecting a column of the table;



psfMag_u ▼ XLabel: □ X Log □ X Flip No Line ▼

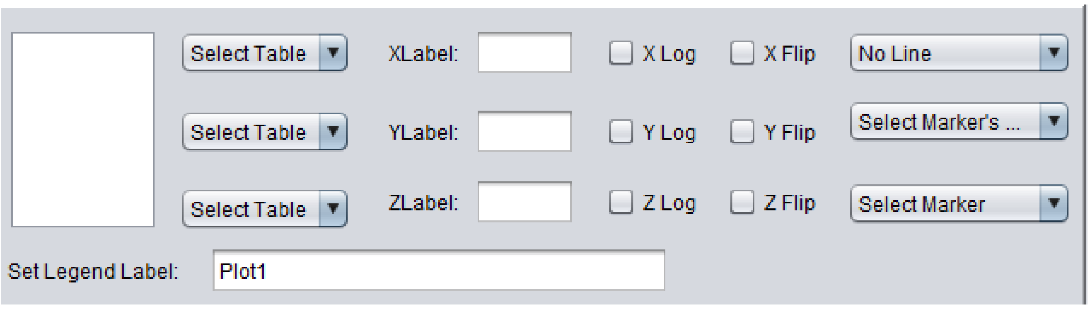
psfMag_r ▼ YLabel: □ Y Log □ Y Flip Red ▼

● ▼

Set Legend Label: Plot1

Figure 4.8: Scatter Plot panel.

- two text fields where it is possible to change the labels for the axis x and y;
- two checkboxes for each axis, one to flip and another to set the axis in logarithmic scale;
- three drop-down menu allow to set the *Line Style*, the *Colour* and the *Marker*;
- below there is another text field where to change the label for the plot legend.



Select Table ▼ XLabel: □ X Log □ X Flip No Line ▼

Select Table ▼ YLabel: □ Y Log □ Y Flip Select Marker's ... ▼

Select Table ▼ ZLabel: □ Z Log □ Z Flip Select Marker ▼

Set Legend Label: Plot1

Figure 4.9: 3D Plot panel.

For **3D Plot** (Fig.4.9) the parameters are:

- a **Table List** that is the same of the main window;
- three drop-down menu to set the **XAxis**, **YAxis** and **ZAxis** selecting a column of the table;
- three text fields where it is possible to change the labels for the axis x, y and z;

- two checkboxes for each axis, one to flip and another to set the axis in logarithmic scale;
- three drop-down menu allow to set the *Line Style*, the *Colour* and the *Marker*;
- below there is another text field where to change the label for the plot legend.

Dragging the mouse on the plot panel, the 3D Plot rotates, so it is enabled also a **Reset** button that allows to redraw the plot in its initial position.

Chapter 5

A Catalogue of Photometric Redshift for 130 million SDSS-DR9 galaxies

The main scientific aim of this thesis was the evaluation of photometric redshifts for the dataset described in the Ch.§2.
In order to achieve this goal, it is real important to remember that all photo-z methods are based on the interpolation of some a priori knowledge, so the first step was to construct the **Knowledge Base** (KB).
The complete dataset contained 133,923,672 object classified as *galaxies* from the SDSS Data Release 9.
The Knowledge Base data were extracted from the spectroscopic subsample of the SDSS-DR9 and 76 objects with missing information (Not a Number or NaN) in any of the five SDSS bands (u-g-r-i-z) were rejected. The resulting KB consisted therefore of 890,119 objects with spectroscopic data.
After the training, the frozen network was applied to all galaxies in the SDSS-DR9 detected in all five SDSS bands.

5.1 Photo-z Estimation

In this section we describe the PhotoRApToR application to the KB to train the method and evaluate the performances of the model. Remembering the descrip-

This section is largely extracted from:

- Brescia M., Cavuoti S., **De Stefano V.**, Longo G., *A catalogue of photometric redshifts for the SDSS-DR9 galaxies*, 2013, **Submitted to A & A**

tion in the Ch.§4, by clicking on the **Open** button a *Load Dialog* opens where will be selected the Knowledge Base dataset: **SDSS_gal_indexed_ psfMagCleaned.csv**.
As you can see from the file extension, it is a COMMA-SEPARATED VALUE file, but, to avoid errors, in the Load Dialog the user must confirm the file type, so the file extension is not really important. This choice is related to the different extensions that is possible to find for any of the file types: for example on line there are FITSTABLEs with *.fit* or *.fits* extension or VOTABLEs with *.vo*, *.vot*, *.votable* and more.

The Knowledge Base dataset **SDSS_gal_indexed_psfMagCleaned.csv** was created by performing the query as listed below:

```
SELECT p.objid, s.specObjID, p.ra, p.dec, p.psfMag_u, p.psfMag_g,
    p.psfMag_r, p.psfMag_i, p.psfMag_z, p.psfmagerr_u,
    p.psfmagerr_g, p.psfmagerr_r, p.psfmagerr_i, p.psfmagerr_z,
    p.fiberMag_u, p.fiberMag_g, p.fiberMag_r, p.fiberMag_i, p.fiberMag_z,
    p.fibermagerr_u, p.fibermagerr_g, p.fibermagerr_r, p.fibermagerr_i,
    p.fibermagerr_z, p.petroMag_u, p.petroMag_g, p.petroMag_r,
    p.petroMag_i, p.petroMag_z, p.petromagerr_u, p.petromagerr_g,
    p.petromagerr_r, p.petromagerr_i, p.petromagerr_z,
    p.modelMag_u, p.modelMag_g, p.modelMag_r, p.modelMag_i,
    p.modelMag_z, p.modelmagerr_u, p.modelmagerr_g, p.modelmagerr_r,
    p.modelmagerr_i, p.modelmagerr_z, p.extinction_u, p.extinction_g,
    p.extinction_r, p.extinction_i, p.extinction_z,
    s.z as zspec, s.zErr as zspec_err,
    s.zWarning, s.class, s.subclass, s.primTarget


FROM PhotoObjAll as p, SpecObj as s

WHERE s.class = ’GALAXY’

AND p.mode = 1

AND dbo.fPhotoFlags(’PEAKCENTER’) != 0

AND dbo.fPhotoFlags(’NOTCHECKED’) != 0

AND dbo.fPhotoFlags(’DEBLEND_NOPEAK’) != 0

AND dbo.fPhotoFlags(’PSF_FLUX_INTERP’) != 0

AND dbo.fPhotoFlags(’BAD_COUNTS_ERROR’) != 0
```

AND dbo.fPhotoFlags('INTERP_CENTER') != 0

AND p.SpecObjID = s.SpecObjID

AND s.zWarning = 0

5.1.1 Manipulating Datasets

After the loading phase, the table is visible into the *Table List* and, when one item is selected, all its properties are displayed into the panel on the right. For the experiment the PSF corrected magnitudes (*psfMag*) had been used, so in the *Edit Table* panel the original table can be edited to choose only necessary features. Selecting only photometric information and spectroscopic redshift with the checkbox (Fig. 5.1), the subtable "**SDSS _gal _indexed _psfMagCleaned _subset**" was created.

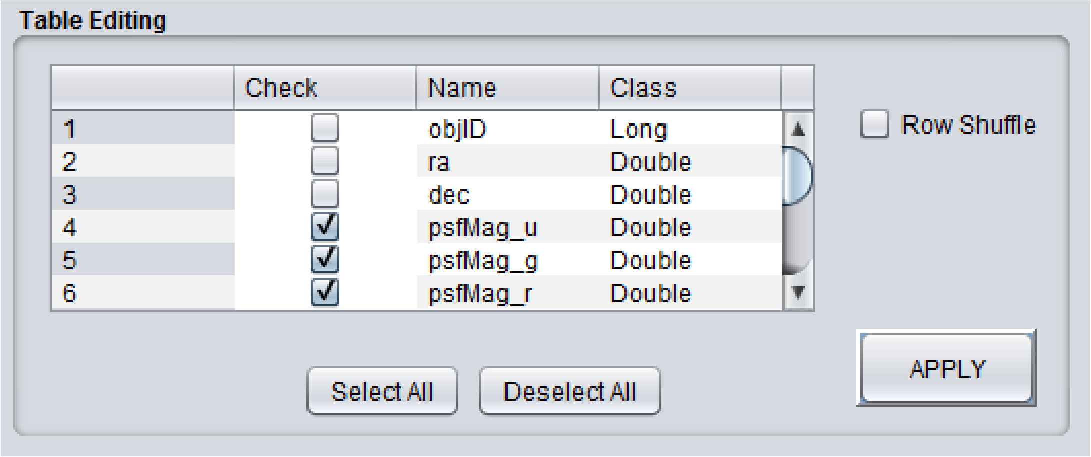


Table Editing

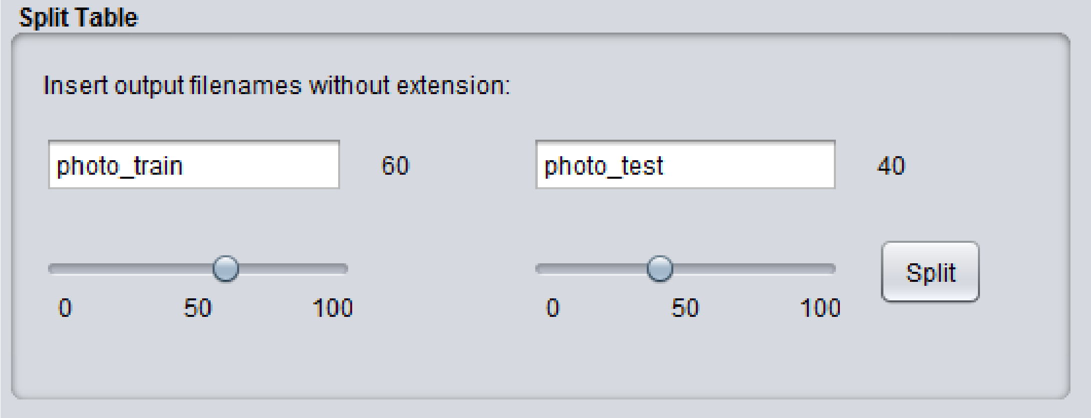
	Check	Name	Class
1	□	objID	Long
2	□	ra	Double
3	□	dec	Double
4	☑	psfMag_u	Double
5	☑	psfMag_g	Double
6	☑	psfMag_r	Double

□ Row Shuffle

Figure 5.1: PhotoRApToR Edit panel. When the desired columns are checked, clicking the **Create Subset** button, a new table is generated and is added to the Table List.

Now it is possible to edit the subtable.
In Ch.§3 it was stressed that for machine learning supervised methods three different subsets for every experiment are generally obtained from the available KB: one (*training set*) to train the method in order to acquire the hidden correlation

In SDSS-III for object 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. The image is sync-shifted so that it is centered on a pixel, and then a Gaussian model of the PSF is fitted to it. This fit is carried out on the local PSF KL model at each position; the difference between the two is then a local aperture correction, which gives a corrected PSF magnitude. Bright stars are used to determine a further aperture correction to a radius of 7.4" as a function of seeing. The resulting magnitude is stored in the quantity *psfMag*.



Split Table

Insert output filenames without extension:

photo_train 60

photo_test 40

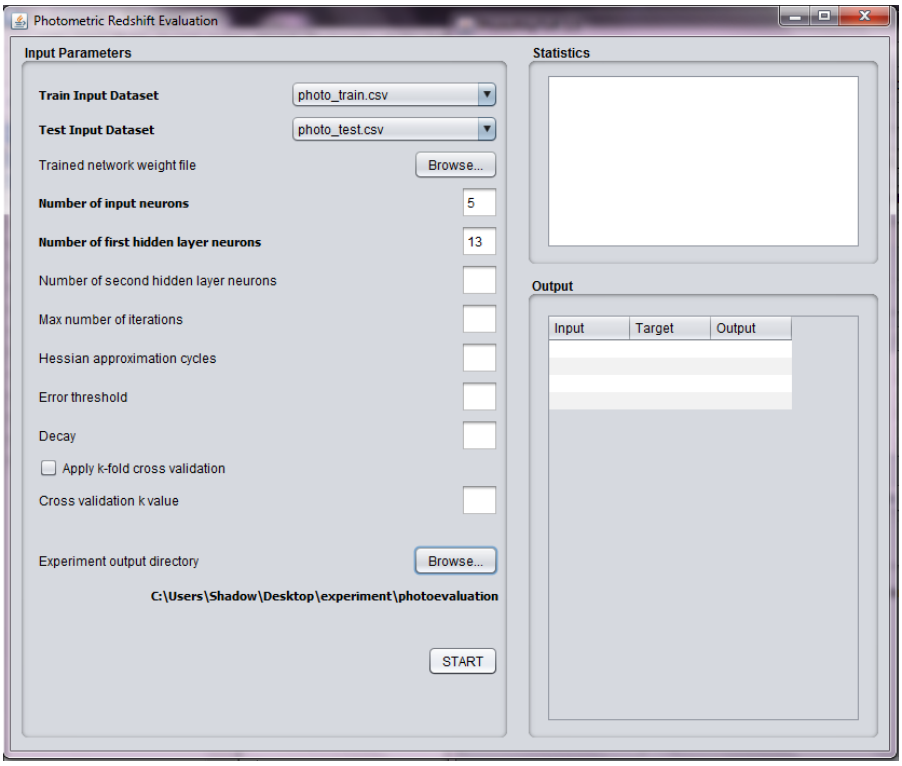
Figure 5.2: PhotoRApToR Split panel. After selecting the table to be split, two different names are choosen for the files and the sliders are dragged to select a different percentage. Clicking on Split button generate the two split datasets.

among the input features which is needed to perform the regression; the second one (*validation set*) is used to check the training, in particular against the loss of generalization capabilities (a phenomenon also known as overfitting); and the third one (*test set*) is used to evaluate the overall performances of the model (Brescia et al. (2013a)).
For the PhotoRApToR application of the MLPQNA method, the validation has been implicitly performed during the training phase, by applying the standard leave-one-out k-fold cross validation mechanism (Geisser (1975)).
So, before the photo-z evaluation, it is necessary to split the dataset in a TRAIN subset and in a TEST subset. According to previous experiences, the two split files (called *photo_train* and *photo_test*) were populated with respectively 60% and 40% of the objects in the KB, obtaining the training set with 535,016 objects and the test one with 356,103 objects (Fig.5.2).

5.1.2 Photo-z Estimation Setting

A click on the **photo-z** button opens a new window (Fig. 5.3). To run photo-z evaluation experiment it was necessary to set the MLPQNA input parameters, as seen for Regression and Classification cases in Ch.§4.
The window was created to have in input MLPQNA requested parameters to run a regression train + test experiment, so that it generates a table where the last column is the estimated photometric redshift. In details, the parameters that have to be setted are the following:

- Two drop-down menu allow to select the TRAIN dataset and the TEST one; (**this parameter is a field required**)



Train Input Dataset: photo_train.csv

Test Input Dataset: photo_test.csv

Trained network weight file

Number of input neurons: 5

Number of first hidden layer neurons: 13

Number of second hidden layer neurons

Max number of iterations

Hessian approximation cycles

Error threshold

Decay

☐ Apply k-fold cross validation

Cross validation k value

Experiment output directory

C:\Users\Shadow\Desktop\experiment\photoevaluation

Figure 5.3: Photometric redshift evaluation window. On the left there are the fields for setting parameters. On the right there are two panels. When the experiment is complete, in the upper panel the regression train statistics and the regression test statistics is displayed. In the lower panel the final table with the photo-z column is reported.

- if we had already done the training phase, it is possible to use the **trained weight file**;
- the **Number of input neurons** is the number of input dataset columns (in our case is 5); (**this parameter is a field required**)
- the **Number of first hidden layer neurons** is the number of neurons of the first hidden layer of the network (in this case 13); (**this parameter is a field required**)
- the **Number of second hidden layer neurons:** described in Ch.§4 (in this case is 4) ;
- **Max number of iteration:** described in Ch.§4 (for our case is 30000);
- **Hessian approximation cycles:** described in Ch.§4 (in our case is 60);
- **Error threshold:** described in Ch.§4 (in this case is 0.0001);

- **Decay:** described in Ch.§4 (for our case is 0.01);
- **Cross validation:** as seen in Experiment description in Ch.§4;
- finally **Experiment output directory** is the parent directory of the output for the experiments and it was called **photoevaluation**.

From the available data it was necessary to check which types of flux combinations were more effective, in terms of magnitudes or related colors. So were performed and compared two kinds of experiments:

- MAG experiment: the five SDSS PSF magnitudes have been used as input features;
- MIXED experiment: the 4 colors (U-G, G-R, R-I, I-Z) and the reference magnitude R have been used as input features;

The optimal combination turned out to be the MIXED type. From the physical point of view this can be easily understood by noticing that even though colors are derived as a subtraction of magnitudes, the content of information is quite different, since an ordering relationship is implicitly assumed, thus increasing the amount of information in the final output. In the MIXED experiment, the network has two hidden layers and stronger QNA parameters, in order to obtain the expected best performance of the model, so as reported in the Table 5.1.1

ITEM	**EXPERIMENTS**	
	MAG	**MIXED**
TRAIN SET	712022 (80%)	535016 (60%)
TEST SET	178097 (20%)	356103 (40%)
INPUT FEATURES	5 magnitudes (ugriz)	4 colors + R mag
hidden layers	1	2
hidden 1 neurons	11	13
hidden 2 neurons	0	4
learnig decay	0.1	0.01
hessian approx. restarts	30	60
error threshold	0.0001	0.0001
max iterations at each restart	4000	30000

Table 5.1.1: Experiment setup comparison.

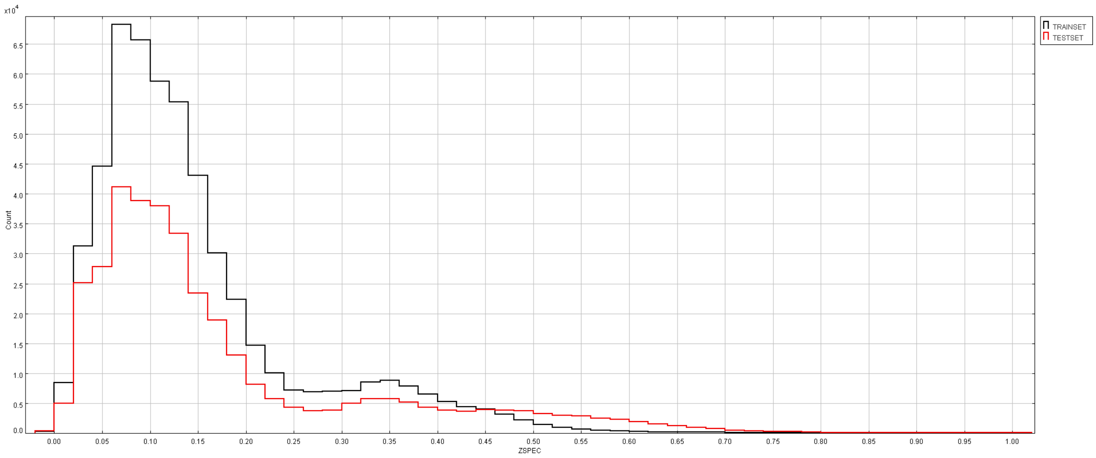


Figure 5.4: Histograms of spectroscopic redshift (z_{spec}) distribution for data used in MIXED experiment. In figure it is possible to see the training set distribution (black line) and the test one (red line).

Statistics

The obtained results of the individual experiments have to be evaluated in a consistent and objective manner through an homogeneous set of statistical indicators. As described in Ch.§4, PhotoRApToR uses a specific algorithm to generate statistics.

For each experiment there is a list of N samples for z_{spec} and z_{phot}. So are defined:

$$\Delta z = z_{spec} - z_{phot} \tag{5.1}$$

$$\Delta z_{norm} = \frac{z_{spec} - z_{phot}}{1 + z_{spec}} \tag{5.2}$$

where Δz_{norm} is the normalized Δz. The notations used into the statistical reports are the same described previously:

$$\begin{aligned}
bias(x) &= \frac{\sum_{i=1}^{N} x_i}{N} \\
\sigma(x) &= \sqrt{\frac{\sum_{i=1}^{N}\left[x_i - \left(\frac{\sum_{i=1}^{N} x_i}{N}\right)\right]^2}{N}} \\
RMS(x) &= \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N}} \\
MAD(x) &= Median(\mid x \mid) \\
NMAD(x) &= 1.48 \times Median(\mid x \mid)
\end{aligned}$$

The term x in all above expressions may be either Δz or Δz_{norm}.
Average statistical indicators such as bias and Standard Deviation, however, provide only part of the information which allows to correctly evaluate the performances of a method. There is a relation between RMS and the Standard Deviation σ: $RMS = \sqrt{mean^2 + \sigma^2}$, but σ^2 is the *variance*, so we have $RMS = \sqrt{mean^2 + variance}$. For a direct comparison of results, in terms of distance of $m\sigma$ (m = 1, 2 ...) from the distribution of Δz, it is much more precise to use the Standard Deviation as main indicator, rather than the simple RMS.
There is often a confusion about the relation between photometric and spectroscopic used to apply the statistical indicators. For instance, the performance could be very different if the simple Δz is used instead of the Δz_{norm}. The idea is that the Δz cannot represent the best choice in the specific case of photometric redshift prediction.
The velocity dispersion error, intrinsically present within the photometric estimation, is not uniform in a wide spectroscopic sample, and the related statistics is not able to give a consistent estimation at all ranges of redshift. On the contrary, the normalized term Δz_{norm} introduces a more uniform information, correlating in a more correct way the variation of photometric estimation, and permitting a more consistent statistical evaluation at all ranges of spectroscopic redshift. More in detail:

$$\begin{aligned} z &= \frac{\Delta\lambda}{\lambda} = \frac{\lambda_{obs} - \lambda_{emit}}{\lambda_{emit}} = \qquad (5.3) \\ &= \frac{\lambda_{obs}}{\lambda_{emit}} - 1 \\ -> 1 &+ z = \frac{\lambda_{obs}}{\lambda_{emit}} \end{aligned}$$

So, differentiating the Eq.5.4:

$$\begin{aligned} dz &= d\left(\frac{\lambda_{obs} - \lambda_{emit}}{\lambda_{emit}}\right) = \frac{d\lambda_{obs}}{\lambda_{emit}} = \\ &= \frac{d\lambda_{obs}}{\lambda_{emit}} \frac{\lambda_{obs}}{\lambda_{obs}} = \frac{d\lambda_{obs}}{\lambda_{obs}}(1 + z) \qquad (5.4) \end{aligned}$$

Finally obtaining:

$$\frac{dz}{1 + z} = \frac{d\lambda_{obs}}{\lambda_{obs}} \qquad (5.5)$$

And the right term of the Eq. 5.5 is exactly the variation between photometric and spectroscopic observed redshift, which is the main focus of the photometric redshift estimation for empirical models which learn its prediction based on the spectroscopic information.

Ref.	$\|bias\|$	σ	NMAD	RMS	bias$_{norm}$	σ_{norm}	NMAD$_{norm}$	RMS$_{norm}$
			TEST DATASET ONLY					
MAG	0.0002	0.016	0.001	0.016	0.0003	0.012	0.0013	0.012
MIXED	0.0004	0.014	0.001	0.014	0.0003	0.011	0.0013	0.011
Laurino et al.	0.015	0.015	0.016	0.021	0.014	0.013	0.013	0.019

Table 5.2.2: Standard statistical indicators used to evaluate the performances of photo-z methods. Also the same indicators for Laurino et al. (2011) are included..

This result is invariant to the redshift range considered. In conclusion the term $\frac{dz}{1+z}$ is the best choice on which to apply the statistical operators.

5.2 Results

For empirical methods, based on machine learning paradigm, the correct way to present the results is to refer to the test set output only, otherwise the performance is altered by the obviously precise bias introduced by considering the training output.
The results of MAG and MIXED experiments are compared with those obtained by Laurino et al. (2011) which achieved the higher accurancy prior 2012: they used a machine learning model with a slightly more complex architecture, named WGE (Weak Gated Experts) method. This comparison is reported in the Table 5.2.2 where, overall, the MIXED experiment provides the best results, with a normalized standard deviation of 0.011.

Table 5.2.3 reports the fraction of outliers, i.e. objects for which the photometric redshift estimate deviates from the spectroscopic value. For the outliers evaluation, the statistical estimation was provided at different multiples of the standard deviation (from 1σ to 4σ). This gives the possibility to evaluate the trend of the prediction scattering and to proceed with a deeper analysis of objects resulting as outliers at different degrees of scattering.
For what the analysis of the catastrophic outliers is concerned, according to Mobasher et al. (2007), the parameter $D_{95} = \Delta_{95}/(1 + z_{phot})$ enables the identification of outliers in photometric redshifts derived through SED fitting methods (usually evaluated through numerical simulations based on mock catalogues). In fact, in the hypothesis that the redshift error $\Delta z_{norm} = (z_{spec} - z_{phot})/(1 + z_{spec})$ is Gaussian, the catastrophic redshift error limit would be constrained by the width of the redshift probability distribution, corresponding to the 95% confidence interval, i.e. with $\Delta_{95} = 2\sigma(\Delta z_{norm})$. In our case, however, photo-z are empirical, i.e. not based on any specific fitting model and it is preferable to use the standard

Ref.	$\|\Delta z\|$	$\|\Delta z\|$	$\|\Delta z\|$	$\|\Delta z\|$	$\|\Delta z\|_{norm}$	$\|\Delta z\|_{norm}$	$\|\Delta z\|_{norm}$	$\|\Delta z\|_{norm}$
	$> 1\sigma$	$> 2\sigma$	$> 3\sigma$	$> 4\sigma$	$> 1\sigma$	$> 2\sigma$	$> 3\sigma$	$> 4\sigma$
MAG	9.6%	4.01 %	1.82 %	0.95%	11.54%	4.52%	2.09%	0.92%
MIXED	9.65%	3.76%	1.76%	0.93%	10.62 %	4.33%	2.03 %	1.05%

Table 5.2.3: Fraction (in percentage) of outliers computed using a 1, 2, 3, and 4 σ clipping threshold.

deviation value $\sigma(\Delta z_{norm})$ derived from the photometric cross matched samples, although it could overestimate the theoretical Gaussian σ, due to the residual spectroscopic uncertainty as well as to the method training error. Therefore, we consider as catastrophic outliers the objects with $|\Delta z_{norm}| > 2\sigma(\Delta z_{norm})$.

The MIXED experiment provides the best results with a very low fraction of outliers ($\sim$ 4% at 2σ and $\sim$ 1% at 4σ).
The catastrophic outliers, i.e. those objects for which $| \Delta z_{norm} |> 2\sigma(\Delta z_{norm})$, were rejected and so a σ_{norm} of $\sim$ 0.0050, larger than $NMAD_{norm}$, was obtained. This result exactly corresponds with what stated in Mobasher et al. (2007). In fact, in the case where photo-z are empirical, it is always useful to analyze the direct correlation between the $NMAD_{norm}$ and the standard deviation σ_{norm} calculated on data which are not catastrophic outliers. In these cases a correct photo-z prediction occurs whenever the quantity $NMAD_{norm}$ is lower than the σ_{norm} for the cleaned sample, induced by the residual spectroscopical uncertainly as well as by the method training error.

The MIXED configuration was therefore used to produce the final catalogue of photo-z. Before to calculate the photo-z for all the objects in the catalogue, it can be shown the plot of the estimated photometric redshift versus the spectroscopic redshift values for all objects in the test set of the MIXED experiment (Fig.5.5).

5.2.1 The Catalogue

After training, the frozen network generated during the MIXED experiment was applied to all galaxies in the SDSS-DR9 detected in all five SDSS bands. As described in the Ch.§2 the final downloaded SDSS-DR9 catalogue contains all objects within the declination range [-30°; +85°] and detected in all SDSS bands classified as galaxies. For convenience, the whole catalogue was split in 38 files containing a total of 133,923,672 objects, for which the photo-zs were evaluated.

The final catalogue consists therefore of 38 files corresponding to different declination ranges (Tab:5.2.4), each of them being structured in 19 columns containing:

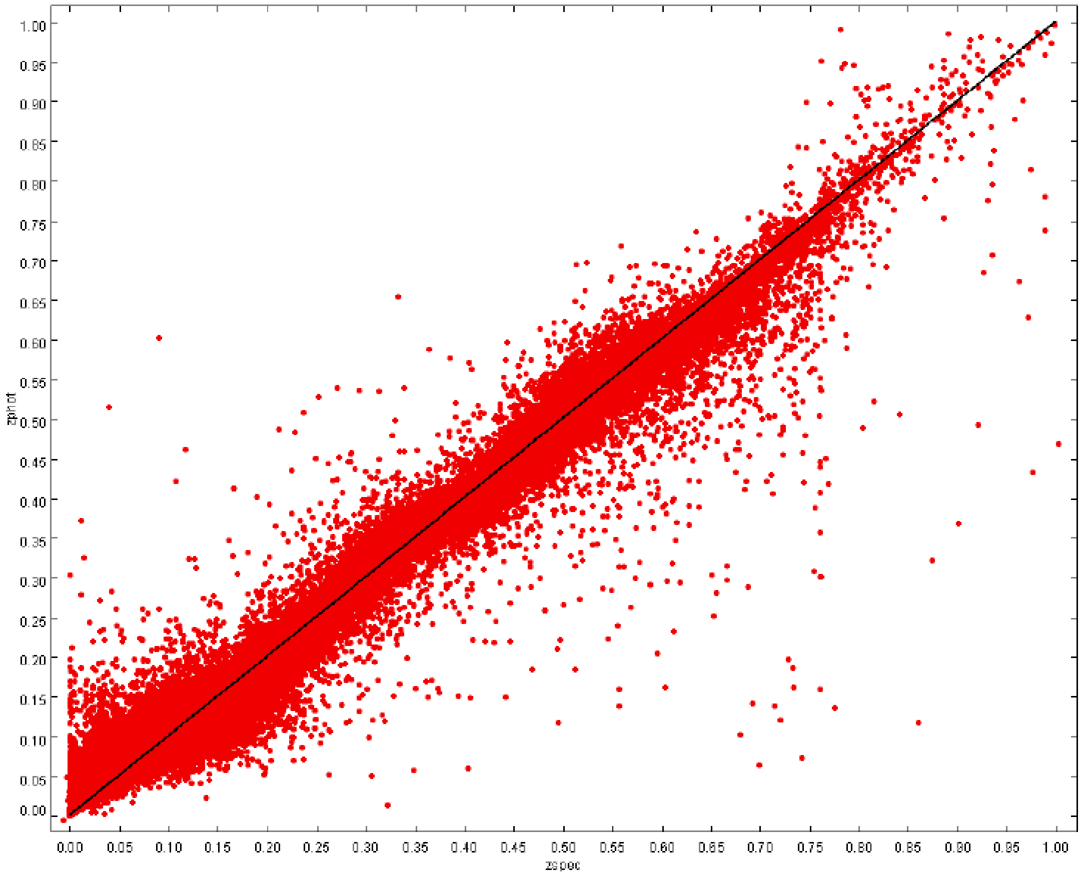


Figure 5.5: Spectroscopic versus photometric redshifts for the 356103 galaxies in the blind test set. As it can be seen, no systematic trends, besides the well known bias at low redshifts ($z < 0.1$), are present.

- column 1: the SDSS-DR9 object identification;
- columns 2 and 3: right ascension and declination;
- columns 4-8: the u, g, r, i, and z PSF magnitudes;
- columns 9-14: the $PSFMag_{err}$ for all magnitudes;
- columns 14-18: the extinction for each magnitude;
- column 19: the estimated photo-z;

The catalogue is publicly available at the URL: `http://dame.dsf.unina.it/catalog/DR9PHOTOZ/`.

5.3 Conclusions

The photometric redshift is a valid and necessary instrument for a variety of cosmological applications. In the introductory chapter I summarized how photo-z's have become crucial in last decades due to the increasingly rapid growth of astronomical data archives.
The many ongoing and planned photometric surveys produce huge datasets that would not be analyzed without the data mining methods deriving from the emerging field of Astroinformatics. In this thesis are described different methods (theoretical and empirical) to evaluate photo-z, but the attention was focused on an empirical method that uses Neural Networks. This is a Machine Learning supervised method that uses a Quasi Newtonian learning rule: MLPQNA.
To demonstrate the accuracy with which photo-z are estimated by MLPQNA, it was applied to all galaxies contained in the Sloan Digital Sky Survey (SDSS) Data Release 9 using a newly implemented desktop application realized in Java language: PhotoRApToR.
With PhotoRApToR it is possible to open tables in most diffused format, to edit them for the application of MLPQNA to run experiments and to plot data in different plot styles. These functions have been applied on the SDSS-DR9 dataset of galaxies.
After the training phase on a Knowledge Base data extracted from the spectroscopical subset of SDSS-DR9, the best results were obtained with a two hidden layer network with a combination of the 4 SDSS colors (obtained from the SDSS *psfMag*) plus the *psfMag* in the r band. This experiment leads to a negligible bias, to a low fraction of outliers and to a normalized standard deviation of the residuals $\sigma_{norm} = 0.011$, which decreases to 0.005 after the rejection of catastrophic outliers. This result is better or comparable with what was already available in the literature and presents a smaller number of catastrophic outliers.
The MLPQNA method was then applied on the complete dataset containing 133,923,672 object classified as galaxies and the resulting catalogue is available at the address reported above.

The application of the method presented in this work, has produced a refereed article, currently under revision by the editor Brescia et al. (2013b). Furthermore, there is also another manuscript in preparation, whose topic is the PhotoRApToR application [De Stefano et al. 2013, in preparation].

Catalog File	Objects	Photo-z range	
(DEC range)		MIN	MAX
DEC(-30_-20)	601937	0	0.774
DEC(-20_-15)	1481223	0	0.768
DEC(-15_-08)	3680856	0	0.774
DEC(-08_-05)	3592638	0	0.772
DEC(-05_-03)	2604467	0	0.772
DEC(-03_-01)	4219465	0	0.776
DEC(-01_00)	2998182	0	0.776
DEC(00_01)	3031914	0	0.771
DEC(01_02)	2452737	0	0.776
DEC(02_04)	4610011	0	0.782
DEC(04_06)	4783281	0	0.781
DEC(06_08)	4859457	0	0.779
DEC(08_09)	2288882	0	0.765
DEC(09_10)	2240852	0	0.772
DEC(10_11)	2222814	0	0.779
DEC(11_12)	2159481	0	0.779
DEC(12_14)	4466309	0	0.776
DEC(14_16)	4670156	0	0.779
DEC(16_18)	4586431	0	0.778
DEC(18_20)	4500819	0	0.779
DEC(20_22)	4362577	0	0.780
DEC(22_24)	4428770	0	0.777
DEC(24_26)	4494908	0	0.782
DEC(26_28)	4314896	0	0.775
DEC(28_30)	4174234	0	0.777
DEC(30_32)	3957315	0	0.774
DEC(32_34)	3693299	0	0.775
DEC(34_37)	5093588	0	0.774
DEC(37_40)	4894905	0	0.771
DEC(40_43)	4606852	0	0.770
DEC(43_46)	4339472	0	0.780
DEC(46_50)	4589908	0	0.777
DEC(50_55)	4766871	0	0.775
DEC(55_60)	4013131	0	0.776
DEC(60_65)	3481072	0	0.778
DEC(65_70)	1516315	0	0.769
DEC(70_75)	351332	0	0.767
DEC(75_85)	792315	0	0.764
TOTAL	**133923672**	0	0.782

Table 5.2.4: Information on the final photo-z catalogue produced by MLPQNA.

Bibliography

Abazajian, Adelman-McCarthy, Agueros et al., *ApJS*, 182, 543 arXiv:0812.0649 (2009)

Abdalla, Banerji, Lahav, & Rashkov, arXiv:0812.3831 (2008)

Adelman-McCarthy J. K., et al.,*ApJS*, 175, 297 arXiv:0707.3413 (2008)

Ahn C. P., et al, *ApJS*, 203, 21 arXiv:1207.7137 (2012)

Aihara, et al., *ApJS*, 193, 29 arXiv:1101.1559 (2011a)

Aihara, et al., *ApJS*, 195, 26 (2011b)

Albrecht, Bernstein, Cahn, et al., arXiv:astro-ph/0609591 (2006)

Arnouts, Moscardini, Vanzella et al., *MNRAS*, 329, 355 (2002)

Assef, Kochanek, Brodwin et al., *ApJ*, 676, 286 arXiv:0708.1513 (2008)

Assef, Kochanek, Brodwin et al., *ApJ*, 713, 970 arXiv:0909.3849 (2010)

Banerji, Abdalla, Lahav & Lin, *MNRAS*, 386, 1219 arXiv:0711.1059 (2008)

Baum, *Proceedings from IAU Symposium*, ed. G.C. McVittie, 15, 390 (1962)

Benitez, *ApJ*, 536, 571 arXiv:astro-ph/9811189 (2000)

Bishop, Pattern Recognition and Machine Learning, Springer ISBN 0-387-31073-8 (2006)

Blanton & Roweis, *AJ*, 133, 734 arXiv:astro-ph/0606170 (2007)

Bolton A. S., et al., *AJ*, 144, 144 (2012)

Bolzonella, Miralles & Pello, *A&A*, 363, 476 (2000)

Borne K., *Astroinformatics: A 21st Century Approach to Astronomy*, Position Papers, no.6 (2009)

Brammer, van Dokkum & Coppi, *ApJ*, 686, 1503 (2008)

Brescia M., Longo G., Djorgovski G.S., Cavuoti S., D'Abrusco R., Donalek C., arXiv:1010.4843 (2010)

Brescia M., Cavuoti S., Paolillo M., Longo G., Puzia T., *MNRAS*, 421, 2, 1155 arXiv:1110.2144 (2012)

Brescia M., *New Trends in E-Science: Machine Learning and Knowledge Discovery in Databases, Horizons in Computer Science Research*, Series Horizons in Computer Science Vol. 7, Nova Science Publishers, ISBN: 978-1-61942-774-7 (2012)

Brescia M., Cavuoti S., DAbrusco R., Longo G. & Mercurio A., *ApJ*, 772, 140 arXiv:1305.5641 (2013)

Brescia M., Cavuoti S., De Stefano V., Longo G., Submitted to A&A (2013)

Brunner, Connolly & Szalay, *ApJ*, 516, 563 (1999)

Brunner R., Djorgovski S. G., Prince T., Szalay A., Massive Data Sets in Astronomy *Handbook of Massive Data Sets*, p. 931 (2001)

Bruzual & Charlot, *ApJ*, 405, 538 (1993)

Buzzoni, *IAU Symposium*, ed. K. Sato, Dordrecht, Kluwer, 183, 134 (1998)

Capozzi, De Filippis, Paolillo, D'Abrusco, Longo, *MNRAS*, 396, 900 (2009)

Capozziello S. & Funaro M., *Introduzione alla Relativitá Generale* ISBN: 978-88-207-3872-3 (2005)

Carliles, Budavari, Heinis, Priebe & Szalay, *ApJ*, 712, 511 (2010)

Cavuoti S., Brescia M., Longo G., Mercurio A., *A&A*, 546, A13, 1 (2012) arXiv:1206.0876

Cavuoti S., Brescia M., Longo G., Garofalo M. & Nocella A., *Science - Image in Action*, World Scientific Publishing, 241 (2012)

Cavuoti S., *Data Rich Astronomy: Mining Synoptic Sky Surveys*, PhD in Physics, XXV Cycle, University Federico II of Naples, available at `http://dame.dsf.unina.it/documents/Stefano_Cavuoti_PhD_Thesis.pdf` (2013)

Charlot, Worthey, Bressan, *ApJ*, 457, 625 (1996)

Cole S., et al., *MNRAS*, 362, 505 (2005)

Coleman, Wu & Weedman, *ApJS* 43, 393 (1980)

Collister & Lahav, *PASP*, 116, 345 arXiv:astro-ph/0311058 (2004)

Collister, Lahav, Blake et al., *MNRAS*, 375, 68 (2007)

Connolly, Csabai, Szalay et al., *AJ*, 110, 2655 arXiv:astro-ph/9508100 (1995)

Connolly et al., *ApJ*, 486, 11 (1997)

Csabai et al., *AJ*, 125, 580 arXiv:astro-ph/0211080 (2003)

Csabai, Dobos, Trencseni et al., *Astronomische Nachrichten*, 328, 852 (2007)

Davidon, *Computational Journal* 10, 406. (1968)

Dawson K. S., et al., *AJ*, 145, 10 (2013)

De Lucia & Blaizot, *MNRS*, 375, 2 (2007)

Djorgovski S. G. et al., *Some Pattern Recognition Challenges in Data-Intensive Astronomy*(ICPR 2006), Vol. 1, eds. Y. Y. Tang et al. IEEE Press, p. 856, arXiv:astroph/0608638 (2006)

Donalek C., Graham M., Mahabal A. & Djorgovski S.G., Tools for Data to Knowledge, *VAO Technical Report* (2011)

Eisenstein D.J., et al., *AJ*, 122, 2267 arXiv:astro-ph/0108153 (2001)

Eisenstein D.J., et al., *ApJ*, 633, 560 arXiv:astro-ph/0501171 (2005)

Eisenstein D. J., et al., *AJ*, 142, 72 arXiv:1101.1529 (2011)

Euclid Red Book, ESA Technical Document, ESA/SRE 12, Issue 1.1 (2011)

Feldmann, Carollo, Porciani et al., *MNRAS*, 372, 565 (2006)

Firth A. E. et al., *MNRAS*, 339, 1195 (2003)

Fletcher R. & Reeves C. M., *Computational Journal* 7, 2, 149 MR 0187375 (1964)

Fukugita M., Ichikawa T., Gunn J. E., Doi M., Shimasaku K. & Schneider D. P., *AJ*, 111, 1748

Geisser S., *Journal of the American Statistical Association*, 70 (350), 320-328 (1975)

Gerdes, Sypniewski, McKay et al., *ApJ*, 715, 823 (2010)

Golub G.H. & Ye Q., *SIAM Journal of Scientific Computation*, Vol. 21, 1305 (1999)

Gunn J. E., et al., *AJ*, 116, 3040 (1998)

Gunn J. E., et al., *AJ*, 131, 2332 (2006)

Guyon I. & Elisseeff A., *Journal of Machine Learning Research*, Vol. 3, pp. 1157-1182. (2003)

Guyon I. & Elisseeff A., *In Feature Extraction, Foundations and Applications*, L. A. Editors, Series: Studies in Fuzziness and Soft Computing, Springer, Vol. 207 (2006)

Gwyn& Hartwick, *ApJ*, 468, L77 (1996)

Hestenes M. R. & Stiefel E., *J. Res. Nat. Bur. Standards* 49 , 6, 409 MR 0060307 (1952)

Hey T., Tansley S. & Tolle K., *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Redmond, WA: Microsoft Research. (2009)

Hildebrandt, Wolf & Benitez, *A&A*, 480, 703 (2008)

Hildebrandt, Arnouts, Capak, Wolf et al., *A&A*, 523, 31 (2010)

Hogg, Cohen, Blandford et al., *ApJ*, 115, 1418 (1998)

Ilbert, Arnouts, McCracken et al., *A&A*, 457, 841 (2006)

Jain A.K. et al., *Data Clustering: A Review*, ACM Computing Surveys, 31, 3, 264 (1999)

Keiichi, Medezinski, Nonino, et al.,*ApJ*, Submitted (2012)

Komatsu E., et al., *ApJS*, 192, 18 (2011)

Koo, *AJ*, 90, 418 (1985)

Koo, *Astronomical Society of the Pacific Conference Series, ed. Weymann, Storrie-Lombardi, Sawicki & Brunner.*, Vol. 191, 3 (1999)

Lanzetta, Yahil & Fernandez-Soto, *Nature*, 381, 759 (1996)

Laurino O., D'Abrusco R., Longo G., & Riccio G., *MNRAS*, 418, 2165 (2011)

Loh & Spillar, *ApJ*, 303, 154 (1986)

Luo, Brandt, Xue et al., *ApJS*, 187, 560 (2010)

Lupton R., Blanton M.R., Fekete G., Hogg D.W., O'Mullane W., Szalav A. & Wherry N., *PASP*, 116, 133 (2004)

Maraston C., Stršombšack G., Thomas D., Wake D. A. & Nichol, R. C., *MNRAS*, 394, L107 (2009)

Massarotti, Iovino & Buzzoni, *A&A*, 368, 74 (2001)

Massarotti, Iovino, Buzzoni & Valls-Gabaud, *A&A*, 380, 425 (2001)

Mobasher B., Capak P., Scoville N. Z. et al. *Astroph. Journal Suppl. Series*, 172, Issue 1, 117 (2007)

Noll, Mehlert, Appenzeller et al., *A&A*, 418, 885 (2004)

Oesch, Carollo, Feldmann et al., *ApJ*, 714, L47 (2010)

Padmanabhan N., Xu X., Eisenstein D. J., Scalzo R., Cuesta A.J., Mehta K.T. & Kazin, E. arXiv:1202.0090 (2012)

Pasian F. et al., Astronomical Images and Data Mining in the International Virtual Observatory Context, *Proceedings of Data Analysis in Astronomy* (2011)

Peacock, Schneider, Efstathiou et al., *ESA-ESO Working Group on Fundamental Cosmology, Tech. Rep.* (2006)

Pedregosa, Varoquaux, Gramfort et al., *JMRL*, 12, 2825 (2011), arXiv:1201.0490

Percival W.J. et al., *MNRAS*, 401, 2148 (2010)

Polak E. & Ribiere G., 'Note sur la convergence de methodes des directions conjugees', *Revue Fr. Inf. Rech. Oper.*, 16-R1, 35 MR 0255025 (1969)

Puschell, Owen & Laing, *ApJ*, 257, L57 (1982)

Reid I. N., Brewer C., Brucato R. J. et al. *Publications of the Astronomical Society of the Pacific*, 103, 661 (1991)

Richards G.T. et al., *AJ*, 123, 2945 (2002)

Ross N.P. et al. *ApJS*, 199, 3 (2012)

Sawicki, Lin & Yee, *AJ*, 113, 1 (1997)

Sha, Lin, Saul & Lee, *Neural Computation*, 19, 2004 (2007)

Shadbolt N., Hall W. & Berners-Lee, *IEEE Intelligent Systems*, 21, 3, 96101 (2006)

Smee S.A. et al., *AJ*, 146, 32 (2013)

Strauss, M. A., et al., *AJ* 124, 1810 arXiv:astro-ph/0206225 (2002)

Tagliaferri R., Longo G., Andreon S., Capozziello S., Donalek C. & Giordano G., *Neural Networks and Photometric Redshifts* (2002)

Taylor I.J. et al., *Workflows for e-Science: Scientific Workflows for Grids*, Springer, London (2007)

Tegmark M. et al., *Phys. Rev. D*, 74, 123507 (2006)

Weinberg D.H., Mortonson M.J., Eisenstein D.J., Hirata C., Riess A.G. & Rozo

Weir N., Fayyad U. & Djorgovski S., *AJ*, 109, 2401 (1995)

Williams, Blacker, Dickinson et al., *AJ*, 112, 1335 (1996)

Yanny B., et al., *AJ*, 137, 4377 arXiv:0902.1781 (2009)

York D. G., et al., *AJ*, 120, 1579 (2000)