Università degli Studi di Napoli "Federico II"

Facoltà di Scienze Matematiche, Fisiche e Naturali

Corso di Laurea in Astrofisica e Scienze dello Spazio

Anno Accademico 2014-2015

A new photometric method to determine the metallicity of galaxies.

Candidato:
Carlo M. De Masi
Matricola:
N91/000009

Relatori:
Prof. Giuseppe Longo
Dott. Massimo Brescia
Dott. Stefano Cavuoti
Prof. Claudia Maraston
Dott. Amata Mercurio

Contents

Preamble

Metallicity is a fundamental parameter in the study of galactic evolution, providing us with information on the object's age and on its star formation history; together with Initial Mass Function (IMF), age and Star Formation Rate (SFR), metallicity allows us to fully constrain a galaxy Spectral Energy Distribution (SED), which is the theoretical basis of modern evolutionary population synthesis models.
As widely discussed in literature, moreover, metallicity is correlated to other physical quantities as stellar mass, luminosity and Star Formation Rate; the determination of these quantities, can provide an estimate of the expected metallicity value, which in turn can be compared to the observed ones to investigate the effects of feedback mechanisms (stellar winds, Supernovae explosions or the inflow/outflow of material) on the chemical composition of a galaxy.

The methods currently used to determine metallicity in a galaxy involve the observation of emission or absorption lines of the galactic spectrum; specifically, gas-phase metallicity in active or star-forming galaxies is obtained by the observation of nebular optical emission lines, while stellar metallicity can be estimated by the study of some peculiar spectral absorption features, whose strength is measured by the use of the so-called Spectral Indices SI.
However, because of the difficulties in observing reliable spectral features outside nearby galaxies and of the advantages presented by photometric observations over spectroscopic ones (in terms of telescope time efficiency and of the possibility to observe fainter objects), it appears suitable to develop ways to determine galactic metallicities through the observation of photometric properties alone.
In this sense, the aim of this work is to present the application of Machine Learning techniques to the problem of photometric metallicity determination; using datasets of galaxies of known metal content, we trained a Multi- Layer Perceptron (MLP) neural network with a Quasi-Newton Algorithm (QNA) as learning rule, in order to make the algorithm able to derive metallicity from observable photometric quantities.

The work is organized as follows.

In Chapter 1, we introduced the concept of metallicity from a theoretical point of view, presenting the difference between stellar and gas-phase metallicity in a galaxy and the spectroscopic methods actually used to estimate them; then, after presenting

a model describing the influence of star formation activity on the gas metal content in a galaxy, we explored its consequences in terms of the correlation existing between metallicity and other significant physical parameters of galaxies.

Chapter 2 is dedicated to the description of the Machine Learning methods employed. Specifically, we focused on the Multi-Layer Perceptron neural network model, explaining its properties and describing the Quasi-Newton Algorithm used as optimization rule to train it.

In Chapter 3, we detailed the work we carried out on the photometric catalogues. After presenting the datasets, we described the experiments performed with the neural network, and exposed the corresponding results.
In particular, we used the Bruzual&Charlot 2003 stellar evolutionary population synthesis code to generate a mock catalogue of galaxies to test our algorithm, in order to find the best configuration of the network; we presented the results obtained on the mock-catalogue, and their application to the real galaxies.
Then, we dealt with gas-phase metallicity. We first presented the data used, and then we illustrated the results of the experiments with the network; finally, we compared our method to the photometric metallicity determination used by Sanders et al. (2013).

Finally, Chapter 4 contains a summary of the whole work and the discussion of the results.

Chapter 1

Metallicity - Theoretical framework

In Chemistry, the term "**metal**" usually indicates particular elements or alloys presenting peculiar properties (such as a high electrical and thermal conductivity); in Astrophysics and Cosmology, it assumes a less strict meaning.
Since stars and the InterStellar Medium (ISM) are mainly composed of Hydrogen and Helium, all the other elements only account for a small percentage of the mass of the Universe, so that any element heavier than He is generally referred to as "heavy element" or "metal".

In this sense, we define the **metallicity** of an object (usually designated by the letter **Z**) as the ratio of the object's mass composed by any element heavier than He (m_i) to the total mass (M):

$$Z = \sum_{i>He} \frac{m_i}{M} = 1 - X - Y \tag{1.1}$$

where we defined X and Y as the Hydrogen and Helium mass fraction, respectively. The values of X, Y and Z of an object are often referred to the solar values, which

Parameter	Solar value
H mass fraction	$X_\odot \approx 0.73$
He mass fraction	$Y_\odot \approx 0.25$
Metallicity	$Z_\odot \approx 0.02$

Table 1.1: Solar values of Hydrogen, Helium and metal mass fractions.

are reported in table 1.1 (Binney and Tremaine, 2011)

Metallicity is usually expressed using the quantity [Z/H], which compares the metal to Hydrogen ratio of the object to the solar one according to the relation

$$[M/H] \equiv \frac{Log\left(\frac{Z}{X}\right)_*}{Log\left(\frac{Z}{X}\right)_\odot} = Log\left(\frac{N_M}{N_H}\right)_* - Log\left(\frac{N_M}{N_H}\right)_\odot \tag{1.2}$$

where N_M and N_H are the number of metal and Hydrogen atoms per unit of volume, respectively.
In practice, we usually express the metal content of an object in terms of the abundance of a particular element, which is easier to detect and which is assumed to be a good metallicity indicator for that specific object; the most commonly used element for gas-phase is oxygen, while for stars we usually choose Fe or Mg.

1.1 Metallicity and star populations

The study of the metallicity of an object can give us an important insight on its age and evolution.
According to the cosmological models accepted to date, the only elements formed after the Big Bang were H and He, so that the first stars born in the early Universe, the so-called **Population III stars**, are supposed to have been completely metal-free (Binney and Tremaine, 2011).
Modern stellar models suggest that such stars, formed with no heavy elements and in a warmer ISM, would have been several hundreds of times more massive than the

Sun, i.e. more massive than any star observed today; for this reason, they would have created heavier elements up to Iron via nucleosynthesis, and they would have ended their lives in extremely energetic supernovae explosions, thus enriching the ISM of these metals (Binney and Tremaine, 2011).
The expected short lifespan of Population III stars, a consequence of their extremely high mass, accounts for the fact that no such star has been observed up to date (indirect evidence for the existence of Population III stars has been found in gravitationally lensed galaxies - Fosbury et al., 2003).

The following generation of stars, born out of the ISM enriched by elements produced by metal-free stars, would contain a small, yet detectable amount of heavy elements.
These so-called **Population II stars** are the oldest observed stars in the Universe, with ages ranging from 10 to 13 billion years, and present a low metallicity ([Fe/H]<-1); they are mainly found in the spheroidal component (the stellar halo and bulge) of galaxies, and generally present randomly oriented and highly elliptical orbits (Binney and Tremaine, 2011).
In particular, in the Milky Way we distinguish between intermediate Population II stars, common in the bulge near the center of the galaxy, and late Population II stars, found in the galactic halo and globular clusters, which are older and even more metal-poor.
Population II stars are believed to have formed all the elements of the periodic table up from Iron, and to have eventually returned these elements to the ISM out of which newer, more metal-enriched stars formed ever further (Binney and Tremaine, 2011).
These youngest stars, including our Sun, have the highest metal content, and are known as **Population I stars**.
They are common in the spiral arms of the Milky Way galaxy, and they present roughly circular orbits close to the mid-plane of the galactic disk. As for Population II stars, they can be roughly divided into two groups, with the youngest extreme Population I stars located closer to the plane of the galaxy than intermediate Population I stars (the Sun is an intermediate Population I star).

1.2 Stellar and gas-phase metallicity

When talking about the metallicity of a galaxy, it is appropriate to distinguish between the *stellar* and *gas-phase* values.
These quantities are indeed strictly connected to each other by the process of stellar evolution; stars are created from the collapse of denser regions within molecular clouds in the ISM, and through their lifespan they create heavier elements via the process of nucleosynthesis, only to return these processed materials to the ISM during the latest stages of their evolution. As a result, the initial metal content of each generation of stars depends on the metallicity of the ISM at the time and place of star formation, so that at each moment stellar metallicity gives us indications on the star formation history of the galaxy.

As for the gas-phase metal abundance, looking at the effect of stellar evolution only we expect it to be higher than in stars, and to increase with the succession of star formation periods.
The situation, however, is much more complex, since the chemical composition of ISM is the result of the so-called "feedback" process (Erb et al., 2006; Pettini et al., 2001), a complex interaction between:

- metal production by stars;
- the inflow of gas from the intergalactic medium, which dilutes the metallicity, but at the same time provides the fuel for new star formation;
- the effect of stellar winds and supernovae, which influence the star formation rate by heating the ISM and create outflows of metal-enriched gas into the intergalactic medium.

1.2.1 Gas-phase metallicity determination

As we mentioned earlier, in ISM studies the term metallicity is generally used to refer to its oxygen content, which is used as the reference element since it is one of the most abundant elements and all the lines produced by its ions are generally observable (otherwise, we have to correct for unobserved ions using appropriate correction

factors; see Stasinska, 2004).
So, we usually express gas metallicity as:

$$Z = 12 + Log(O/H) \tag{1.3}$$

where the ratio O/H is the abundance of oxygen relative to hydrogen (the Sun has Z= 8.69).

Generally speaking, the study of ISM metallicity presents some advantages over its stellar counterpart.
Specifically, most of the methods used are based on the examination of nebular optical emission lines, which are characterized by a S/N ratio usually higher than in the continuum (Tremonti et al., 2004, hereafter T04); moreover, the analysis is not affected by the age-metallicity degeneracy, which plagues stellar metallicity determination instead (see next sections).
On the other hand, as we will describe in the following, there exists in literature a great amount of different methods to determine ISM metal abundance, each of which generally provides slightly different results, thus making it difficult to compare them in a straightforward way (Kewley and Dopita, 2002; Kewley and Ellison, 2008).

Direct (T_e) method

The most reliable method to determine the nebular metallicity of emission-line galaxies is the so-called "T_e", or "*direct*" method, based on the assumption of a HII-region model for the galaxy; hence, the abundance of a given element is obtained as the sum of the abundances of its ions, which in turn can be expressed as a function of the ratio of lines emitted by these ions and of the electron temperature T_e (Erb et al., 2006; Stasinska, 2004; Pettini and Pagel, 2004; Pilyugin, 2001).

The results provided by this analysis are largely influenced by line intensity errors, and by the different model adopted for the HII region (it can be a one-zone model with a single characteristic T_e or a two-zone model with two T_e values).
Izotov et al. (2006), for example, obtain the electron temperature from an iterative procedure, using the equations 1.4a and 1.4b

$$\begin{cases} t = \dfrac{1.432}{Log\left(\dfrac{[OIII]\lambda 4959 + \lambda 5007}{[OIII]\lambda 4363}\right) - LogC_T} & (1.4a) \\ C_T = (8.44 - 1.09\,t + 0.5\,t^2 - 0.08\,t^3)\dfrac{1 + 0.0004x}{1 + 0.044x} & (1.4b) \end{cases}$$

with

$$\begin{cases} t = 10^{-4}\,T_e & (1.5a) \\ x = 10^{-4}\,N_e\,t^{-0.5} & (1.5b) \end{cases}$$

where the electron density (N_e) can be measured from the density sensitive [SII]$\lambda\lambda$6716,6731 doublet; then, the ions abundances can be obtained as:

$$\begin{cases} 12 + Log(O^+/H^+) = Log\dfrac{[OII]\lambda 3727}{H\beta} + 5.961 + \dfrac{1.676}{t} + \\ \qquad\qquad - 0.40\,Log(t) - 0.034\,t + Log(1 + 1.35\,x) & (1.6a) \\ 12 + Log(O^{++}/H^+) = Log\left(\dfrac{[OIII]\lambda 4959 + [OIII]\lambda 5007}{H\beta}\right) + 6200 + \\ \qquad\qquad + \dfrac{1.251}{t} - 0.55\,Log(t) - 0.014\,t & (1.6b) \end{cases}$$

and the total oxygen abundance is given by the sum of the abundances of all its ions,

$$(O/H) = O^+/H^+ + O^{++}/H^+ \tag{1.7}$$

The biggest flaw of the T_e method is that the [OIII]λ4363 line gets weaker in higher metallicity ($Z > 0.5\,Z_\odot$, or $12+log(O/H) \geq 8.4$) galaxies, so that the method cannot be used unless we have extremely high S/N spectra, or we use other lines. Moreover,

in the presence of temperature fluctuations the line emissions are not uniform, but they increase in high temperature, low metallicity regions, thus leading to an overall metallicity underestimate (Stasinska, 2004; Stasińska, 2005).

Strong lines methods

The limits of the direct T_e method led to the development of the so-called "strong-line" abundance diagnostics (Erb et al., 2006; Pilyugin and Thuan, 2005; Kewley and Dopita, 2002;T04), which are based on the fit of strong line emission ratios to the T_e method or to theoretical photoionization models (these are often referred to as "empirical methods", though the photoionization models are theoretical, so that the name can be somewhat misleading).

An indicator commonly used in many metallicity diagnostics is the R_{23}, defined as:

$$R_{23} \equiv \frac{[OII]\lambda 3727 + [OIII]\lambda 4959, 5007}{H\beta} \tag{1.8}$$

which basically provides an estimate of the total cooling due to oxygen, and so, since oxygen is one of the principle nebular coolants, of the oxygen abundance itself (McGaugh, 1991; Pilyugin, 2001).
This indicator however (as many other emission-line abundance diagnostics, actually) is degenerate, so we first have to determine which solution branch applies with an initial guess of the metallicity based on an alternative diagnostic; Pilyugin and Thuan (2005) calibrated it versus some T_e metallicity models, finding an "upper" ($12 + log(O/H) > 8.25$), and a "lower" branch ($12 + log(O/H) < 8.0$) :

$$\begin{cases} 12 + Log(O/H)_{upper} = \dfrac{R_{23} + 726.1 + 842.2\,P + 337.5\,P^2}{85.96 + 82.76\,P + 43.98\,P^2 + 1.793\,R_{23}} & \text{(1.9a)} \\ 12 + Log(O/H)_{lower} = \dfrac{R_{23} + 106.4 + 106.8\,P - 3.40\,P^2}{17.72 + 6.60\,P + 6.95\,P^2 - 0.302\,R_{23}} & \text{(1.9b)} \end{cases}$$

where P accounts for the ionization parameter

$$P = \frac{\frac{[OIII]\lambda\lambda 4959,5007}{H\beta}}{R_{23}} \tag{1.10}$$

Other methods

Other methods are based on the combination of two or more diagnostics, according to the metallicity range.
Kewley and Dopita (2002) (hereafter KD02) combined more diagnostics according to the metallicity range; they calibrated various strong line indicators against a set of photoionization models with ionization parameters varying from $q = 5 \times 10^6$ to $q = 3 \times 10^8\ cm\, s^{-1}$ (q characterizes the ionization state of the gas, and is defined as the number of ionizing photons per second per unit area divided by the H density):

- for high metallicity values ($Z > 0.5Z_\odot$, or $12 + log(O/H) > 8.6$) they use the [NII]λ6584/[OII]λ3727 ratio, finding a best-fit quadratic relation of the form

$$12 + Log(O/H) = Log(1.54020 + 1.26602\, R + 0.167977\, R^2) + 8.93 \tag{1.11}$$

  where

$$R = Log\left(\frac{[NII]}{[OII]}\right)$$

- for intermediate metallicity values ($8.5 < 12 + log(O/H) < 8.6$), KD02 use the Zaritsky et al. (1994) (hereafter Z94) method and the calibration by Kobulnicky et al. (1999) of the McGaugh (1991) (hereafter M91) method, i.e. they assume

$$12 + Log(O/H) = \frac{1}{2}(Z_{M91} + Z_{Z94}) \tag{1.12}$$

where

$$\begin{cases} Z_{Z94} = 9.625 - 0.33\,R_{23} - 0.202\,R_{23}^2 + \\ \qquad - 0.207\,R_{23}^3 - 0.333\,R_{23}^4 \end{cases} \tag{1.13a}$$

$$\begin{cases} Z_{M91} = 12.0 - 4.944 + 0.767\,R_{23} + 0.602\,R_{23}^2 + \\ \qquad - y(0.29 + 0.332\,R_{23} - 0.331\,R_{23}^2) \end{cases} \tag{1.13b}$$

with

$$y = Log\left(\frac{[OIII]\lambda\lambda 4959, 5007}{[OII]\lambda 3727}\right) \tag{1.14}$$

- for low metallicity ($12 + log(O/H) < 8.5$) they use the average of the Charlot and Longhetti (2001) (hereafter C01) method and of their calibration of the R_{23} method (see below):

$$12 + Log(O/H) = \frac{1}{2}(Z_{C01} + Z_{R_{23}}) \tag{1.15}$$

where

  - the C01 metallicity is given by

$$\begin{aligned} Z_{C01} = Log\Bigg\{5.09 \times 10^{-4} \left[\left(\frac{[OII]/[OIII]}{1.5}\right)^{0.17}\right] \times \\ \times \left[\left(\frac{[NII]/[SII]}{1.17}\right)^{0.85}\right]\Bigg\} + 12 \end{aligned} \tag{1.16}$$

– KD02 R_{23} calibration involves solving iteratively the system

$$\begin{cases} q = 10^{k_0+k_1R+k_2R^2} & \text{(1.17a)} \\ 12 + Log(O/H) = \dfrac{-\xi_1 + \sqrt{\xi_1^2 - 4\,\xi_2(k_0 - R_{23})}}{2\,\xi_2} & \text{(1.17b)} \end{cases}$$

where

$$R = Log\left(\frac{[OIII]}{[OII]}\right)$$

the constants $k_0, ..k_2$ depend on the metallicity value, while the constants $\xi_0, ..\xi_2$ depend on the ionization parameter q.

Finally, T04 obtain metallicity estimates by simultaneously fitting a set of various diagnostics ($[OII], H\beta, [OIII], H\alpha, [NII], [SII]$) to a library of models produced by the combination of population synthesis and photoionization codes, each characterized by a different choice of the main parameters (galaxy averaged metallicity, ionization parameter, dust attenuation at 5500 Å, and dust-to-metal ratio).
Then, for each galaxy they obtain the likelihood distribution of the metallicity, and assume its median as the final value, with a median 1 σ error of 0.03 dex.

1.2.2 Star metallicity determination

Gas-phase metallicity cannot be determined, of course, in gas-free galaxies, or in non active galaxies lacking the necessary emission lines.
Hence, when talking about the metallicity of these galaxies we are referring to their *stellar* metal content (usually represented through the iron abundance), which can be estimated by the study of the galactic SED. The integrated spectra of unresolved populations, in fact, display a variety of absorption features (whose strength depends on the abundance of specific elements in stellar atmospheres), that can be compared to theoretical predictions to constrain the ages and chemical composition of the ob-

served population.

The main tool used to theoretically model the spectro-photometric properties of galaxies in relation to their stellar content are the so-called Evolutionary Population Synthesis (EPS) models (Bruzual and Charlot, 2003; Maraston, 2003; Thomas et al., 2003; Peletier, 2013); by using the simplifying hypothesis that the populations in galaxies can be expressed as the sum of Single Stellar Populations (SSP) (groups of stars born at the same time, with the same metallicity) and making some assumptions on the IMF and SFR, these models allow us to calculate the age-dependent distribution of stars in a Hertzsprung-Russell (HR) diagram, from which the integrated spectral evolution of the stellar population can be obtained (Bruzual and Charlot, 2003; Maraston, 2003).
The two main "ingredients" of any stellar population synthesis models are (Maraston, 2003)

1. **stellar evolutionary tracks:** providing the evolution of the main parameters (magnitude, color and surface gravity) of a star of given mass and chemical composition as a function of the evolutionary time. Most model tracks in literature agree on the main stages of stellar evolution, whereas there are significant differences regarding the later evolutionary stages (especially the Horizontal Branch morphology); in particular, we can roughly divide tracks according to whether they take into account convective overshooting or not (Maraston, 2003).

2. **stellar spectral libraries:** which, assuming a model atmosphere, yield the flux of the stars corresponding to the values of the parameters provided by the evolutionary tracks, thus allowing the computation of the SED.

So, the SED at time t of a stellar population characterized by a SFR $\psi(t)$ and a metal-enrichment law $\xi(t)$ can be written as

$$F_\lambda(t) = \int_0^t \psi(t-t') S_\lambda[t', \xi(t-t')] dt' \tag{1.18}$$

where $S_\lambda[t', \xi(t-t')]$ is the power radiated per unit wavelength per unit initial mass by an SSP of age t and metallicity $\xi(t-t')$.

The strength of the various absorption features can be measured by the use of the so-called Spectral Indices (SI) (Cassisi and Salaris, 2013).
For narrow lines, the SI is measured in Angstroms and is basically the line's equivalent width, whose value is obtained by evaluating the difference between the line flux F_λ and a continuum value F_0 (represented by the straight line connecting the mean flux in the two bands delimiting the line), i.e.:

$$I_A = \int_{\lambda_1}^{\lambda_2} \left(1 - \frac{F_\lambda}{F_0}\right) d\lambda \tag{1.19}$$

while for broader features (as in molecular bands) the index is measured in magnitudes and expressed as

$$I_{mag} = -2.5\, Log \left[\left(\frac{1}{\lambda_2 - \lambda_1}\right) \int_{\lambda_1}^{\lambda_2} \frac{F_\lambda}{F_0}\, d\lambda \right] \tag{1.20}$$

The main issue arising with the use of spectral indices is the so-called Age-Metallicity Degeneracy (AMD), i.e. the fact that the same SI generally depends on both metallicity and age in the same way, so that it is hard to discriminate between the effects of these two parameters on the strength of the lines (Worthey et al., 1994; Cassisi and Salaris, 2013).
The task is not easier if we recur to the comparison of the photometric properties of the observed galaxies and the model, since they are also influenced by the AMD. Increasing either the age or the metallicity of the stars in a galaxy has a similar reddening effect on its broadband colors, so that young and metal-rich stellar populations produce optical colors which are indistinguishable from those produced by old and metal-poor populations. As the galaxy ages, in fact, more stars move to the giant branch on the Color-Magnitude Diagram (CMD), and in the same fashion increasing the metallicity of stars leads to an increase in their average photospheric opacity, which causes their effective temperatures (on which colors depend) to drop

(Worthey et al., 1994).
Worthey (1994) was the first to point out this effect, estimating that increasing the age of a factor 2 has the same effect on optical colors than increasing the metallicity of a factor 3 (the so-called *2/3 rule*).

One of the most frequently used sets of spectral indices is the Lick system, which mainly describes features in the blue-optical part of the spectrum (figure 1.1) and has been calibrated (Schiavon, 2007; Worthey et al., 1994) as a function of the main stellar parameters (T_{eff}, g, [Fe/H]). The Lick system somewhat provides the reasons

Name	Index band	Blue continuum	Red continuum	Units	Measures
$H\delta_A$	4083.500–4122.250	4041.600–4079.750	4128.500–4161.000	Å	
$H\delta_F$	4091.000–4112.250	4057.250–4088.500	4114.750–4137.250	Å	
CN_1	4142.125–4177.125	4080.125–4117.625	4244.125–4284.125	mag	C,N,(O)
CN_2	4142.125–4177.125	4083.875–4096.375	4244.125–4284.125	mag	C,N,(O)
Ca4227	4222.250–4234.750	4211.000–4219.750	4241.000–4251.000	Å	Ca,(C)
G4300	4281.375–4316.375	4266.375–4282.625	4318.875–4335.125	Å	C,(O)
$H\gamma_A$	4319.750–4363.500	4283.500–4319.750	4367.250–4419.750	Å	
$H\gamma_F$	4331.250–4352.250	4283.500–4319.750	4354.750–4384.750	Å	
Fe4383	4369.125–4420.375	4359.125–4370.375	4442.875–4455.375	Å	Fe,C,(Mg)
Ca4455	4452.125–4474.625	4445.875–4454.625	4477.125–4492.125	Å	(Fe),(C),Cr
Fe4531	4514.250–4559.250	4504.250–4514.250	4560.500–4579.250	Å	Ti, (Si)
$C_2$4668	4634.000–4720.250	4611.500–4630.250	4742.750–4756.500	Å	C,(O),(Si)
H_β	4847.875–4876.625	4827.875–4847.875	4876.625–4891.625	Å	
Fe5015	4977.750–5054.000	4946.500–4977.750	5054.000–5065.250	Å	(Mg),Ti,Fe
Mg_1	5069.125–5134.125	4895.125–4957.625	5301.125–5366.125	mag	C,Mg,(O),(Fe)
Mg_2	5154.125–5196.625	4895.125–4957.625	5301.125–5366.125	mag	Mg,C,(Fe),(O)
Mg_b	5160.125–5192.625	5142.625–5161.375	5191.375–5206.375	Å	Mg,(C),(Cr)
Fe5270	5245.650–5285.650	5233.150–5248.150	5285.650–5318.150	Å	Fe,C,(Mg)
Fe5335	5312.125–5352.125	5304.625–5315.875	5353.375–5363.375	Å	Fe,(C),(Mg),Cr
Fe5406	5387.500–5415.000	5376.250–5387.500	5415.000–5425.000	Å	Fe
Fe5709	5696.625–5720.375	5672.875–5696.625	5722.875–5736.625	Å	(C),Fe
Fe5782	5776.625–5796.625	5765.375–5775.375	5797.875–5811.625	Å	Cr
Na_D	5876.875–5909.375	5860.625–5875.625	5922.125–5948.125	Å	Na,C,(Mg)
TiO_1	5936.625–5994.125	5816.625–5849.125	6038.625–6103.625	mag	C
TiO_2	6189.625–6272.125	6066.625–6141.625	6372.625–6415.125	mag	C,V,Sc

Figure 1.1: From Cassisi and Salaris (2013) - Lick indices properties.

to break the degeneracy, since some of the indices have a different sensitivity to Z and age ($H\beta$ is more sensitive to age, while metallic indices like Fe5406 to Z); in a two-dimensional index-index diagram, lines of constant age and lines of constant Z (or [Fe/H]) are roughly orthogonal, meaning that the AMD is strongly softened (figure 1.2).

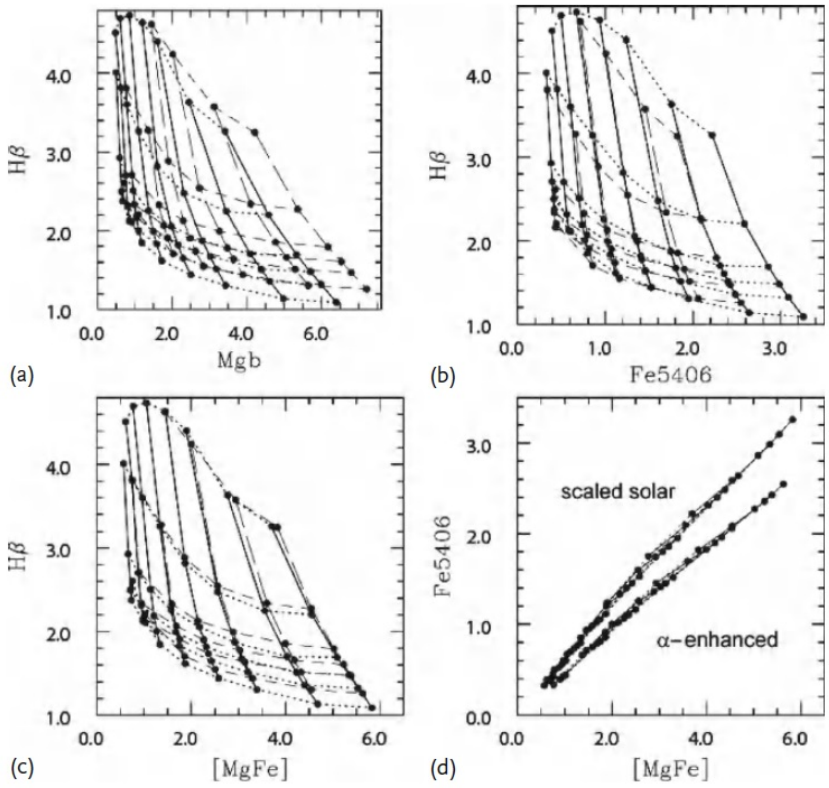


Figure 1.2: From Cassisi and Salaris (2013) - Index-index diagrams can be used to break the AMD.

In the same fashion, recent studies (Li et al., 2007; Li and Han, 2008) have investigated favourable colour combinations to minimize the AMD.

1.3 Correlation between metallicity and physical parameters

In this section, we investigate the relations existing between metallicity and other fundamental galactic parameters.
After providing a description of the simple "closed-box model", we proceed to discuss its consequences in terms of the existing relation between metallicity, stellar mass and luminosity.

1.3.1 Closed-box model of ISM enrichment

We will now briefly introduce a simple theoretical model describing the chemical enrichment of the ISM due to the production of heavy elements by stars; for the sake of simplicity, we will initially neglect the inflow/outflow of gas into/from the galaxy, so that this model is actually best suited to describe the gas-phase metallicity evolution of a *portion* of a galaxy, such as the Solar neighbourhood (Binney and Tremaine, 2011; Van den Bergh, 1962).

So, let us consider a region whose initial composition is ruled by gas, with no heavier elements, and assume that the gas distribution is kept homogeneous by turbulent motions; as time goes by, stars are formed, produce metals and return the processed materials to the interstellar medium, so that, since there are no inflows/outflows, the quantity of gas steadily reduces, and the ISM gets enriched with heavy elements.
By defining M_h and M_g as the metal and gas mass in the ISM respectively, and M_s to be the stellar mass at each time, the average gas metallicity will then be given by:

$$Z = \frac{M_h}{M_g} \tag{1.21}$$

We adopt the so-called **instantaneous recycling approximation** (Binney and Tremaine, 2011; Erb et al., 2006), i.e. we ignore the interval of time between the

formation of new stars and the ejection of the heavy elements produced into the ISM, which appears reasonable, if we consider that the lifespan of heavy mass stars producing metals is extremely short compared to the time scales of other evolutionary processes in the galaxy (Binney and Tremaine, 2011).
We assume that the quantity of metals returned to the ISM after the stars' death is proportional to the mass δM_s that remains locked in stellar remnants trough a coefficient y, the so-called **yield** of the generation of stars

$$\delta M_h^+ = y\,\delta M_s \tag{1.22}$$

So, the total mass of heavy elements present in the ISM changes by an amount of

$$\delta M_h = y\,\delta M_s - Z\,\delta M_s = (y - Z)\,\delta M_s \tag{1.23}$$

where we considered that M_h is increased due to the ejection of metals from the stars, but it is also reduced, since a part of the metals already present in the ISM remains locked in stellar remnants.
Using the condition

$$\delta M_s = -\delta M_g \tag{1.24}$$

which is a simple consequence of mass conservation, the latest equation can be rewritten as:

$$\delta M_h = (y - Z)\,\delta M_s = (Z - y)\,\delta M_g \tag{1.25}$$

Differentiating equation (1.21), we see that the metallicity changes according to

$$\delta Z = \delta\left(\frac{M_h}{M_g}\right) = \frac{\delta M_h}{M_g} - \frac{M_h}{M_g^2}\,\delta M_g = \frac{1}{M_g}(\delta M_h - Z\,\delta M_g) \tag{1.26}$$

so that, putting together equations (1.25) and (1.26), we obtain for the variation of Z:

$$\delta Z = -y \frac{\delta M_g}{M_g} \tag{1.27}$$

If the **yield** can be considered constant throughout the various generations of stars, we can integrate the latest equation from t=0 to t, thus obtaining:

$$Z(t) - Z(0) = -y\, ln \left(\frac{M_g(t)}{M_g(0)} \right) \tag{1.28}$$

where $M_g(0)$ is the total initial gas mass.
Since we assumed that the gas was initially metal-free (Z(0)=0), we can write

$$Z(t) = y\, ln \left(\frac{M_g(0)}{M_g(t)} \right) \tag{1.29}$$

So, the metallicity of the interstellar gas depends on the logarithm of the gas mass fraction μ, defined as:

$$\mu \equiv \frac{M_g}{M_{tot}} \tag{1.30}$$

i.e. the ratio between the total gas-phase mass and the initial total mass.

1.3.2 Mass-metallicity relation

By summarizing, under the assumptions that

- there are no inflows/outflows of gas,
- the system is initially composed by gas only, with no heavy elements,
- the gas is homogeneous,

- we ignore the time gap between the formation of a star generation and the ejection of metals,

the metallicity of the interstellar gas is a simple function of the stellar **yield** (the mass of metals produced in units of the mass that remains locked in stellar remnants) and of the gas mass μ_{gas} fraction, according to equation (1.31)

$$Z = y\, ln\left(\frac{1}{\mu_{gas}}\right) \tag{1.31}$$

An immediate consequence of this model is that it predicts the existence of a correlation between the gas-phase metallicity of a galaxy and its stellar mass since, as time passes, more and more gas is converted into stars and reprocessed as metals, so that we expect both the stellar mass and the average gas metallicity to increase.
Such a correlation has indeed been first observed by Lequeux et al. (1979), and is now well established in nearby galaxies; T04 quantified the relation for a sample of $\approx$ 53000 galaxies from the Sloan Digital Sky Survey (SDSS), while Erb et al. (2006) (hereafter E06) extended the relation for galaxies with $z \geq 2$.

It is worth to point out that, though we derived the mass-metallicity relation in the context of a simple closed-box model, the correlation between these quantities can be accounted for even considering a more realistic scenario, including the inflow/outflow of gas:

- low mass galaxies are observed to have higher gas fractions than high mass galaxies (McGaugh and De Blok, 1997), which could be imputed to a low gas surface density and a resulting inefficiency at turning gas into stars (Kennicutt Jr, 1998);

- stellar winds and supernovae explosions (SNe) could account for the correlation between mass and metallicity by expelling a greater amount of material (and, consequently, of metals) in low-mass galaxies, presumably because of the lower potential in these systems.
  In this sense, the closed-box model predicts that the metallicity of the interstellar gas is given by equation (1.31), so that by inverting this equation we can obtain an estimate of the "effective yield" from the observed metallicity and the gas mass fraction.

According to the closed-box model, we expect the yield to be constant, independently of the galaxy mass, but Garnett (2002) observed a decrease in the yield for galaxies with a rotational velocity less than $\approx$ 150 km/s which T04 interpreted as a consequence of metal loss via galactic winds; assuming that the depth of the galaxy's potential well scales with the rotational velocity as V_c^2, and that the galaxy's baryonic mass is given by $M_{bar} \propto V_c^{3.5}$, they obtained for the effective yield:

$$y_{eff} = \frac{y_0}{1 + (M_0/M_{bar})^{0.57}} \tag{1.32}$$

where y_0 is the true yield (the one obtained without metals outflow) and M_0 is the mass at which the galaxy loses 1/2 of its metals.

T04 mass-metallicity (MZ) relation

T04 obtained an estimate for the MZ relation on a sample of $\approx$ 53000 SDSS star-forming galaxies.
The result of their work is shown in figure 1.3. A correlation is evident, which is roughly linear from $10^{8.5}$ to $10^{10.5}$ $M_\odot$, after which a gradual flattening occurs.
Most remarkable is the tightness of the correlation: the 1 σ spread of the data around the median is $\pm$ 0.10 dex, with only a handful of extreme outliers present.
The relationship is well fitted by a polynomial of the form

$$12 + log(O/H) = -1.492 + 1.847(logM_*) - 0.08026(logM_*)^2 \tag{1.33}$$

where M_* is the stellar mass in units of solar masses.

E06 mass-metallicity (MZ) relation

E06 observed the MZ relation in high redshift galaxies at $z \geq$2.
The stellar masses used in their work are the integrated SFR over the lifetime of the galaxies, and so they represent the total mass of stars formed in this period and not the stellar mass at the moment of the observations.

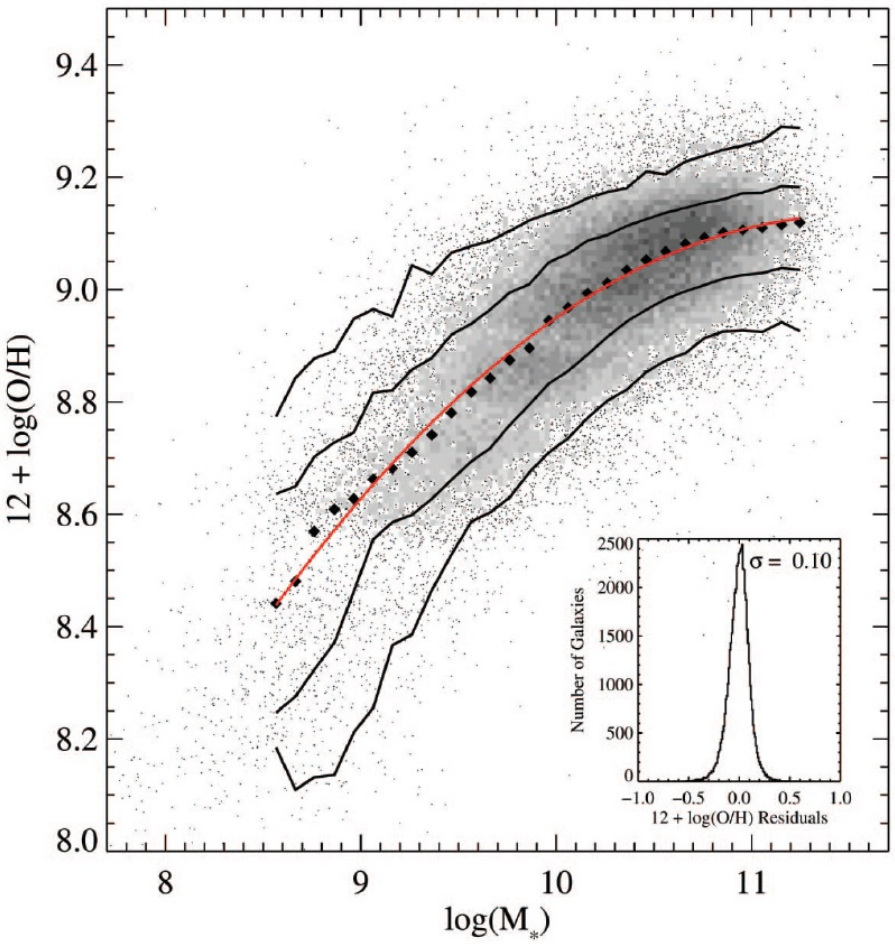


Figure 1.3: From T04 - Relation between stellar mass, in units of $M_{\odot}$, and gas-phase oxygen abundance for $\approx$ 53,000 star-forming galaxies in the SDSS. The large black filled diamonds represent the median in bins of 0.1 dex in mass that include at least 100 data points. The solid lines are the contours that enclose 68% and 95% of the data. The red line shows a polynomial fit to the data.

Figure 1.4 shows the plot of the mean metallicity (estimated from the N2 diagnos-

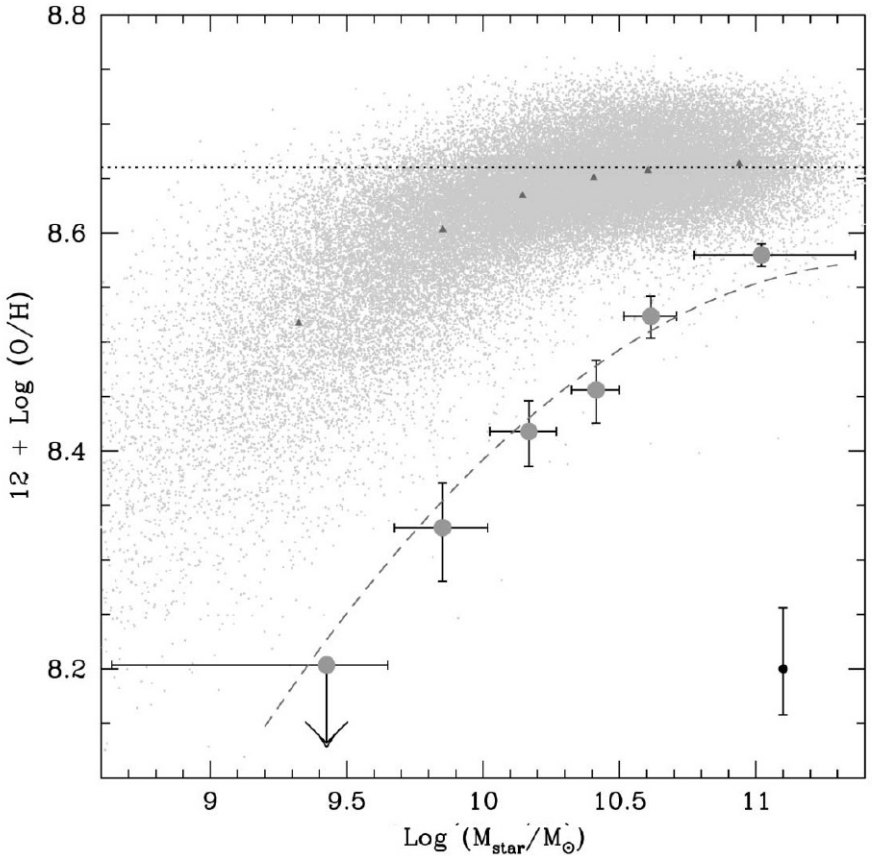


Figure 1.4: From Erb et al. (2006) - Observed relation between stellar mass and oxygen abundance at z≈2.

tic, the $[NII]/[H\alpha]$ ratio) vs. stellar mass (large grey filled circles); galaxies have been divided into mass bins, each containing 14-15 galaxies, and the horizontal bars represent the mass disperion in each of these bins.
The vertical error bars show the uncertainty in [12+log(O/H)], and an additional vertical error bar (the lower right corner) shows the uncertainty due to the intrinsic scatter in N2 calibration.
The dashed line shows the mass-metallicity relation determined by T04, with an

arbitrary downward shift of 0.56 dex, accounting for the different metallicity diagnostics used; with this shift, the SDSS relation matches the z=2 galaxies remarkably well.
Finally, the small grey dots show the metallicities of the T04 sample determined with the same N2 index used in E06, and the small triangles the metallicity in the same mass bins used in E06.

According to the two lowest mass bins, which can be considered the most reliable ones, galaxies at z $\approx$ 2 are $\approx$ 0.3 dex lower in metallicity than galaxies of the same stellar mass today.

Mass-Metallicity-Star Formation Rate (MZSFR) relation

As discussed above, the SFR plays a fundamental role in the balance determining the chemical composition of the interstellar gas in a galaxy.
Therefore, a relation between metallicity and SFR is likely to exist.

Mannucci et al. (2010) (hereafter M10) investigated this correlation by studying several samples of galaxies at different redshifts whose metallicity, stellar masses and SFRs have been measured; an estimate of the SFR is obtained by the $H\alpha$ luminosity, corrected for dust extinction, following Kennicutt Jr (1998):

$$SFR(M_\odot\, yr^{-1}) = \frac{L(H\alpha)}{1.26 \times 10^{41}\, erg\, s^{-1}} \tag{1.34}$$

M10 plotted the observed metallicity, M_* and SFR of their galaxies in a 3D space (figure 1.5), where the points distribution forms a tight surface in the space, the so-called FMR (Fundamental Metallicity Relation).
The scatter of the FMR is significantly smaller than the one characterizing most simple mass-metallicity relations ($\approx$0.08 dex), with a dispersion of individual galaxies around the FMR, that is about $\approx$0.06 dex when computed across the full FMR, and reduces to $\approx$0.05 dex in the central, most populated region; this seems to suggest that about half of the total scatter of the MZ relation is actually due to the systematic effect of SFR.

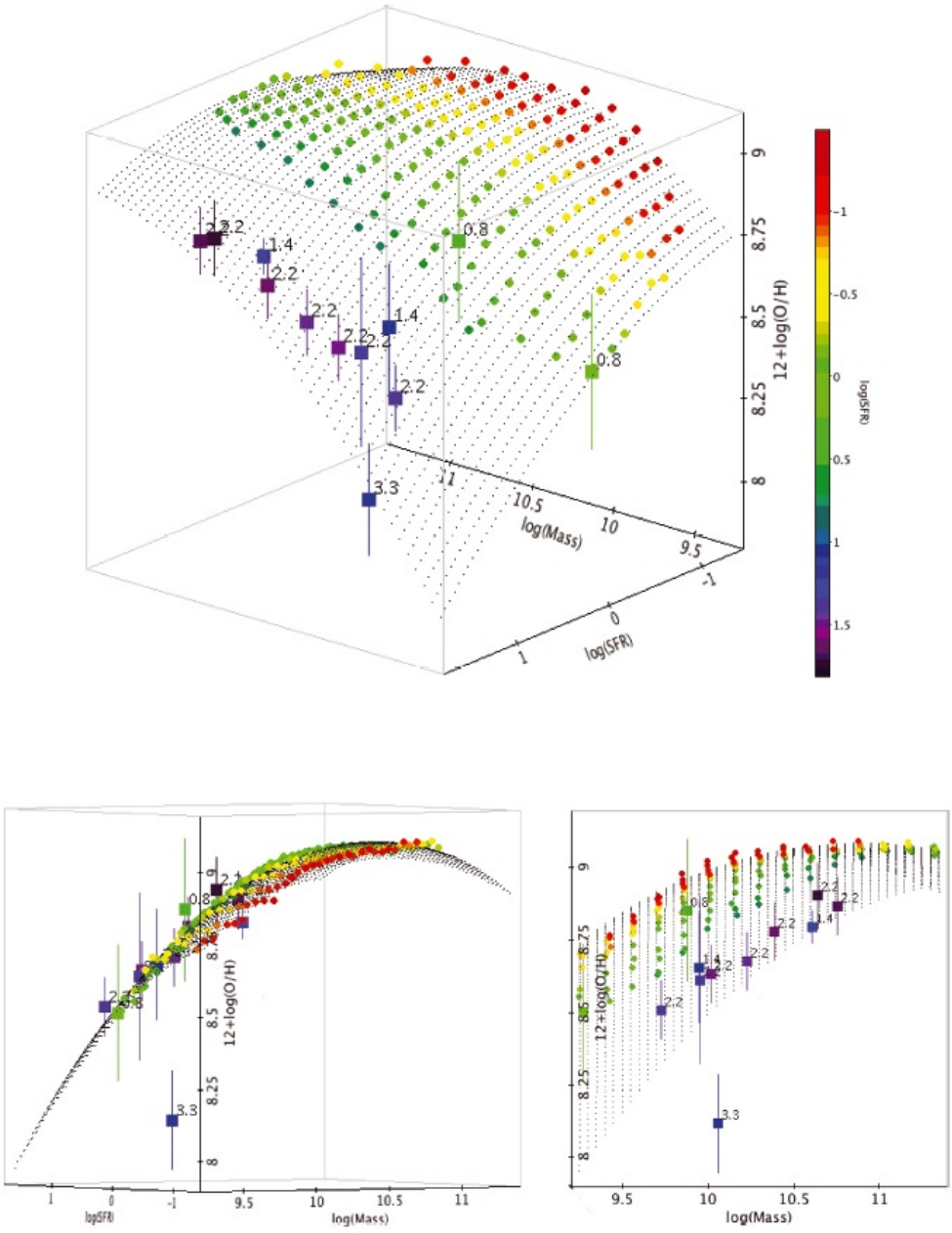


Figure 1.5: From M10 - The FMR relation (top), and two projections.

M10 fitted the FMR by the relation:

$$12 + log(O/H) = 8.90 + 0.37m - 0.14s - 0.19m^2 + 0.12ms - 0.054s^2 \tag{1.35}$$

with

$$\begin{cases} m = log(M_*) - 10 & \text{(1.36a)} \\ s = log(SFR) & \text{(1.36b)} \end{cases}$$

and also projected the FMR on an axis μ_α combining mass and SFR as

$$\mu_\alpha = log(M_*) - \alpha log(SFR) \tag{1.37}$$

where α is a free parameter (the best dispersion is obtained for $\alpha \approx 0.32$), thus obtaining

$$12 + log(O/H) = 8.90 + 0.39x - 0.20x^2 - 0.077x^3 + 0.064x^4 \tag{1.38}$$

with

$$x = \mu_{0.32} - 10 \tag{1.39}$$

Finally, the FMR somewhat shows a more fundamental nature than similar relations like the MZ or M-SFR, in that the former has been found to remain roughly constant when explored on galaxies at higher redshift (up to z $\approx$ 2.5).

1.3.3 Luminosity-Metallicity (LZ) relation

Because of the difficulties in deriving galaxy masses, there are much more luminosity-metallicity (LZ) relations than MZ relations in the literature, spanning up to 11 orders of magnitude in L and 2 dex in metallicity and observed in galaxies of all types (Erb et al., 2006; Garnett, 2002; Zaritsky et al., 1994; Lequeux et al., 1979).
In principle, these relations could allow us to derive metallicity only on the basis of photometric properties; in practice, however, the observed correlation presents a dispersion too great to provide good metal abundance estimates from luminosity alone (Sanders et al., 2013).

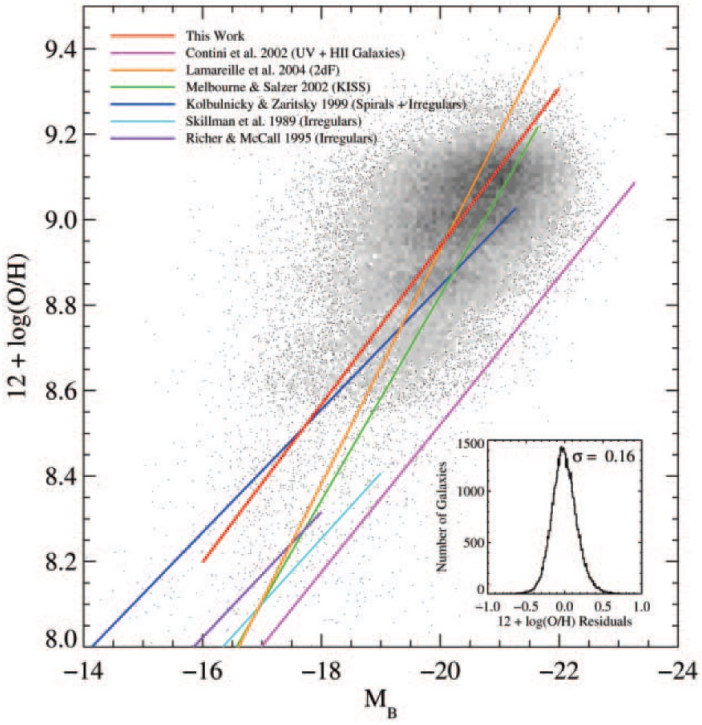


Figure 1.6: From T04 - LZ relations for SDSS galaxies and various galaxy samples drawn from the literature (see legend). The red line represents the linear least-squares bisector fit to the T04 SDSS data. The inset plot shows the residuals of the fit

T04 examined the LZ relation in their sample, compared to other works, using a linear least-squares technique. The measurement errors in luminosity (≈0.01 dex) and metallicity (≈0.1 dex) are small compared to the observed scatter of $\sigma \approx 0.16$ (see figure 1.6).
The luminosity-metallicity relation for their sample is

$$12 + log(O/H) = -0.185(\pm 0.001)M_B + 5.238(\pm 0.018) \tag{1.40}$$

At higher redshift, E06 constructed a LZ relation analogous to their MZ relation by dividing their sample galaxies into six bins by rest-frame absolute B magnitude M_B.
As shown in figure 1.7, the correlation between luminosity and metallicity is weaker than that between mass and metallicity, although the faintest galaxies do have the lowest metallicities.
In general, the comparison with the SDSS galaxies shows how high-redshift galaxies have both lower metallicities and higher luminosities than most of the local sample (in the more reliable lower metallicity bins the $z \approx 2$ galaxies are approximately 3 mag brighter than local galaxies with the same oxygen abundance).
The comparison, however, should be taken with a grain of salt due to the saturation of the lines used by E06 to determine metallicity around solar values.

Luminosity-Color-Metallicity (LCZ) relation

We saw that the LZ relation has a large intrinsic scatter, in terms of metallicity residuals, which limits the utility of this relation as an effective indicator of metallicity.
There are two primary causes for the scatter in the luminosity-metallicity relation:

- while the scatter in the mass-metallicity relation is fairly small ($\sigma \approx 0.10$ dex, T04), luminosity is not a perfect proxy for mass, since the M/L ratio of galaxies is highly correlated with galaxy color (redder galaxies at a fixed luminosity are more massive);

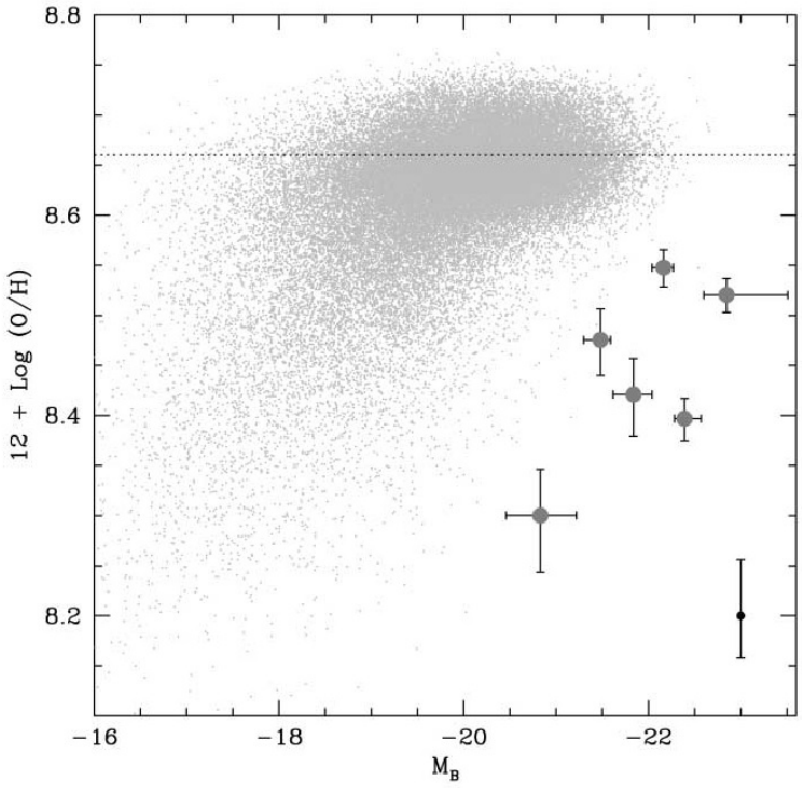


Figure 1.7: From Erb et al. (2006) - LZ relation at z≈2. The sample has been divided into six bins by rest-frame absolute B magnitude, and the metallicity has been estimated in each bin.

- as seen above, a more fundamental relation has been uncovered between mass (M), metallicity (Z), and SFR which has remarkably small residual scatter ($\sigma \approx$0.05 dex), indicating that variations in SFR are responsible for much of the scatter in the mass-metallicity relation.

Sanders et al. (2013), hereafter S13, showed that the addition of color information significantly decreases the scatter in photometric metallicity estimates.
Following M10, they projected the so found LZC relation onto an axis μ with components of color and luminosity, defining

$$\mu = M_i - \alpha \times (m_i - m_j) \tag{1.41}$$

where i, j are the selected bands, and obtained

$$12 + log(O/H) = p_0 + p_1\mu + p_2\mu^2 + p_3\mu^3 \tag{1.42}$$

S13 found the optimal value of the parameter α for different choices of the metallicity diagnostic, luminosity band and color. Figure (1.8) demonstrates this optimization for the g-r filter, the g band magnitude and three Z diagnostic (T04, KD02, and Pettini and Pagel (2004) calibration of O3N2). The metallicity scatter shows a clear improvement as compared to the LZ relation alone ($\sigma_Z = 0.10$ for α= 0), with the best results obtained for the O3N2 metallicity diagnostic ($\sigma_Z = 0.07$ dex).
Interestingly, regardless of the choice of diagnostic or filters, the residual scatter is lower for asymptotically high values of α than for $\alpha = 0$, which means that, in general, **color is more effective than luminosity as a predictor of metallicity.**

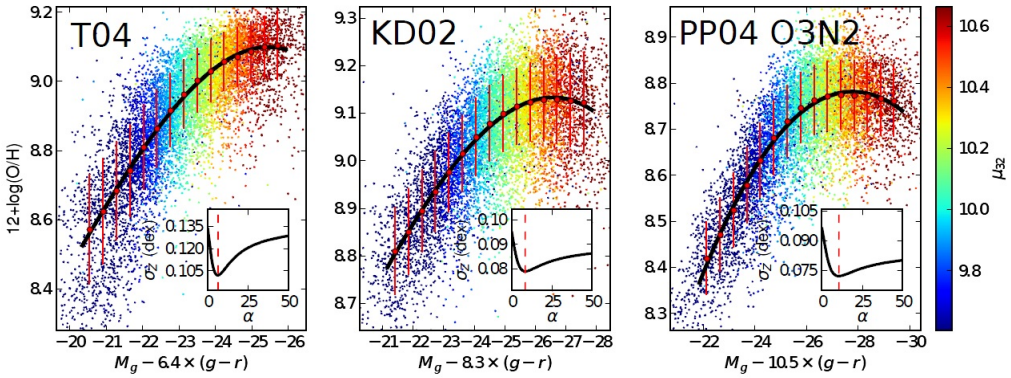


Figure 1.8: From Sanders et al. (2013) - The optimal projection of the LZC relation for M_g, g-r color, and three different metallicity diagnostics. The red points and lines show the median and standard deviation of the metallicity for galaxies in 15 bins. The projected LZC relation is shown for the optimal value of α, with the best fit third order polynomial LZC relation in black. The color coding shows the optical physical parameter (μ_{32}) from M10.

Chapter 2

Machine Learning methods

In this work, we describe the application of Machine Learning techniques to the problem of photometric metallicity determination; specifically, we used datasets of galaxies, whose gas-phase and stellar metal content has been already determined by means of classic spectroscopic methods, as Knowledge Base (KB) to train a Multi-Layer Perceptron (MLP) neural network with a Quasi-Newton Algorithm (QNA).
A theoretical overall description of Machine Learning methods and their application to Astrophysics can be found in Brescia (2012), Brescia and Longo (2013) and Cavuoti (2013); these techniques have been already successfully applied to solving astrophysical problems, like the determination of photometric redshifts (Brescia et al., 2013; Cavuoti et al., 2012), the detection of globular clusters (Brescia et al., 2012) or the photometric classification of emission line galaxies (Cavuoti et al., 2014).

Astronomy, as well as many other disciplines, is undergoing a process of deep evolution, driven by the advent of large digital sky surveys providing us with an unprecedented amount of data, either in the form of tables, images, spectra or datacubes.
The sheer volume of available data has steadily been increasing, and this trend is reasonably bound to continue; the present catalogues contain hundreds of millions of objects, and the explored wavelength regions are constantly expanding, now encompassing almost the whole electromagnetic spectrum. The greatest scientific potential of surveys, however, lies well beyond the sheer increase in the *quantity* of available data. The possibility to cross-match objects among different catalogues allows us to investigate processes that can only be understood via a multi-band analysis, like the

classification of quasars, the study of ultraluminous starbursts or the interpretation of γ-ray bursts, while the use of surveys that cover large areas of the sky repeatedly paves the way to a time-dependent analysis of the sky (Djorgovski et al., 2013).

A definitely not comprehensive summary of some popular surveys is shown in figure

Survey	Type	Duration	Bandpasses	Lim. flux	Area coverage	$N_{sources}$	Notes
DSS scans	Visible	1950's-1990's	B (~ 450 nm) R (~ 650 nm) I (~ 800 nm)	21 – 22 mag 20 – 21 mag 19 – 20 mag	Full sky	~ 10^9	Scans of plates from the POSS and ESO/SERC surveys
SDSS-I SDSS-II SDSS-III	Visible	2000-2005 2005-2008 2009-2014	u (~ 800 nm) g (~ 800 nm) r (~ 800 nm) i (~ 800 nm) z (~ 800 nm)	22.0 mag 22.2 mag 22.2 mag 21.3 mag 20.5 mag	14,500 deg^2	4.7×10^8	Numbers quoted for DR8 (2011). In addition, spectra of ~ 1.6 million objects
2MASS	Near IR	1997-2001	J (~ 1.25 μm) H (~ 1.65 μm) K_s (~ 2.15 μm)	15.8 mag 15.1 mag 14.3 mag	Full sky	4.7×10^8	
UKIDSS	Near IR	2005-2012	Y (~ 1.05 μm) J (~ 1.25 μm) H (~ 1.65 μm) K (~ 2.2 μm)	20.5 mag 20.0 mag 18.8 mag 18.4 mag	7,500 deg^2	~ 10^9	Estim. final numbers quoted for the LAS; deeper surveys over smaller areas also done
IRAS	Mid/Far IR (space)	1983-1986	12 μm 25 μm 60 μm 100 μm	0.5 Jy 0.5 Jy 0.5 Jy 1.5 Jy	Full sky	1.7×10^5	
NVSS	Radio	1993-1996	1.4 GHz	2.5 mJy	32,800 deg^2	1.8×10^6	Beam ~ 45 arcsec
FIRST	Radio	1993-2004	1.4 GHz	1 mJy	10,000 deg^2	8.2×10^5	Beam ~ 5 arcsec
PMN	Radio	1990	4.85 GHz	~ 30 mJy	16,400 deg^2	1.3×10^4	Combines several surveys
GALEX	UV (space)	2003-2012	135 – 175 nm 175 – 275 nm	20.5 mag AIS 23 mag MIS	AIS 29,000 MIS 3,900	6.5×10^7 1.3×10^7	As of GR6 (2011); also some deeper data
Rosat	X-ray (space)	1990-1999	0.04 – 2 keV	~ 10^{-14} erg cm^{-2} s^{-1}	Full sky	1.5×10^5	Deeper, small area surveys also done
Fermi LAT	γ-ray (space)	2008-?	20 MeV to 30 GeV	4×10^{-9} erg cm^{-2} s^{-1}	Full sky	~ 2×10^3	LAT instrument only; in addition, GRBM

Figure 2.1: From Djorgovski et al. (2013) - Basic properties of some of the popular wide-field surveys.

2.1.

It appears evident how this "data tsunami", though providing us with the clues to get to a better knowledge of our Universe, also establishes the need for the development of new instruments to deal with this huge incoming flow of information, to the point that Data Mining (also known as Knowledge Discovery in Databases, KDD), i.e. extraction of knowledge from massive data archives, has been indicated by Hey et al. (2009) as the fourth pillar of modern science, after theory, experiments and

simulations.
The need to tackle this task in an efficient way has led to the emergence of a new discipline called Astro-informatics (in general, we talk about X-informatics, where X can be any science, such as bio, astro, eco...; Brescia and Longo, 2013; Brescia, 2012).
By combining astronomy, statistics and computer science, the purpose of Astro-informatics is to provide us with automated tools for:

- archiving and organizing huge amounts of data;
- extracting, processing and integrating data from different sources;
- representing and accessing data;
- most importantly, reproducing the processes of a human brain in terms of learning from and make predictions on data (Machine Learning). Specifically, the greatest challenge lies in the attempt to replicate the human ability to learn from examples used to train the algorithms, and to use the acquired knowledge to generalize the process of analysis to unknown data.

2.1 Neural networks

Neural Networks are a powerful machine learning method for dealing with supervised data mining problems, such as classification and regression.
Replicating the working principles of human brain, these algorithms create a network of "connections" between a group of input features and their corresponding output, and try to reproduce the hidden relationship which is assumed to exist between these input features and the output by modifying the strength of these connections.

The Multi-Layer Perceptron model we used differs from simpler implementations, like the Single Layer Perceptron (Brescia, 2012; McCulloch and Pitts, 1943), in that the input and output neurons are not directly connected, but are separated by one or more intermediate hidden layers, each of which can be formed by an arbitrary number of neurons (it can be shown that the addition of an intermediate layer introduces a new hyperplane able to correctly separate data in classification problems - Cybenko, 1989); the neurons in each layer are connected to all the other nodes in

the adjacent layers (such a model is said to be *fully connected*).
Figure (2.2), for instance, represents a MLP with two input neurons, one hidden layer formed by three nodes (in white) and an output neuron.

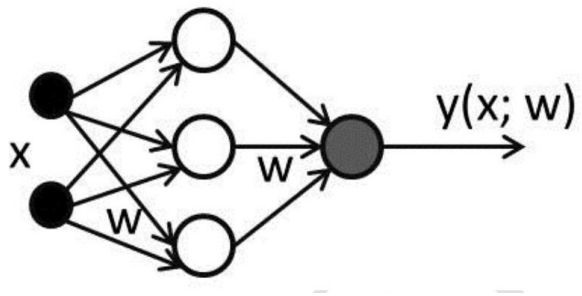


Figure 2.2: Schematic representation of a Multi-Layer Perceptron.

The network shown in figure 2.2 allows us to represent the nonlinear function of many variables by a composition of nonlinear activation functions of one variable.
By indicating with d the number of input neurons and with M the number of nodes in the intermediate layer, the output y_k (the value assumed by the k-th output neuron) can be formally expressed as

$$y_k = \sum_{j=0}^{M} w_{kj}^{(2)}\, g\left(\sum_{i=0}^{d} w_{ji}^{(1)}\, x_i\right) \tag{2.1}$$

where

- $w_{jk}^{(i)}$ represents the weight of the connection between the j-th neuron in the (i+1)-th layer and the k-th neuron in the i-th layer;
- g is the activation function, representing the response of the intermediate layer neurons to the input ones (the response of the output layer is assumed to be linear).

As for the activation function, we usually choose a smoothed version of a step function like the hyperbolic tangent (fig 2.3); though somewhat preserving the "on-off" response to the input of the step function, where the neuron is activated if the input

is greater than the threshold, and remains turned off otherwise, a sigmoid keeps information, at least to a certain degree, on the input values, which is essential to the training process.

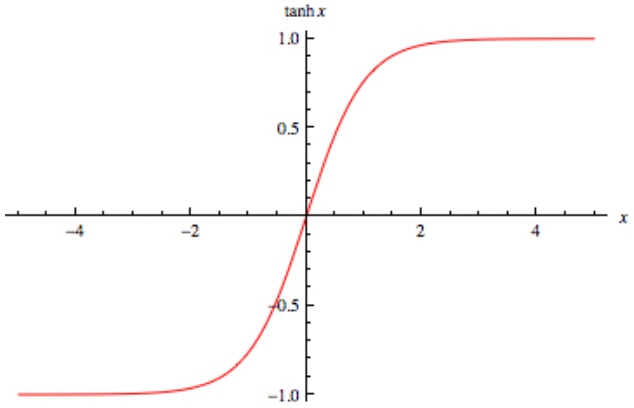


Figure 2.3: Hyperbolic tangent activation function.

The training process of a neural network involves the use of a Knowledge Base (KB), i.e. a dataset of objects whose input and output features (in our case, photometric properties and metallicity respectively) are already known; such an approach is called supervised learning (Brescia, 2012).
Using the KB, for each training pattern of features p it is possible to define an *error function* E_p, representing the difference between the expected and calculated output of the network.
Notice that the weights are generally initialized to small random values, so that the initial state of the network is random as well.
Once the error value is evaluated, a backward phase starts where an optimization rule is applied to modify the weights of the network, with the aim to find the values that minimize E_p; this process is reiterated until the error becomes lower than a predetermined threshold, or when a maximum number of iterations is reached.
At the end of the training phase, the network will act as a simple function of the input features; it can be presented with new input data never used in the training phase, and it will provide the corresponding correct output (if the training phase was successful).

2.1.1 Quasi-Newton Algorithm (QNA) optimization rule

There exist many different optimization rules that can be applied to train a neural network.
Generally, they consist on iterative processes where at each step the weights vector $\boldsymbol{w}_{t+1}$ is modified with respect to its previous value $\boldsymbol{w}_t$ by

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \alpha_t \, \boldsymbol{d}_t \tag{2.2}$$

where

- d_t is a generic descent direction, i.e. a direction satisfying the condition

$$\boldsymbol{d_t} \cdot \boldsymbol{\nabla}_w E < 0 \tag{2.3}$$

  which guarantees that E_p is reduced along d_t.
  In the most general case, we assume d_t to take the form

$$\boldsymbol{d}_t = -B_t^{-1} \, \boldsymbol{\nabla} E_t \tag{2.4}$$

  where B_t is a symmetric nonsingular matrix.

- the entity of the variation α_t is obtained at each step by a *line search*; given d_t, we study the behavior of the function and find its minimum along this direction, thus basically reducing the problem to a single-dimension minimization. Specifically, the parameter α_t is chosen so that

$$\frac{\partial}{\partial \alpha} \left[E(\boldsymbol{w}_t + \alpha_t \, \boldsymbol{d}_t) \right] = 0 \tag{2.5}$$

The learning rule we applied in the training phase of the neural network is the Quasi-Newton Algorithm, a modified version of the Newton Method (Brescia, 2012; Brescia et al., 2013).
The latter is an optimization algorithm used to find the stationary points of a twice-differentiable function, which is achieved by making a local square approximation of the function using a Taylor expansion.
The Newton method uses the inverse Hessian matrix, which is the coefficient of the quadratic term in the Taylor expansion, as the matrix B_t to determine the descent direction in equation (2.4), so that we have

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \alpha_t \left[H(E_t)\right]^{-1} \boldsymbol{\nabla} E_t \tag{2.6}$$

where, as usual, the step length α_t is found by a line search.
The Newton algorithm is usually extremely slow; in particular, if the function is quadratic the exact extremum is found in only one step, and it can be shown that every local minimum has a neighbourhood such that, if we start the minimization process from a point w_0 within it, the process converges quadratically.
Using second derivatives combined with the gradient in determining the descent direction has the effect of giving the method a "natural" ability to avoid local extrema and find the absolute minimum of the function.
However, the Hessian of a function is not always available (in many cases it is simply too complex to be calculated analytically), and even when it can be evaluated, by storing its value at each iteration, it can be extremely memory consuming.
The use of the Quasi-Newton Algorithm (QNA) allows us to overcome these issues and to save resources by generating a series of matrices G_t, obtained starting from the identity matrix I by gradient calculations only, that are increasingly good approximations of H; we expect that the matrices G_t could preserve the main features of the Hessian, for instance:

- symmetry: the Hessian matrix (and its inverse) are symmetric, and so we want G_t to be symmetric as well;
- positive definiteness: this ensures that d_t will be a descent direction;
- quasi-Newton condition: the local square approximation for the error function, which is particularly accurate in the proximity of a minimum w^*, allows us to

express the minimum position as

$$\boldsymbol{w}^* = \boldsymbol{w} - H^{-1} \times \boldsymbol{\nabla} E \tag{2.7}$$

where the vector $H^{-1} \times \boldsymbol{\nabla} E$ is known as a **Newton direction**.
Using this equation, we can express the difference between the weight vectors on steps t and (t+1) as

$$\boldsymbol{w}_{t+1} - \boldsymbol{w}_t = -[H(E_{(t+1)})]^{-1} \times (\boldsymbol{\nabla} E_{t+1} - \boldsymbol{\nabla} E_t) \tag{2.8}$$

which is known as *quasi-Newton condition.*

There are various implementations of the QNA; a very known one is the **BFGS** (by the names of its inventors - Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970), for which the expression of G is given by

$$G_{(t+1)} = G_{(t)} + \frac{\boldsymbol{p}\,\boldsymbol{p}^T}{\boldsymbol{p}^T\,\boldsymbol{\nu}} - \frac{(G_{(t)}\boldsymbol{\nu})\,\boldsymbol{\nu}^T\,G_{(t)}}{\boldsymbol{\nu}^T\,G_{(t\,\boldsymbol{\nu})}} + \left(\boldsymbol{\nu}^T\,G_{(t)}\,\boldsymbol{\nu}\right)\boldsymbol{u}\,\boldsymbol{u}^T \tag{2.9}$$

where we introduced the vectors

$$\begin{cases} \boldsymbol{p} = \boldsymbol{w}_{t+1} - \boldsymbol{w}_t; & \text{(2.10a)} \\ \boldsymbol{\nu} = \boldsymbol{\nabla} E_{t+1} - \boldsymbol{\nabla} E_t; & \text{(2.10b)} \\ \boldsymbol{u} = \dfrac{\boldsymbol{p}}{\boldsymbol{p}^T\,\boldsymbol{\nu}} - \dfrac{G_t\,\boldsymbol{\nu}}{\boldsymbol{\nu}^T\,G_t\,\boldsymbol{\nu}} & \text{(2.10c)} \end{cases}$$

and the weight correction is

$$\boldsymbol{w}_{(t+1)} = \boldsymbol{w}_t + \alpha_{(t)}\,G_{(t)}\,\boldsymbol{\nabla} E_{(t)}$$

where the step length α is obtained by line search.

Unfortunately, however, even calculating the matrix G becomes very memory consuming for large values of w.
A slightly modified version of the BFGS algorithm (**L-QNA**, or Limited memory QNA) allows to overcome this problem, by replacing at each step the matrix G with a unitary matrix and multiplying by the gradient, thus obtaining

$$\boldsymbol{d}_{(t+1)} = -\boldsymbol{\nabla E}_{(t)} + A\boldsymbol{p} + B\boldsymbol{\nu} \tag{2.11}$$

with

$$\begin{cases} A = -\left(1 + \dfrac{\boldsymbol{\nu}^T \boldsymbol{\nu}}{\boldsymbol{p}^T \boldsymbol{\nu}}\right) \dfrac{\boldsymbol{p}^T \boldsymbol{\nabla} E_{(t+1)}}{\boldsymbol{p}^T \boldsymbol{\nu}} + \dfrac{\boldsymbol{\nu}^T \boldsymbol{\nabla} E_{(t+1)}}{\boldsymbol{p}^T \boldsymbol{\nu}} & \text{(2.12a)} \\ B = \dfrac{\boldsymbol{p}^T \boldsymbol{\nabla} E_{(t+1)}}{\boldsymbol{p}^T \boldsymbol{\nu}} & \text{(2.12b)} \end{cases}$$

So, the algorithm to train a MLP with the L-QNA learning rule can be summarized as:

1. initialize the weights with small random values, tipically normalized in [-1,1], and extracted by the uniform probability distribution;
2. present to the network all the training sets using equation (2.1), and calculate the error function E between output and target values;
3. select the next search direction:
   - if t=0, then use
   $$\boldsymbol{d}_t = -\boldsymbol{\nabla} E_t$$
   - otherwise
   $$\boldsymbol{d_t} = -\boldsymbol{\nabla} E_t + A\boldsymbol{p} + B\boldsymbol{\nu}$$

where p, ν, A and B are given by equations (2.10a), (2.10b), (2.12a) and (2.12b), respectively;

4. update the weight vector by

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \alpha_t \, \boldsymbol{d}_t$$

where the value of α_t is determined by a line search;

5. calculate A and B for the next iteration;

6. if the error is less than some predetermined threshold ϵ stop the algorithm, otherwise go to 2.

2.2 Statistics for experiments

The results of regression experiments performed using a neural network can be evaluated using the following statistical indicators:

- **bias:** the mean of the differences ΔZ between the expected and calculated metallicity value;
- **scatter (σ):** the standard deviation of the differences ΔZ.

In case of classification problems, a simple way to evaluate the classifier's performance is to consider its confusion matrix (also known as *contingency table*, or *error matrix*), reporting the number of objects belonging to various classes and the categories they have been classified to (Brescia, 2012; Cavuoti et al., 2014); looking at table 2.1, we have that:

- N_{AA}: number of objects belonging to Class A, correctly classified as Class A;
- N_{AB}: number of objects belonging to Class A, wrongly classified as Class B;
- N_{BA}: number of objects belonging to Class B, wrongly classified as Class A;

		Classified as	
		Class A	Class B
Known class	Class A	N_{AA}	N_{AB}
	Class B	N_{BA}	N_{BB}

Table 2.1: Confusion matrix defined for a classification experiment.

- N_{BB}: number of objects belonging to Class B, correctly classified as Class B.

From these quantities, we can define the following useful statistical indicators:

- **average efficiency** (t_e). It is defined as the ratio between the number of correctly classified objects and the total number of objects in the data set. In our confusion matrix example, it would be:

$$t_e = \frac{N_{AA} + N_{BB}}{N_{AA} + N_{AB} + N_{BA} + N_{BB}} \tag{2.13}$$

- **purity of a class (pcN)**: Defined as the ratio between the number of correctly classified objects of a class and the number of objects that have been classified in that class, also known as *efficiency* or *precision* of a class. In our confusion matrix example it would be:

$$\begin{cases} pcA = \dfrac{N_{AA}}{N_{AA} + N_{BA}} & \text{(2.14a)} \\ pcB = \dfrac{N_{BB}}{N_{AB} + N_{BB}} & \text{(2.14b)} \end{cases}$$

- **completeness of a class (cmpN)**: Defined as the ratio between the number of correctly classified objects in that class and the total number of objects that are supposed to be in that class (actually belonging to that class), also known

as *recall*. In our confusion matrix example it would be:

$$\begin{cases} cmpA = \dfrac{N_{AA}}{N_{AA} + N_{AB}} & \text{(2.15a)} \\ cmpB = \dfrac{N_{BB}}{N_{BA} + N_{BB}} & \text{(2.15b)} \end{cases}$$

- **contamination of a class (cntN)**: It is the dual of the purity, namely it is the ratio of misclassified object in a class and the number of objects that have been classified in that class. In our example, it is given by:

$$\begin{cases} cntA = 1 - pcA = \dfrac{N_{BA}}{N_{AA} + N_{BA}} & \text{(2.16a)} \\ cntB = 1 - pcB = \dfrac{N_{AB}}{N_{AB} + N_{BB}} & \text{(2.16b)} \end{cases}$$

2.3 The DAMEWARE web application

The Data Mining and Exploration (DAMEWARE) web application resource is one of the products of a joint effort between the Astroinformatics groups at University Federico II, the Italian National Institute of Astrophysics and the California Institute of Technology (Brescia et al., 2014).
Conceived and engineered in 2007, DAMEWARE has been offered to the public since early 2012 (at the URL `http://dame.dsf.unina.it/dameware.html`), to enable a generic user to perform data mining and exploratory experiments on large data sets (of the order of a few tens of gigabytes). By exploiting web 2.0 technologies, it offers several tools which can be seen as working environments within which to choose data analysis functionalities such as clustering, classification, regression, feature selection, etc., together with models and algorithms.

Using DAMEWARE, any user can setup, configure, and execute experiments on his own data, on top of a virtualized computing infrastructure, without the need to install any software on his local machines.
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, DAMEWARE 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. (2014).

Chapter 3

Metallicity experiments

3.1 Experiments overview

In this chapter, we describe our work on the photometric catalogues, carried out in order to train the neural network to produce reliable metallicity predictions.
The network model used is a MLP with *two* hidden layers; for the determination of the ideal topology for the network (i.e. the best number of hidden layers and of nodes in each layer), we applied the rule of thumb that, being N the number of nodes in the input layer, the first hidden layer is formed by (2N+1) neurons, while the second by (N-1).
As described before, the learning rule adopted in the training phase was the QNA. More details on the model can be found in Appendix A.

The work flow we followed is represented in figure 3.1, and can be roughly summarized as follows.

First, we focused on the problem of *stellar* metallicity determination; as reported

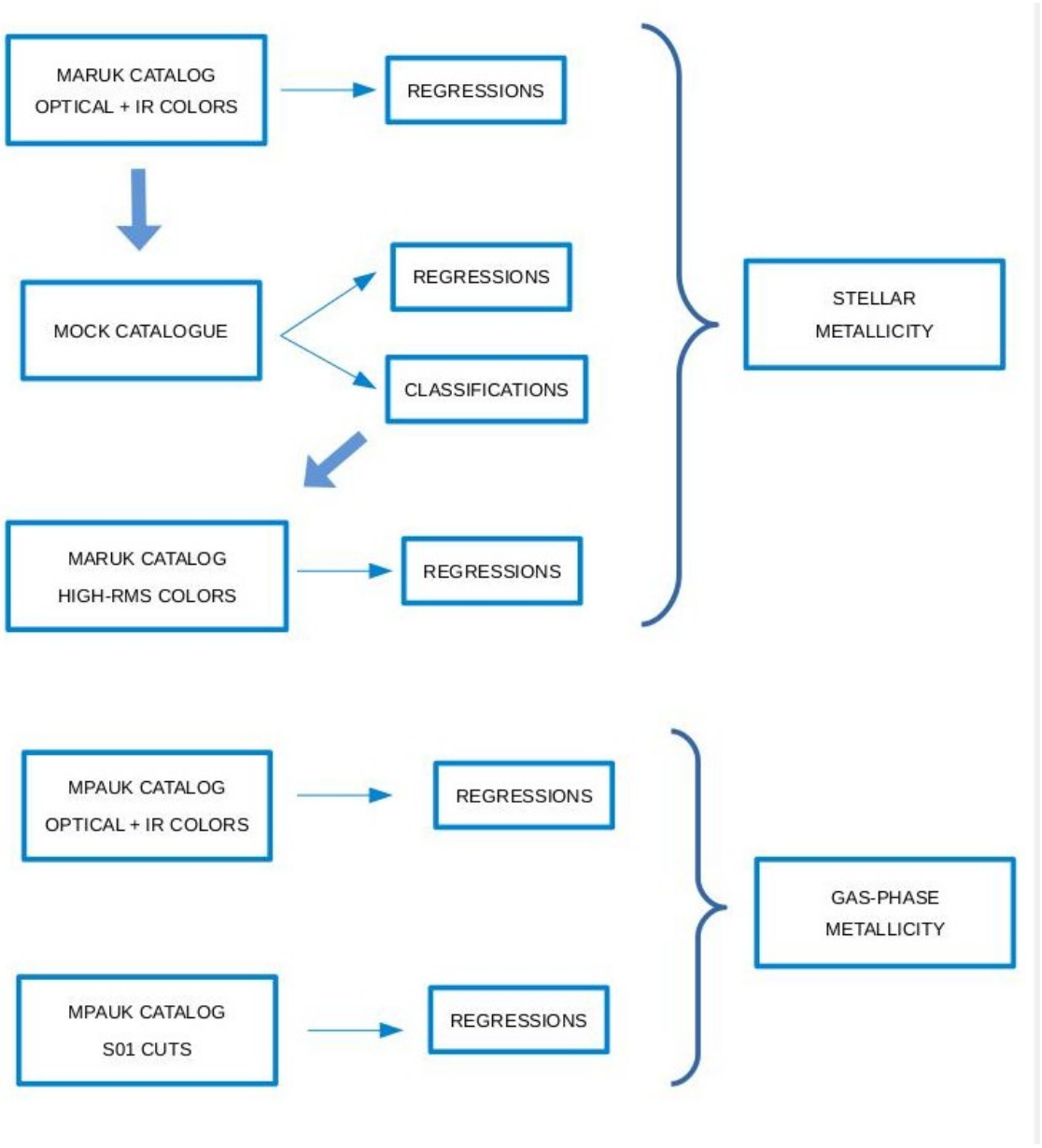


Figure 3.1: Work overview.

in section 3.3.1, we performed a preliminary series of regression experiments using optical and infrared colors as input features on the MARUK catalogue, which we described in section 3.2.1.
The results obtained in this phase, however, showed some difficulties of the neural network in predicting the expected metallicity values; this led us to create a mock catalogue of galaxies, and to test the performances of the network on these objects, hoping that having the complete control on the physical parameters of the dataset could help us to find the right configuration of the network, as well as to verify its real capabilities to predict metallicity. General details on the code used to create the simulated data are provided in Appendix B, while the mock catalogue itself is described in section 3.2.2.
In particular, for the experiments on the simulated dataset, we determined the most metallicity-sensitive colors by following Li et al. (2007) and Li and Han (2008), and used them as input features; these regressions, as described in section 3.3.2, showed a significant improvement over the previous ones. Moreover, since the metallicity of simulated galaxies could only assume discrete values, we further tested the network on the mock catalogue by performing a series of classification experiments, as detailed in section 3.3.3.
Finally, we tried to exploit the positive results obtained, by using the colors which had proven to be the most effective on the simulated data as input features for a new series of regressions performed on the MARUK dataset again; the experiments and their results are described in section 3.3.4.

Then, we moved on to gas-phase metallicity determination, using the MPAUK dataset described in section 3.2.3.
First, we performed a set of regression experiments using optical and infrared colors as input features on the whole database (section 3.4.1); then, we tried to refine the obtained results, and applied a series of cuts on the datasets by following S13. These experiments and their output are describe in section 3.4.2.

3.2 Datasets

In the following, we described the three main photometric datasets used in our work; the MARUK stellar metallicity dataset, the mock catalogues created by using the BC03 EPS code, and the MPAUK dataset used to determine gas-phase metallicity.

3.2.1 Stellar metallicity catalogue

The **Sloan Digital Sky Survey (SDSS)** is an imaging and spectroscopic survey, actually providing deep multi-color images of one third of the sky, and spectra for more than three million astronomical objects. The survey uses a dedicated 2.5-m wide-angle optical telescope at Apache Point Observatory in New Mexico, United States (Gunn et al., 2006) to obtain photometric imaging in five optical bands (Fukugita et al., 1996; Smith et al., 2002) with a drift scan camera (Gunn et al., 1998); the SDSS original optical spectrograph and the BOSS upgrade are described in Smee et al. (2013).
Operating since 2000, the survey has now reached its third phase of operations (SDSS-III) with its twelfth Data Release (DR12); see (York et al., 2000) for an overall technical description of the SDSS-I and SDSS-II (DR 1-7) and (Eisenstein et al., 2011) for the SDSS-III (DR 8-12).

The first dataset we used was provided to us by the Portsmouth Institute of Cosmology and Gravitation (ICG) group led by Prof. Claudia Maraston (Maraston et al., 2009, 2013). It contained the age, estimated E(B-V), stellar mass and metallicity for 465,289 objects from the DR9 (Ahn et al., 2012); in addition, we retrieved from the SDSS archive photometric information (model, cmodel, petrosian and 3" fiber magnitudes in the ugriz bands) and the spectroscopic redshifts for all these objects. A further examination of the dataset revealed the presence of stars and QSOs, so we cross-matched the catalogue with the SDSS again and only kept the 459,624 objects spectroscopically identified as galaxies; of these, we eliminated all the non occurrences (Not a Number), thus reducing the dataset to 395,112 elements.

In order to obtain information in the infrared (IR) bands, we performed a positional cross-match of our objects with the **UKIDSS** Large Area Survey (LAS) catalog, selecting for each SDSS object the nearest UKIDSS one within a radius of 0.4 arcseconds.
The UKIDSS project is defined in Lawrence et al. (2007). UKIDSS uses the UKIRT Wide Field Camera (WFCAM; Casali et al. (2007)); the photometric system is described in Hewett et al. (2006), the calibration in Hodgkin et al. (2009), while the pipeline processing and science archive are described in Hambly et al. (2008).

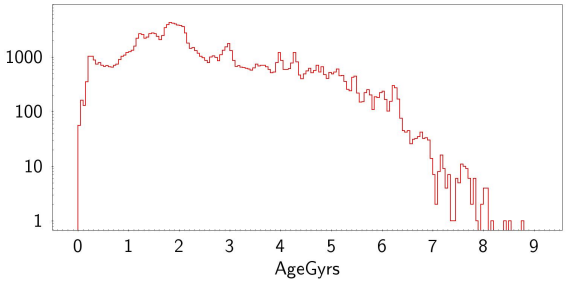
Thus, the final Maraston/UKIDSS (MARUK) catalogue was formed by 134,837 galaxies of known:

- age, E(B-V), stellar mass and metallicity;
- model, cmodel, petrosian and fiber magnitude in the ugriz SDSS bands;
- Y, J, H, K UKIDSS *apermags* (the UKIDSS default point/extended source aperture-corrected magnitude);

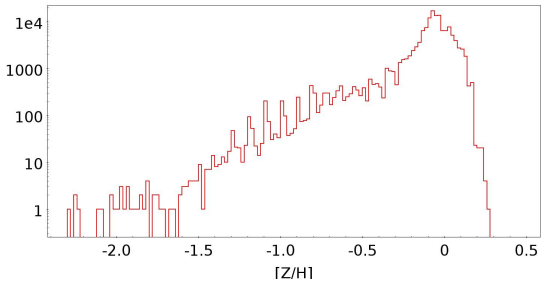
Figures 3.2 and 3.3 show the redshift, age, metallicity and magnitude distributions for the objects in the cleaned MARUK dataset.

3.2.2 Mock catalogue

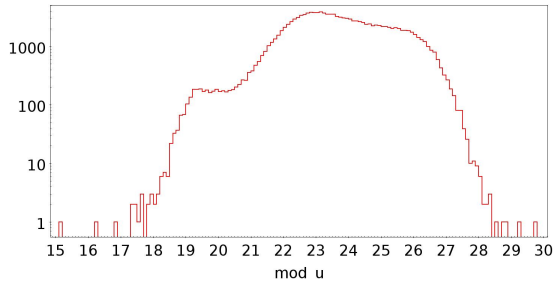
To create the mock catalogue, we used the standard parameters of the BC03 model, i.e. a Chabrier IMF and the Padova 1994 evolutionary tracks (Bruzual and Charlot, 2003); details on the BC03 code and the catalogues created are reported in Appendix B.
At redshift z=0.00, of the 221 default age values we eliminated the smaller ones (Age< 0.1 Gyr), which left us with the 106 age values reported in table B.1; for each of these values, we derived 6 SSP models, one for each of the six available metallicity classes, thus obtaining a catalogue of 636 SSP models of known rest-frame absolute magnitudes in the 13 ugrizUBVRIJHK bands; from this catalogue, we randomly extracted two subsets to be used for training and testing the network respectively, equally containing half of the SSP model catalogue (318 objects).
As described in section 3.3.2, we determined the metallicity sensitivity of all the 66 colors available in the mock catalogue by following Li et al. (2007) and Li and Han (2008); the distributions of the 10 most metallicity sensitive colors are shown in figures 3.4 and 3.5, while the distributions of the least sensitive ones is shown in figures 3.6 and 3.7. Figure 3.8, finally, shows the distribution of the SDSS colors, used in one of the experiments, in the mock catalogue.



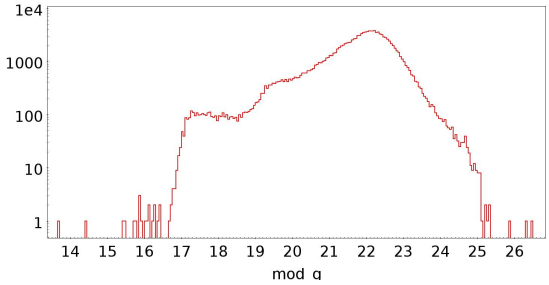
(a) Age (Gyr)



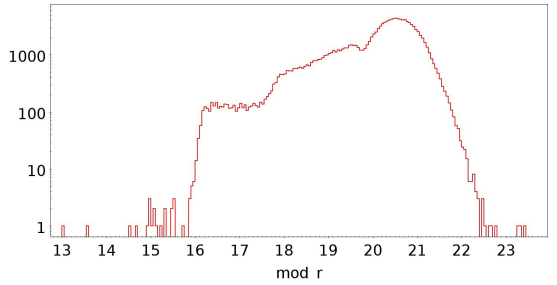
(b) Metallicity (Z/H)



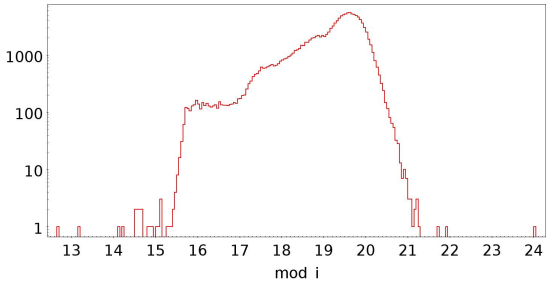
(c) model magnitude, u band distribution.



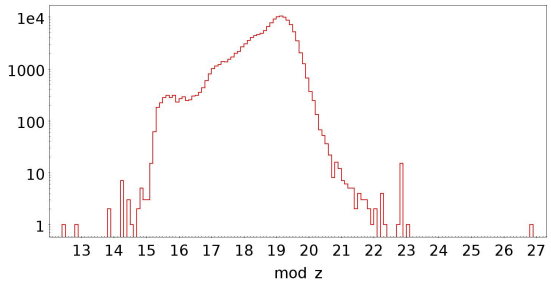
(d) model magnitude, g band distribution.



(e) model magnitude, r band distribution.

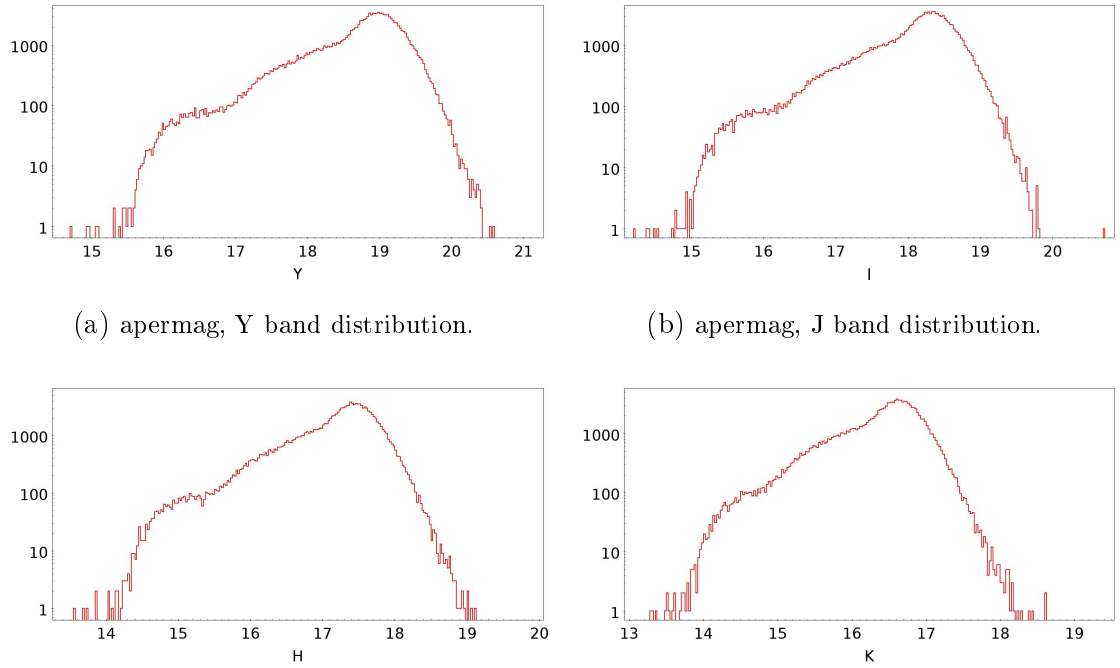


(f) model magnitude, i band distribution.



(g) model magnitude, z band distribution.

Figure 3.2: MARUK catalogue, distribution of: age (a), metallicity (b), SDSS model magnitude in the ugriz bands (panels c to g)



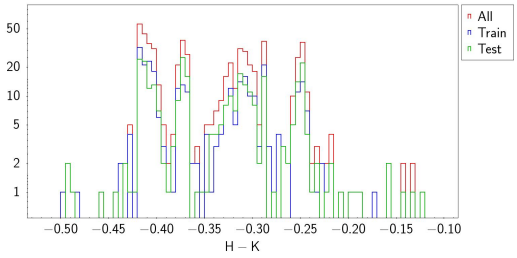
(a) apermag, Y band distribution.

(b) apermag, J band distribution.

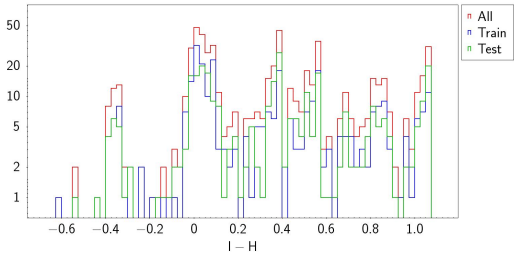
(c) apermag, H band distribution.

(d) apermag, K band distribution.

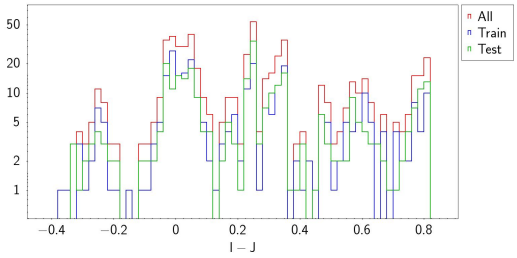
Figure 3.3: MARUK catalog, distribution of UKIDSS apermags in the YJHK bands (from panel a to d)



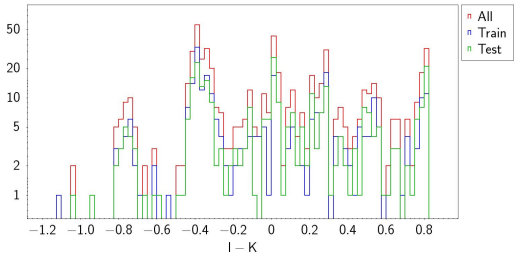
(a) H-K color distribution.



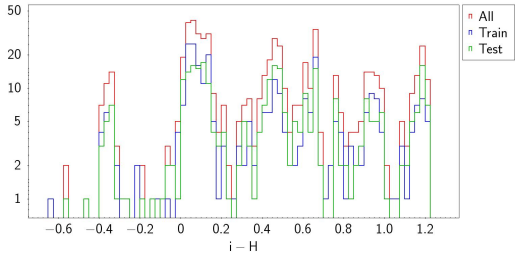
(b) I-H color distribution.



(c) I-J color distribution.

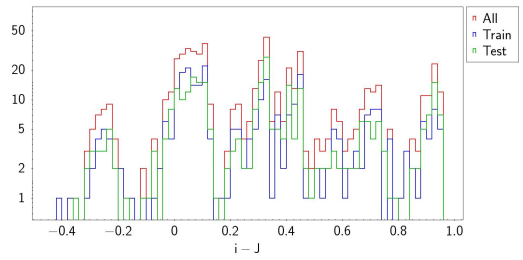


(d) I-K color distribution.

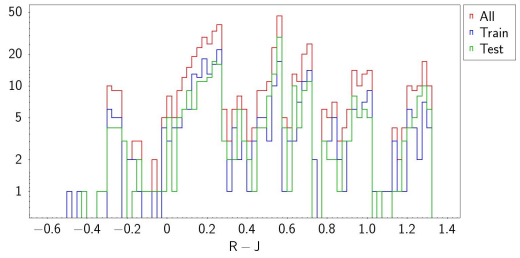


(e) i-H color distribution.

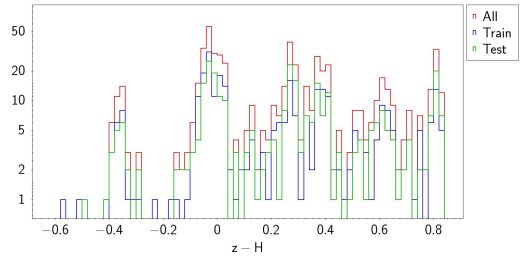
Figure 3.4: Distribution of the first 5 of the 10 more metallicity-sensitive colors in the mock catalogue at redshift z=0.00, as listed in table 3.2. From top left to bottom right: H-K (a), I-H (b), I-J (c), I-K (d), i-H (e)



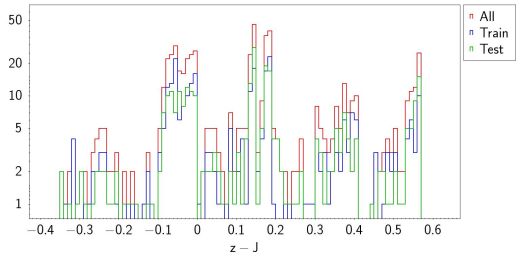
(a) i-J color distribution.



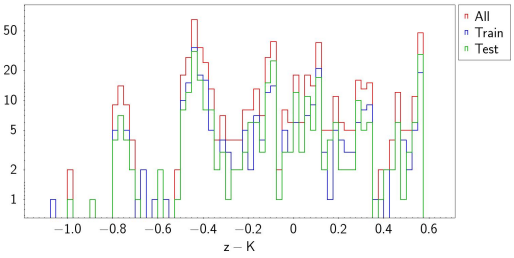
(b) R-J color distribution.



(c) z-H color distribution.

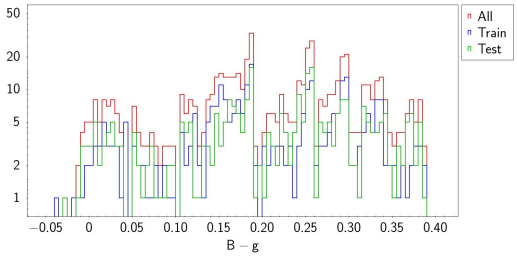


(d) z-J color distribution.

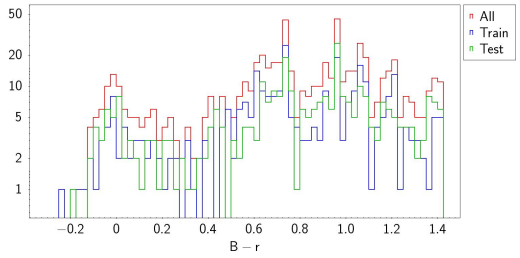


(e) z-K color distribution.

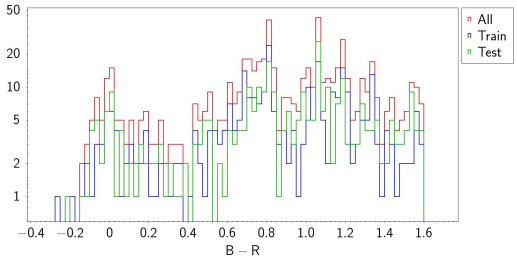
Figure 3.5: Distribution of the last 5 of the 10 more metallicity-sensitive colors in the mock catalogue at redshift z=0.00, as listed in table 3.2. From top left to bottom right: i-J (a), R-J (b), z-H (c), z-J (d), z-K (e)



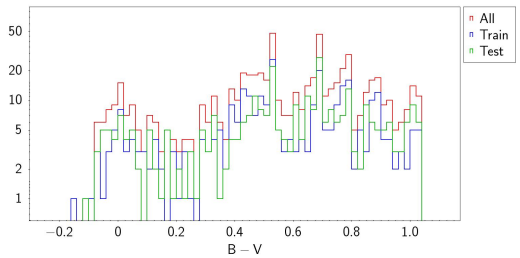
(a) B-g color distribution.



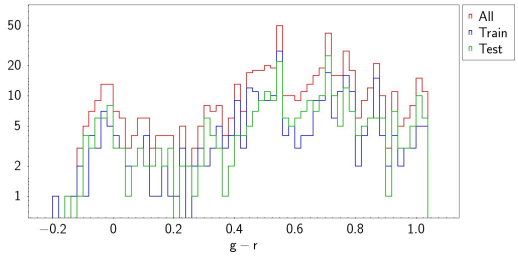
(b) B-r color distribution.



(c) B-R color distribution.

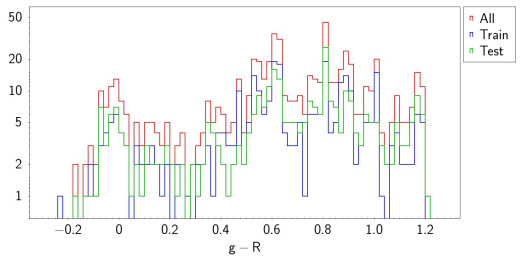


(d) B-V color distribution.

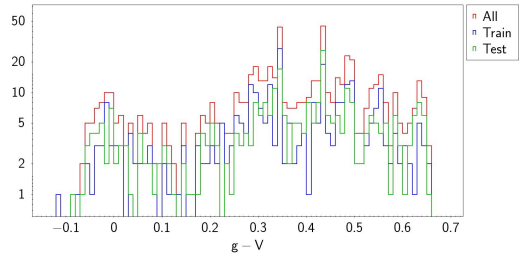


(e) g-r color distribution.

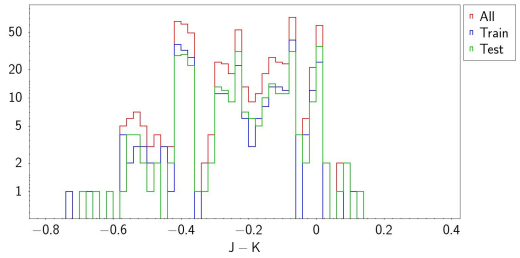
Figure 3.6: Distribution of the first 5 of the 10 least metallicity-sensitive colors in the mock catalogue at redshift z=0.00, as listed in table 3.2. From top left to bottom right: B-g (a), B-r (b), B-R (c), B-V (d), g-r (e)



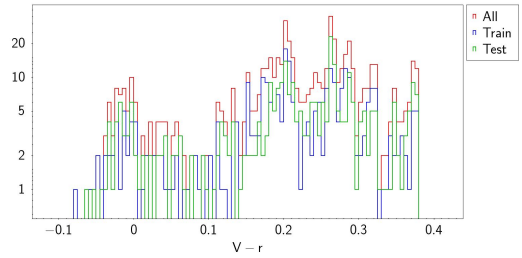
(a) g-R color distribution.



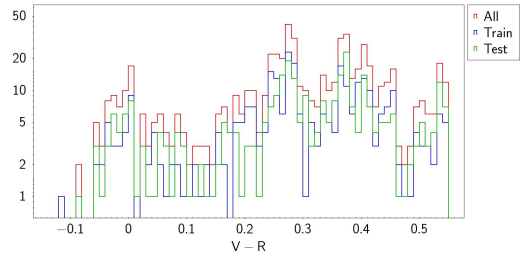
(b) g-V color distribution.



(c) J-K color distribution.

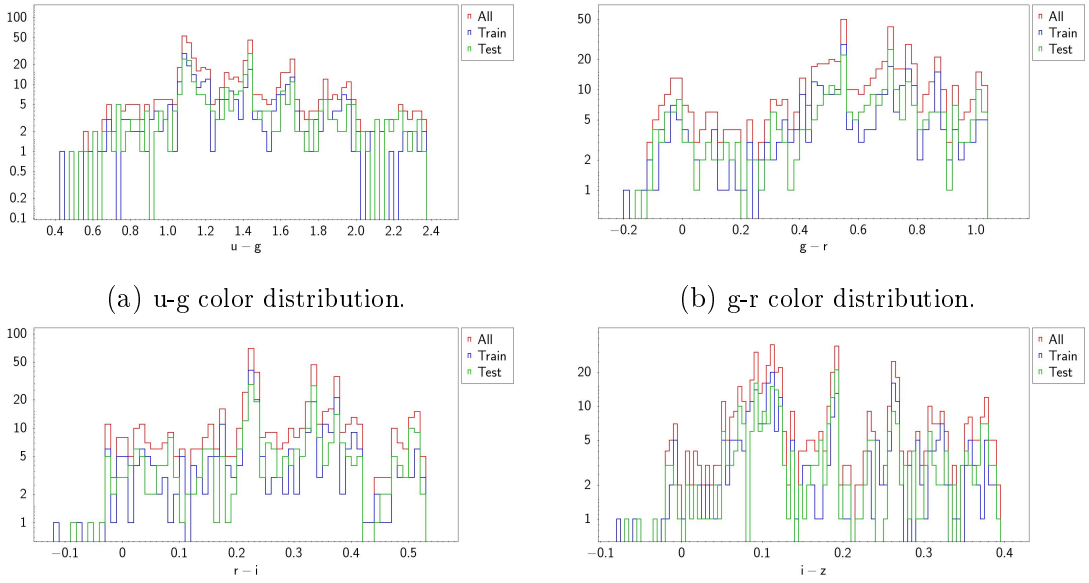


(d) V-r color distribution.



(e) V-R color distribution.

Figure 3.7: Distribution of the last 5 of the 10 least metallicity-sensitive colors in the mock catalogue at redshift z=0.00, as listed in table 3.2. From top left to bottom right: g-R (a), g-V (b), J-K (c), V-r (d), V-R (e)



(a) u-g color distribution.

(b) g-r color distribution.

(c) r-i color distribution.

(d) i-z color distribution.

Figure 3.8: Distribution of the four SDSS colors in the mock catalogue at redshift z=0.00. From top left to bottom right: u-g (a), g-r (b), r-i (c), i-z (d).

3.2.3 Gas-phase metallicity catalogue

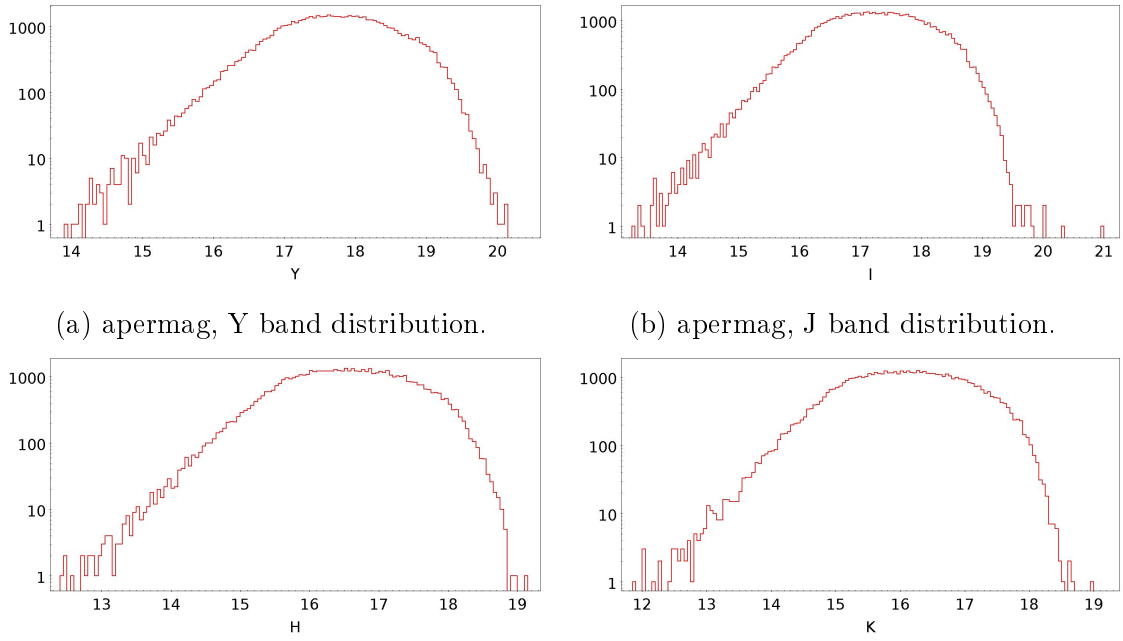
This catalogue contains the galaxy properties derived from the MPA-JHU emission line analysis for the SDSS DR 7 (Abazajian et al., 2009); the gas-phase metallicity of the galaxies, here expressed as the oxygen abundance estimate, is derived using the Charlot and Longhetti (2001) (hereafter CL01) models as discussed in T04.
The data are publicly available at the URL `http://www.mpa-garching.mpg.de/SDSS/DR7/`.
The original catalogue was formed by 927,552 galaxies, out of which, however, only 188,403 presented physical values of the oxygen abundance.
As for the stellar metallicity dataset, we derived the model, cmodel and petrosian optical (ugriz bands) magnitudes and the spectroscopic redshift from the SDSS, thus obtaining a dataset of 188,351 objects; the UKIDSS cross-match (with the usual 0.5 arcsec tolerance) yielded a dataset of 59,787 objects, eventually reduced to 46,598 after the elimination of missing or Not A Number (NaN) values. Hereafter, we refer to this catalogue as the **MPAUK** dataset.
We show the distribution of the physical properties of these galaxies in figures 3.9 and (3.10)

3.3 Stellar metallicity experiments

In this section, we described the experiments performed to train the network to provide *stellar* metallicity predictions; we reported the result obtained on the MARUK objects, on the simulated catalogue and presented the results of the comparison between the two datasets.

3.3.1 Optical colors preliminary experiments.

From the complete MARUK dataset, we extracted the objects in a redshift bin centered around z=0.1 ($0.08 < z < 0.12$), thus obtaining a subset (MARUK_z010) of 1025 galaxies.
Finally, we further separated this dataset in two randomly extracted subsets, one



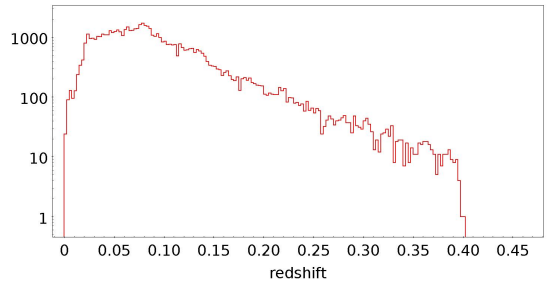
(a) apermag, Y band distribution.

(b) apermag, J band distribution.

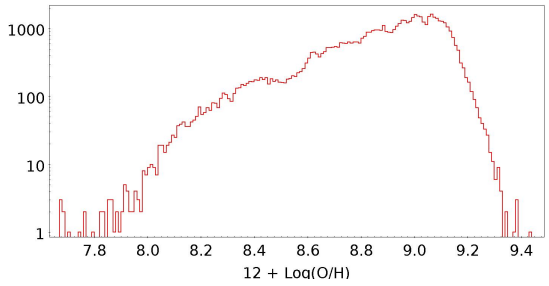
(c) apermag, H band distribution.

(d) apermag, K band distribution.

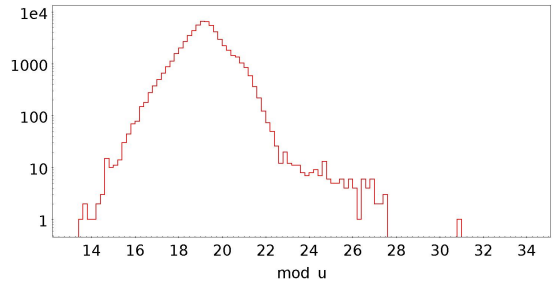
Figure 3.9: MPAUK catalog, distribution of UKIDSS apermag in the YJHK bands (from panel a to d).



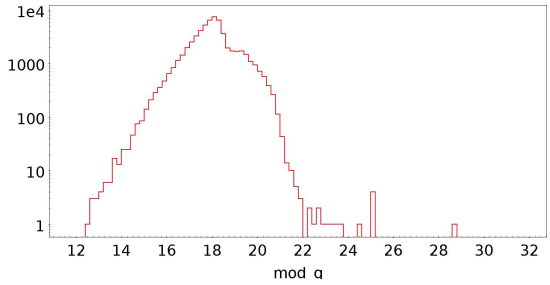
(a) redshift distribution.



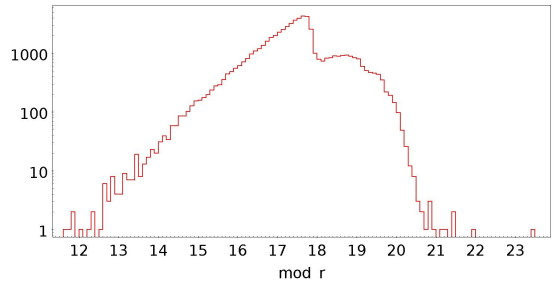
(b) metallicity (12+Log(O/H)) distribution.



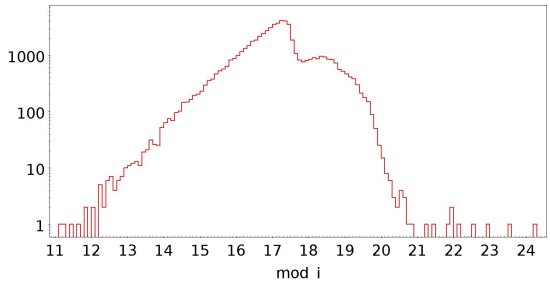
(c) model magnitude, u band distribution.



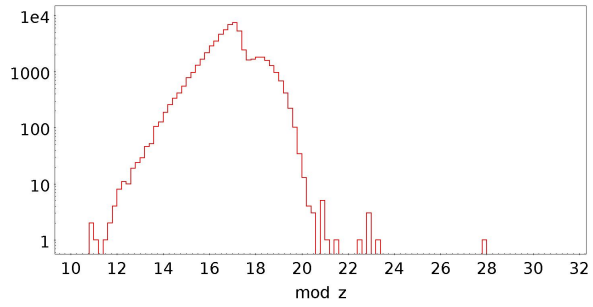
(d) model magnitude, g band distribution.



(e) model magnitude, r band distribution.



(f) model magnitude, i band distribution.



(g) model magnitude, z band distribution.

Figure 3.10: MPAUK catalog, distribution of: redshift (a), metallicity (b), SDSS model magnitude in the ugriz bands (panels c to g)

for training the neural network (710 galaxies) and one (315 galaxies) to test its performances after the training phase, and we performed a first series of regression experiments on the (MARUK_z010) dataset, using only optical colors as input features, and the SC#001 (see A.1).
The input features and the results (summarized by the indicators defined in section 2.2) for each particular experiment are listed in table 3.1.
Figure 3.11 shows the scatter plots (i.e., calculated Vs expected metallicity) of the optical colors experiments, with the blue line identifying the ideal behavior.

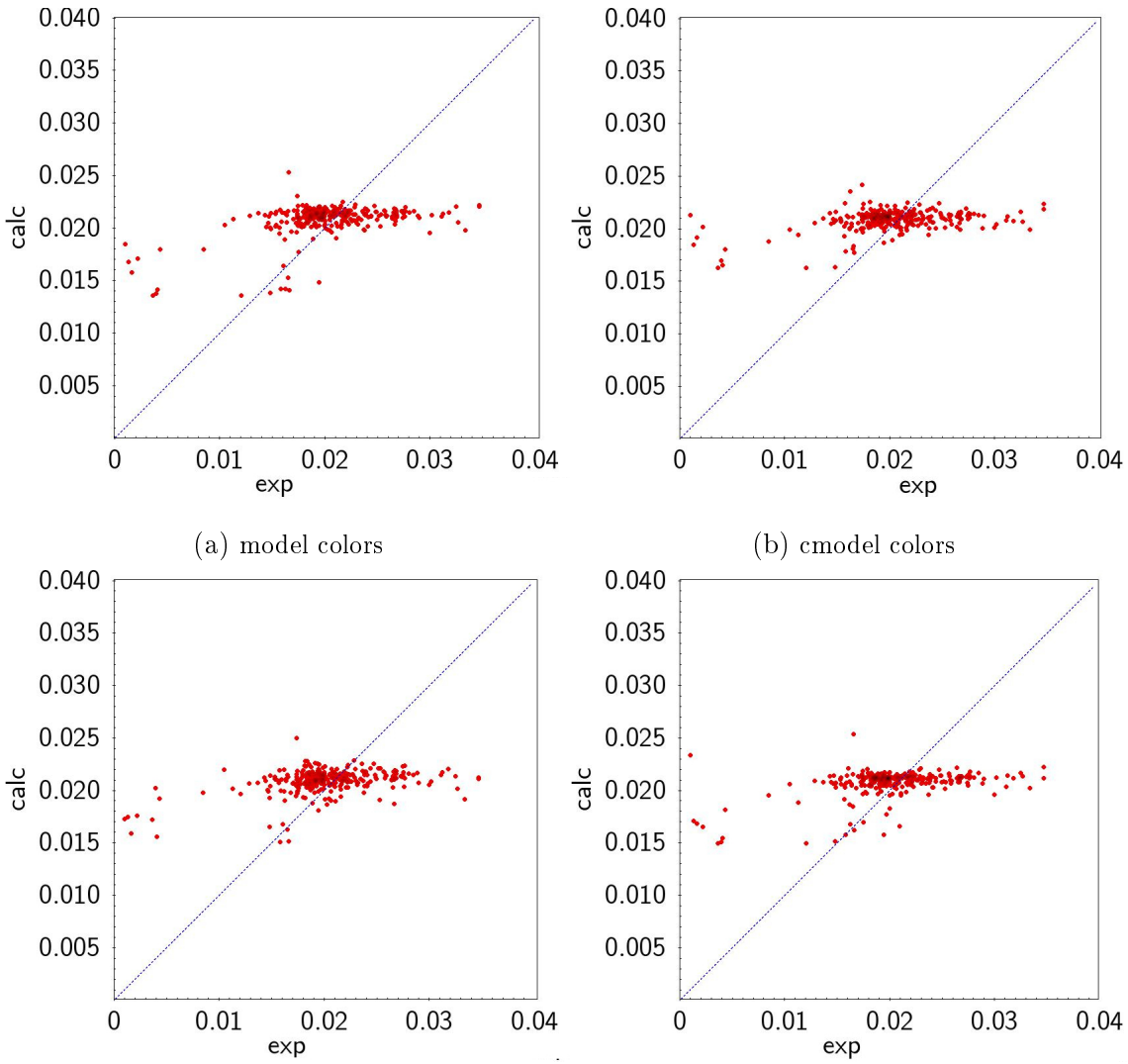
Features	|bias|	σ
model colors (ugriz bands)	0.00052	0.00433
cmodel colors (ugriz bands)	0.00056	0.00456
fiber colors (ugriz bands)	0.00060	0.00448
petrosian colors (ugriz bands)	0.00047	0.00446

Table 3.1: MARUK_z10 catalog, optical colors as input features - results of the regression experiments.

In spite of the good results in terms of the statistical indicators, it appears evident from the scatter plots that using optical colors alone as input features is not enough to produce reliable stellar metallicity estimates; however, we could not tell whether it was a result of the age-metallicity degeneration (or of some other unforeseen effect), or if it was due to an intrinsic inability of the method employed to recreate the relationship between the photometric properties and metallicity.

3.3.2 Mock catalogue regressions

To understand the results obtained on the MARUK_z10 catalog, we decided to test the neural network on a mock catalogue created with the BC03 EPS code, after determining the metallicity sensitivity of all the available colors in order to choose the best photometric features to use.
We now present the method used to determine the metallicity sensitivity, and the results of the experiments on the simulated data.



(a) model colors

(b) cmodel colors

(c) fiber colors

(d) petrosian colors

Figure 3.11: Scatter plots showing calculated Vs expected metallicity (the blue dotted line represents the ideal expected behavior), of the regression experiments performed on the MARUK_z10 catalog, using optical colors obtained from the model (a), cmodel (b), fiber (c) and petrosian (d) SDSS magnitudes respectively as input features.

Relative metallicity sensitivities

Following Worthey (1994), Li et al. (2007) and Li and Han (2008) we investigated the **Relative Metallicity Sensitivity (RMS)** of the 66 colors derived from the available magnitudes to determine their capability of breaking the AMD.
The RMS is defined as

$$RMS \equiv \frac{\frac{\Delta I_Z}{\left(\frac{\Delta Z}{Z_0}\right)}}{\frac{\Delta I_T}{\left(\frac{\Delta T}{T_0}\right)}} \tag{3.1}$$

where

- ΔI_Z is the variation of a color due to a pure change of metallicity ΔZ with respect to the standard metallicity value Z_0
- ΔI_T is the variation of a color due to a pure change of age ΔZ with respect to the standard age value T_0

According to its definition, we expected colors with large RMS (> 1.0) to be more metallicity-sensitive, and those with small RMS (< 1.0) to be good age indicators.

We set the zero point for the age and metallicity values:

- $T_0 = 8$ Gyr;
- $Z_0 = 0.02$ (Solar);

and considered the color changes caused by two different age (ΔT_1 and ΔT_2) and metallicity (ΔZ_1 and ΔZ_2) variations

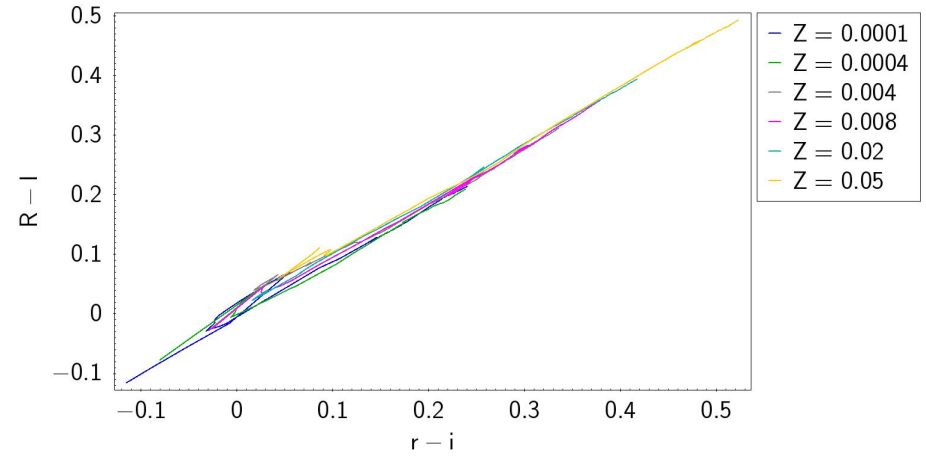
- $T_1 = 3$ Gyr $\Rightarrow \Delta T_1 = T_0 - T_1 = 5$ Gyr;
- $T_2 = 12$ Gyr $\Rightarrow \Delta T_2 = T_2 - T_0 = 4$ Gyr;
- $Z_1 = 0.008$ dex $\Rightarrow \Delta Z_1 = Z_0 - Z_1 = 0.012$ dex;
- $Z_2 = 0.05$ dex $\Rightarrow \Delta T_2 = Z_2 - Z_0 = 0.03$ dex.

OrderID	Color	RMS	OrderID	Color	RMS	OrderID	Color	RMS
1	H-K	4.228	23	I-z	1.423	45	u-g	1.030
2	z-j	2.812	24	B-J	1.419	46	u-I	1.029
3	I-J	2.374	25	i-z	1.398	47	J-H	1.025
4	z-H	2.285	26	B-H	1.385	48	g-I	1.019
5	i-J	2.262	27	i-I	1.384	49	u-i	1.016
6	z-K	2.189	28	B-K	1.382	50	u-V	1.011
7	I-H	2.072	29	u-J	1.290	51	u-R	1.007
8	I-K	2.012	30	u-K	1.277	52	V-i	1.006
9	i-H	2.010	31	u-H	1.275	53	u-r	1.003
10	R-J	1.977	32	R-z	1.241	54	B-I	0.993
11	i-K	1.958	33	r-z	1.201	55	g-i	0.989
12	r-J	1.878	34	R-I	1.167	56	B-i	0.967
13	R-H	1.821	35	r-I	1.131	57	g-R	0.959
14	R-K	1.786	36	V-z	1.121	58	V-R	0.957
15	r-H	1.750	37	R-i	1.089	59	g-V	0.951
16	r-K	1.721	38	r-i	1.070	60	g-r	0.944
17	V-J	1.696	39	u-B	1.069	61	B-R	0.938
18	V-H	1.612	40	g-z	1.067	62	B-r	0.924
19	V-K	1.592	41	V-I	1.053	63	V-r	0.924
20	g-J	1.505	42	u-z	1.051	64	J-K	0.921
21	g-H	1.458	43	B-z	1.035	65	B-V	0.919
22	g-k	1.450	44	r-R	1.031	66	B-g	0.863

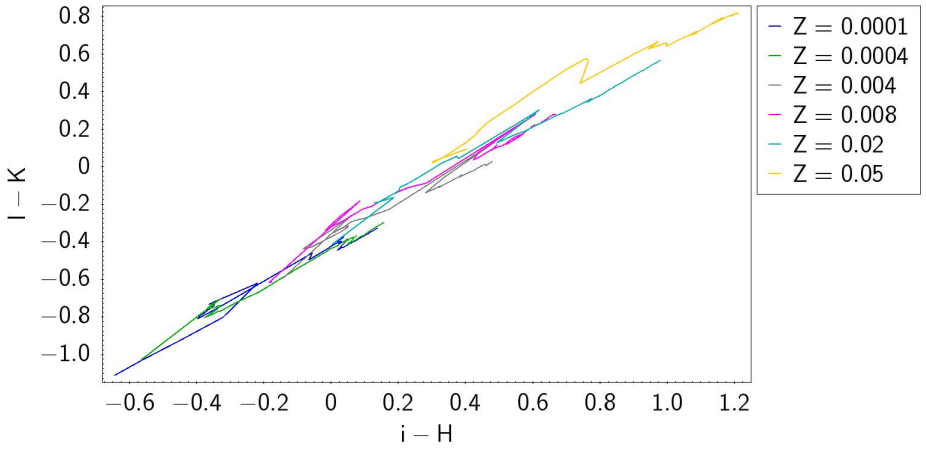
Table 3.2: The 66 available colors in the mock catalogue at z=0.00, ordered by their RMS.

Therefore, for each color we obtained two values of the RMS, and used their average as the final result. Table 3.2 shows the RMS obtained for the 66 available colors.

To evaluate the ability of these color pairs to break the AMD, we can analyze some color-color diagrams.
As shown in figure 3.12, using color pairs with similar rms values is less useful, while if we consider couples of colors with very different metallicity (fig 3.13) sensitivities we are able to separate the various metallicity subsets.

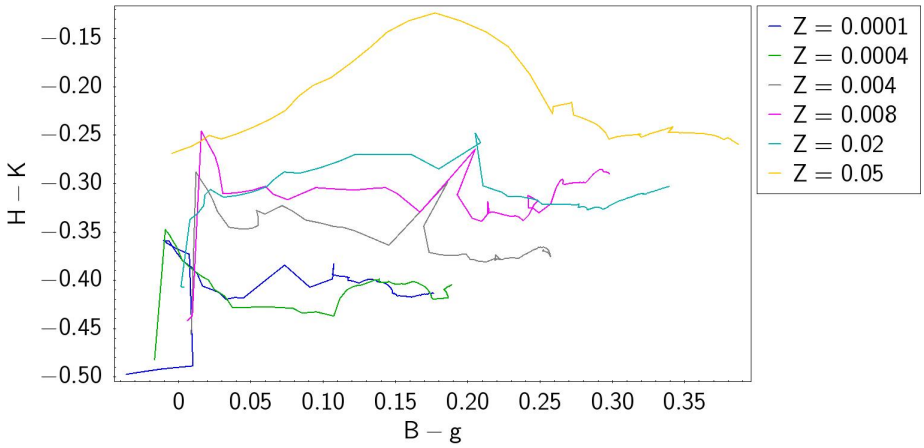


(a)

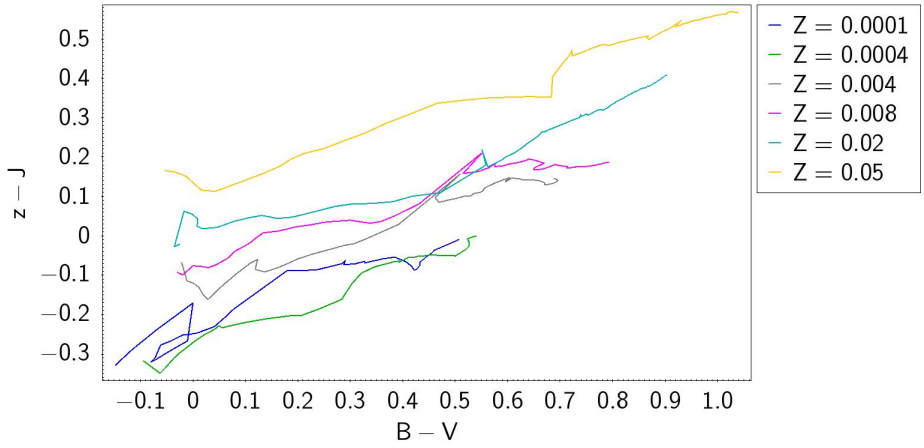


(b)

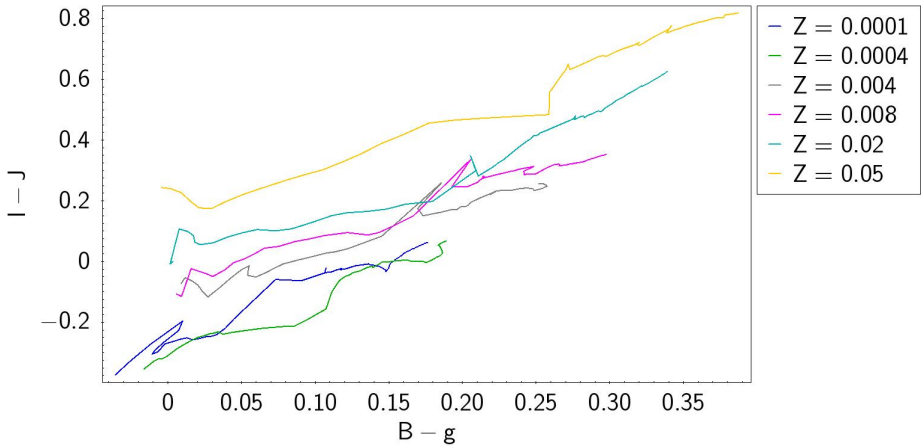
Figure 3.12: Color-color diagrams displaying R-I Vs r-i (a) and I-K Vs i-H (b); these color pairs have similar RMS, so they are not able to break the AMD.



(a)



(b)



(c)

Figure 3.13: Color-color diagrams displaying H-K Vs B-g (a), and z-J Vs B-V (b) and I-J Vs B-g (c); color pairs with very different values of the RMS can be used to break the AMD.

Rest-frame - High and low-RMS colors

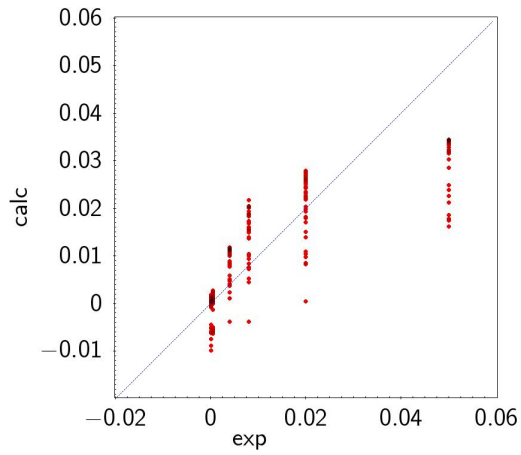
Having a tool to evaluate the sensitivity of the available colors to metallicity, we proceeded to perform a series of tests to verify this correlation and to find the best configuration of the neural network.
To this purpose, we performed four regression experiments:

- two experiments using the 10 colors with the highest values of the RMS (see table 3.2), and two Setup Criteria (SC) reported in table A.1;
- for comparison, we performed two additional regression experiments by using the 10 lowest-RMS colors reported in table 3.2, and both the Setup Criteria used in the previous test.

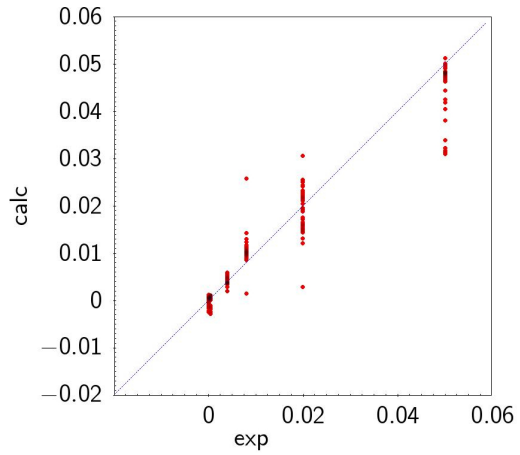
As already mentioned when we presented the catalogue (section 3.2.2), the distributions of the colors used in this phase are shown in figures 3.4 and 3.5 for the high-RMS colors, while figures 3.6 and 3.7 report the distribution of the low-RMS ones.
The scatter plots for the high-RMS features (figure 3.14, first row) readily show a definite improvement with respect to the experiments performed on the real galaxies catalogue; in particular, by considering the best experiment (high-RMS colors and Setup Criteria #001), the σ value (see table 3.3) is smaller than the metallicity difference between all but the first two classes, which allows the network to correctly separate the objects in the right class.
For comparison, in table 3.3 we also showed the results obtained with the least metallicity-sensitive colors and the two sets of Setup Criteria, and we showed the scatter plots in the second row of figure 3.14; it appeared evident that using low-RMS colors yields a worse result with respect to the high-RMS ones, even if not as bad as using the lower decay value of 0.01.

Rest-frame - High-RMS colors + R,K magnitudes

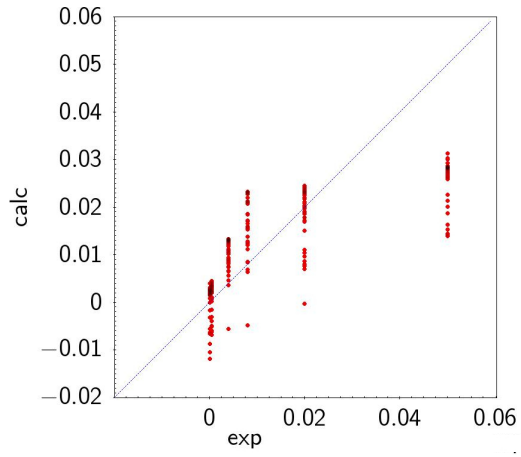
We performed another experiment, by adding the absolute magnitudes in the R and K bands, and using the Setup Criteria #001 which proved to yield the best result in the previous test.
Table 3.3 reports the results of this test, showing that the addition of the magnitudes does not improve the performance of the network. The corresponding scatter plot is shown in figure 3.15.



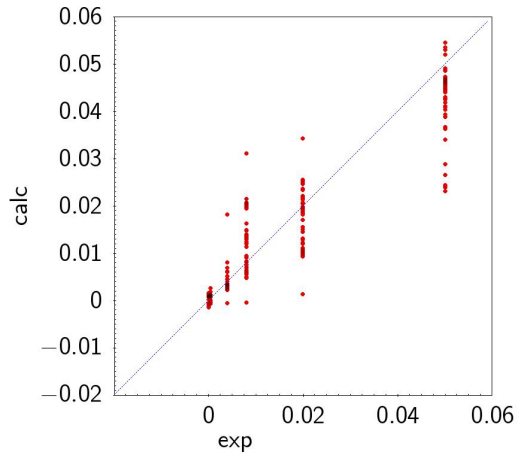
(a) 10 highest-RMS colors, SC#01.



(b) 10 highest-RMS colors, SC#001.



(c) 10 lowest-RMS colors, SC#01.



(d) 10 lowest-RMS colors, SC#001.

Figure 3.14: Scatter plots showing calculated Vs expected metallicity (the blue dotted line represents the ideal expected behavior), of the regression experiments performed on the mock catalogue at z=0.00 using the 10 highest-RMS colors (top row) as input features with Setup Criteria #01 (a) and #001 (b), and 10 lowest-RMS colors (bottom row) with Setup Criteria #01 (c) and #001 (d).

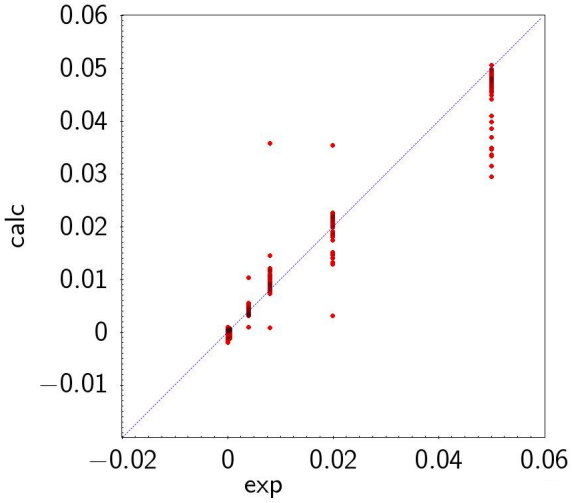


Figure 3.15: Scatter plot showing calculated Vs expected metallicity (the blue dotted line represents the ideal expected behavior) of the regression experiments performed on the mock catalogue at z=0.00 using the 10 highest-RMS colors and the R, K bands absolute magnitudes as input features and the Setup Criteria #001.

Rest-frame - SDSS colors

For a final comparison, we used only the 4 main SDSS colors as input features in this experiment.
The scatter plot of the SDSS experiment is shown in figure 3.16, and the results are listed in table 3.3. As expected, the result obtained using only SDSS colors is worse than the one resulting from using colors in the combined bands (as a matter of fact, it is even worse than the result given by the 10 low-RMS colors).

Comparison at different redshift values

Following the previous results, we performed similar analysis for 6 different values of the redshift.

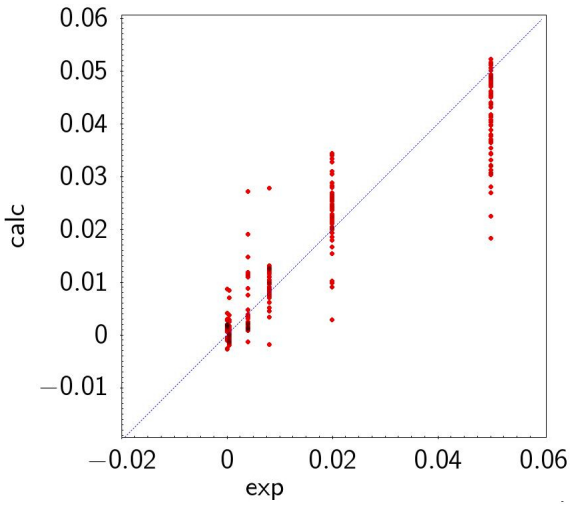


Figure 3.16: Scatter plot showing calculated Vs expected metallicity (the blue dotted line represents the ideal expected behavior) of the regression experiments performed on the mock catalogue at z=0.00 using the 4 SDSS colors as input features and the Setup Criteria #001.

Features	**Setup Criteria**	**|bias|**	σ
10 High-rms colors	#01	0.0014	0.0102
10 High-rms colors	#001	0.0006	0.0037
10 Low-RMS colors	#01	0.0016	0.0123
10 Low-RMS colors	#001	0.0010	0.0062
10 High-rms colors + R,K mags	#001	0.0007	0.0038
SDSS colors	#001	0.0008	0.0067

Table 3.3: Input features, Setup Criteria and results of the regression experiments performed on the mock catalogue at redshift z=0.00.

For each value, we used the BC03 code to create a different catalogue (see Appendix B for the details on the dataset creation), from which we extracted a training and test subset (see table 3.4).

Then, similarly to what we described in the previous section, for each catalogue

z	Galaxies in the catalogue	Galaxies in the training set	Galaxies in the test set
0.10	498	298	200
0.20	426	273	153
0.30	402	222	180
0.40	378	221	157
0.50	360	221	139
0.60	342	201	141

Table 3.4: Number of galaxies in the mock catalogue and the corresponding training and test subsets for each redshift value.

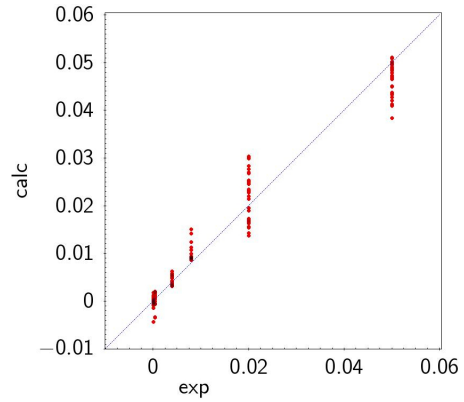
we calculated the RMS of the 66 available colors (again, details are reported in Appendix B), and we carried out two regression experiments, using the 10 highest and the 10 lowest-RMS colors as input features.

Figures 3.17 and 3.18 show the scatter plots of the regressions for each redshift value, and we reported the corresponding statistical indicators in table 3.5; the results, in line with the ones obtained for z=0.00, show the success of the regression experiments, with the network being able to predict the expected metallicity with good accuracy and to separate correctly objects belonging to different classes.

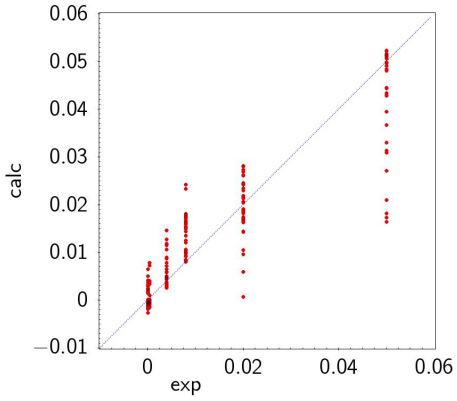
3.3.3 Mock catalogue classifications

In the previous section, we described the regression experiments performed on the mock catalogue, using different photometric sets of input features and various redshift values, and we observed the improvements of these tests in providing reliable metallicity estimates with respect to the preliminary experiments on real data.
However, as it is also readily evident from an inspection of the scatter plots, the objects in the simulated dataset can actually only assume six different *discrete* metallicity values, i.e. each BC03 model belongs to one of the following classes:

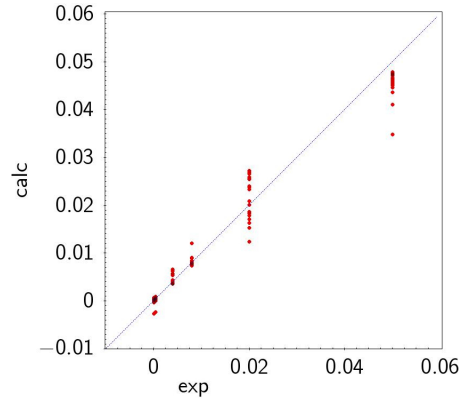
- m22: Z=0.0001;



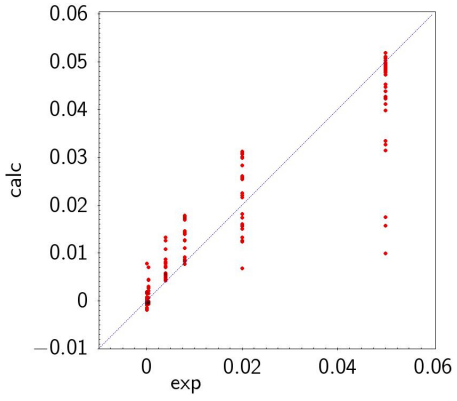
(a) z=0.10, 10 highest-RMS colors



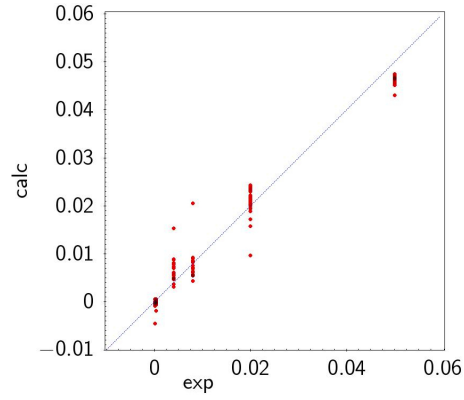
(b) z=0.10, 10 lowest-RMS colors



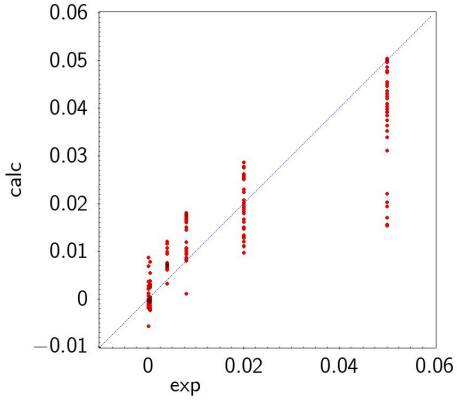
(c) z=0.20, 10 highest-RMS colors



(d) z=0.20, 10 lowest-RMS colors

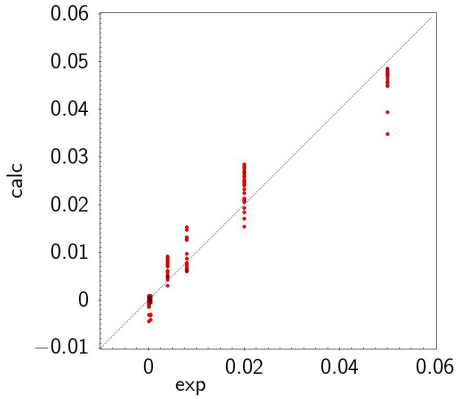


(e) z=0.30, 10 highest-RMS colors

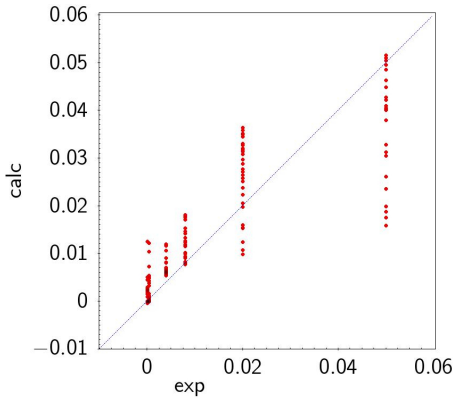


(f) z=0.30, 10 lowest-RMS colors

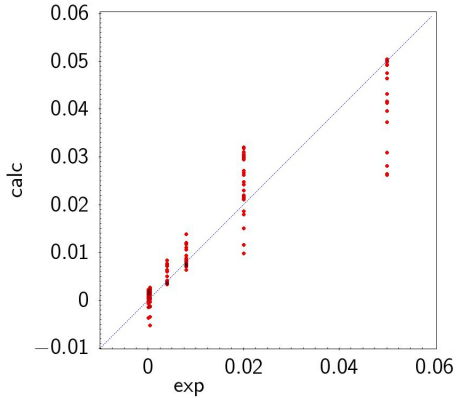
Figure 3.17: Scatter plots showing calculated Vs expected metallicity (the blue dotted line represents the ideal expected behavior) of the regression experiments performed on the mock catalogue for different redshift values. The redshift value increases from top to bottom row (z=0.10, 0.20, 0.30); experiments using the 10 highest-RMS colors as input features are shown on the left size, the ones with the 10 lowest are shown on the right. The adopted Setup Criteria are always the #001.



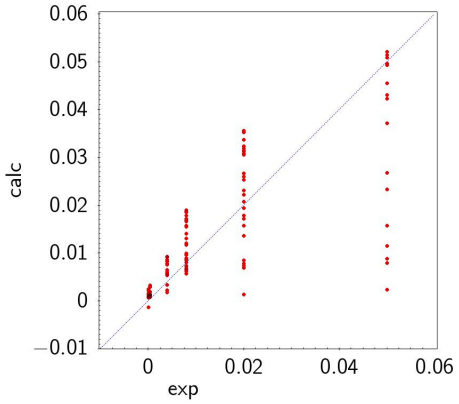
(a) z=0.40, 10 highest-RMS colors



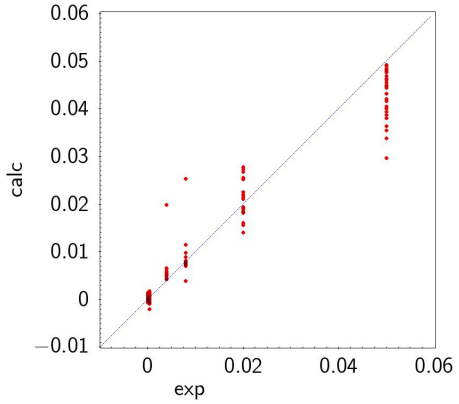
(b) z=0.40, 10 lowest-RMS colors



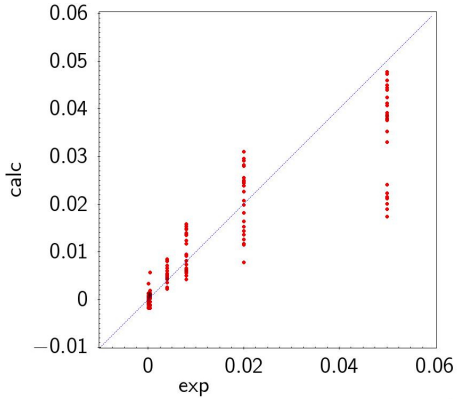
(c) z=0.50, 10 highest-RMS colors



(d) z=0.50, 10 lowest-RMS colors



(e) z=0.60, 10 highest-RMS colors



(f) z=0.60, 10 lowest-RMS colors

Figure 3.18: Scatter plots showing calculated Vs expected metallicity (the blue dotted line represents the ideal expected behavior) of the regression experiments performed on the mock catalogue for different redshift values. The redshift value increases from top to bottom row (z=0.40, 0.50, 0.60); experiments using the 10 highest-RMS colors as input features are shown on the left size, the ones with the 10 lowest are shown on the right. The adopted Setup Criteria are always the #001.

Redshift	**Features**	**Setup Criteria**	**|bias|**	σ
z=0.00	10 High-rms colors	01	0.00142	0.01020
	10 High-rms colors	001	0.00063	**0.00370**
	10 Low-rms colors	01	0.00156	0.01226
	10 Low-rms colors	001	0.00962	0.00617
	10 High-rms colors + R,K mags	001	0.00066	0.00378
	SDSS colors	001	0.00079	0.00665
z=0.10	10 High-rms colors	001	0.00007	**0.00309**
	10 Low-rms colors	001	0.00009	0.00755
z=0.20	10 High-rms colors	001	0.00047	**0.00283**
	10 Low-rms colors	001	0.00036	0.00711
z=0.30	10 High-rms colors	001	0.00059	**0.00269**
	10 Low-rms colors	001	0.00085	0.00793
z=0.40	10 High-rms colors	001	0.00041	**0.00346**
	10 Low-rms colors	001	0.00080	0.00849
z=0.50	10 High-rms colors	001	0.00009	**0.00551**
	10 Low-rms colors	001	0.00051	0.01058
z=0.60	10 High-rms colors	001	0.00100	**0.00451**
	10 Low-rms colors	001	0.00220	0.00800

Table 3.5: Input features, Setup Criteria and results of the regression experiments performed on the mock catalogue for the different redshift values.

- m32: Z=0.0004;
- m42: Z=0.004;
- m52: Z=0.008;
- m62: Z=0.02;
- m72: Z=0.05.

For this reason, we decided it could be useful to carry on a further investigation of the results; thus, by using the input features which provided the most accurate predictions in each regression case, we performed a series of classification experiments on the mock catalogue, in order to estimate the precision with which the network could separate objects belonging to the different classes.

Rest-frame

The better result at redshift z=0.00 was obtained when using the 10 most metallicity-sensitive colors as input features of the network; using the same configuration for classification, we obtained as result the confusion matrix shown in table 3.6, while table 3.7 reports the values of the statistical indicators.

Comparison at different redshift

We report in table 3.8 the confusion matrices of the classifications performed on the mock catalogue at the various redshift values, always using the 10 highest-RMS colors as input features, while the statistics for the experiments are summarized in table 3.9.

3.3.4 Comparison with the MARUK Catalog

Using the results obtained from the mock-catalog, we fell back onto the MARUK_z10 dataset and performed a new series of regression experiments, using the colors with

		Classified as					
		m22	m32	m42	m52	m62	m72
Target	m22	48	2	0	0	0	0
	m32	0	46	0	0	0	0
	m42	2	0	56	0	1	0
	m52	0	0	0	49	2	0
	m62	0	0	0	0	51	0
	m72	0	0	0	0	0	61

Table 3.6: Confusion matrix of the classification experiment for the six BC03 metallicity classes performed on the mock catalogue using the 10 highest-RMS colors as input features at redshift z=0.00.

	cmp	**pc**	**cnt**	t_e
m22	96%	96 %	4%	97.8 %
m32	100 %	95.8%	4.2%	
m42	94.9 %	100 %	0 %	
m52	96.1 %	100 %	0 %	
m62	100 %	94.4 %	5.6 %	
m72	100 %	100 %	0 %	

Table 3.7: Statistical indicators to evaluate the performance of the classification experiment for the six BC03 metallicity classes carried out on the mock catalogue using the 10 highest-RMS colors as input features at redshift z=0.00.

		Classified as					
		m22	m32	m42	m52	m62	m72
Known	m22	38	1	0	0	0	0
	m32	0	26	0	0	0	0
	m42	0	0	33	0	0	0
	m52	0	0	1	33	1	0
	m62	0	0	0	1	33	0
	m72	0	0	0	0	0	33

(a) z=0.10

		Classified as					
		m22	m32	m42	m52	m62	m72
Known	m22	25	1	0	0	0	0
	m32	2	25	0	0	0	0
	m42	0	0	26	0	0	0
	m52	0	0	0	20	0	0
	m62	0	0	0	1	22	0
	m72	0	0	0	0	0	31

(b) z=0.20

		Classified as					
		m22	m32	m42	m52	m62	m72
Known	m22	38	1	0	0	0	0
	m32	0	26	0	0	0	0
	m42	0	0	33	0	0	0
	m52	0	0	1	33	1	0
	m62	0	0	0	1	33	0
	m72	0	0	0	0	0	33

(c) z=0.30

		Classified as					
		m22	m32	m42	m52	m62	m72
Known	m22	25	1	0	0	0	0
	m32	2	25	0	0	0	0
	m42	0	0	26	0	0	0
	m52	0	0	0	20	0	0
	m62	0	0	0	1	22	0
	m72	0	0	0	0	0	31

(d) z=0.40

		Classified as					
		m22	m32	m42	m52	m62	m72
Known	m22	38	1	0	0	0	0
	m32	0	26	0	0	0	0
	m42	0	0	33	0	0	0
	m52	0	0	1	33	1	0
	m62	0	0	0	1	33	0
	m72	0	0	0	0	0	33

(e) z=0.50

		Classified as					
		m22	m32	m42	m52	m62	m72
Known	m22	25	1	0	0	0	0
	m32	2	25	0	0	0	0
	m42	0	0	26	0	0	0
	m52	0	0	0	20	0	0
	m62	0	0	0	1	22	0
	m72	0	0	0	0	0	31

(f) z=0.60

Table 3.8: Confusion matrices of the classification experiments performed on the mock catalogue for each redshift value using the 10 highest-RMS colors.

	cmp	pc	cnt	t_e
m22	97.4 %	100 %	0 %	98.0 %
m32	100 %	96.3 %	3.7 %	
m42	100 %	97.1 %	2.9 %	
m52	94.3 %	97.1 %	2.9 %	
m62	97.1 %	97.1 %	2.9 %	
m72	100 %	100 %	0 %	

(a) z=0.10

	cmp	pc	cnt	t_e
m22	96.2 %	92.6 %	7.4 %	97.4 %
m32	92.6 %	96.2 %	3.8 %	
m42	100 %	100 %	0 %	
m52	100 %	95.2 %	4.8 %	
m62	95.7 %	100 %	0 %	
m72	100 %	100 %	0 %	

(b) z=0.20

	cmp	pc	cnt	t_e
m22	96.4 %	87.1 %	12.9 %	95.6 %
m32	100 %	97.0 %	3.0 %	
m42	89.3 %	96.2 %	3.8 %	
m52	88.9 %	96 %	4 %	
m62	96.4 %	96.4 %	3.6 %	
m72	100 %	100 %	0 %	

(c) z=0.30

	cmp	pc	cnt	t_e
m22	100 %	84.4 %	15.6 %	96.2 %
m32	82.1 %	100 %	0 %	
m42	100 %	100 %	0 %	
m52	95.8 %	100 %	0 %	
m62	100 %	96.6 %	3.4 %	
m72	100 %	100 %	0 %	

(d) z=0.40

	cmp	pc	cnt	t_e
m22	100 %	96 %	4 %	98.6 %
m32	100 %	95.7 %	4.3 %	
m42	95.5 %	100 %	0 %	
m52	100 %	100 %	0 %	
m62	100 %	100 %	0 %	
m72	94.1 %	100 %	0 %	

(e) z=0.50

	cmp	pc	cnt	t_e
m22	100 %	100 %	0 %	96.5 %
m32	100 %	100 %	0 %	
m42	95.2 %	95.2 %	4.8 %	
m52	91.7 %	95.7 %	4.3 %	
m62	95.5 %	87.5 %	12.5 %	
m72	96.3 %	100 %	0 %	

(f) z=0.60

Table 3.9: Statistical indicators used to evaluate the performance of the classification experiments carried out on the mock catalogue for each redshift value using the 10 highest-RMS colors.

the highest RMS as input features.

It should be pointed out, however, that in the MARUK catalogue the I and R Johnson magnitudes are missing, and so are some of the top colors listed in table B.3 (panel a); for this reason, we used the 10 highest colors out of the available ones (see table 3.10).

OrderID	**Color**	**RMS**	**OrderID**	**Color**	**RMS**
1	J-H	2.480	6	i-J	1.949
2	z-H	2.293	7	i-K	1.901
3	z-J	2.268	8	r-H	1.675
4	z-K	2.137	9	r-J	1.642
5	i-H	1.981	10	r-K	1.639

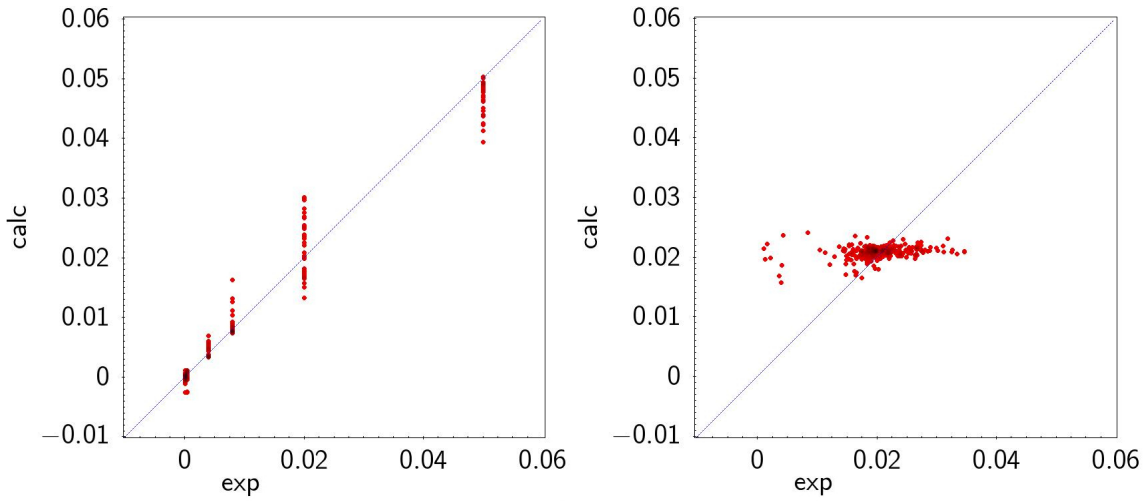
Table 3.10: Colors used as input features with the MARUK_z010 catalogue.

We verified that by eliminating the unavailable colors, the results of the regressions performed on the simulated data would not be affected significantly. As shown in the scatter plot (figure 3.19, panel a), the metallicity values calculated by the network are still consistent with the expected ones, and the statistical indicators (table 3.11, first row) are consistent with the ones obtained in the previous section with the full set of high-RMS colors; as a matter of fact, the σ has even a slightly smaller value (see table 3.5 for a comparison).

The results of the experiments on the MARUK_z010 dataset are compared to the ones obtained on the mock-catalogue in table 3.11, and the scatter plot in figure 3.19 (panel b).
Again, it is evident how even using the 10 high-RMS colors does not provides the expected result.

Dataset	**Features**	**|bias|**	σ
Mock-catalogue - z=0.10	10 High-RMS colors	0.00018	0.00289
MARUK_z010	10 high-rms colors	0.00052	0.00469

Table 3.11: Comparison of the results of a regression experiment at redshift z=0.10, using the 10 highest-RMS colors listed in table 3.10 as input features, performed on the mock-catalogue (first row) and on the MARUK_z010 dataset (second row).



(a) Mock-catalogue (b) MARUK_z010 catalogue

Figure 3.19: Scatter plots, showing calculated Vs expected metallicity (the blue dotted line represents the ideal expected behavior), of the regression experiments performed at redshift z=0.10, using the 10 colors listed in table 3.10, on the mock-catalogue (a) and the MARUK_z010 dataset (b). The adopted Setup Criteria are always the #001.

3.4 Gas-phase metallicity

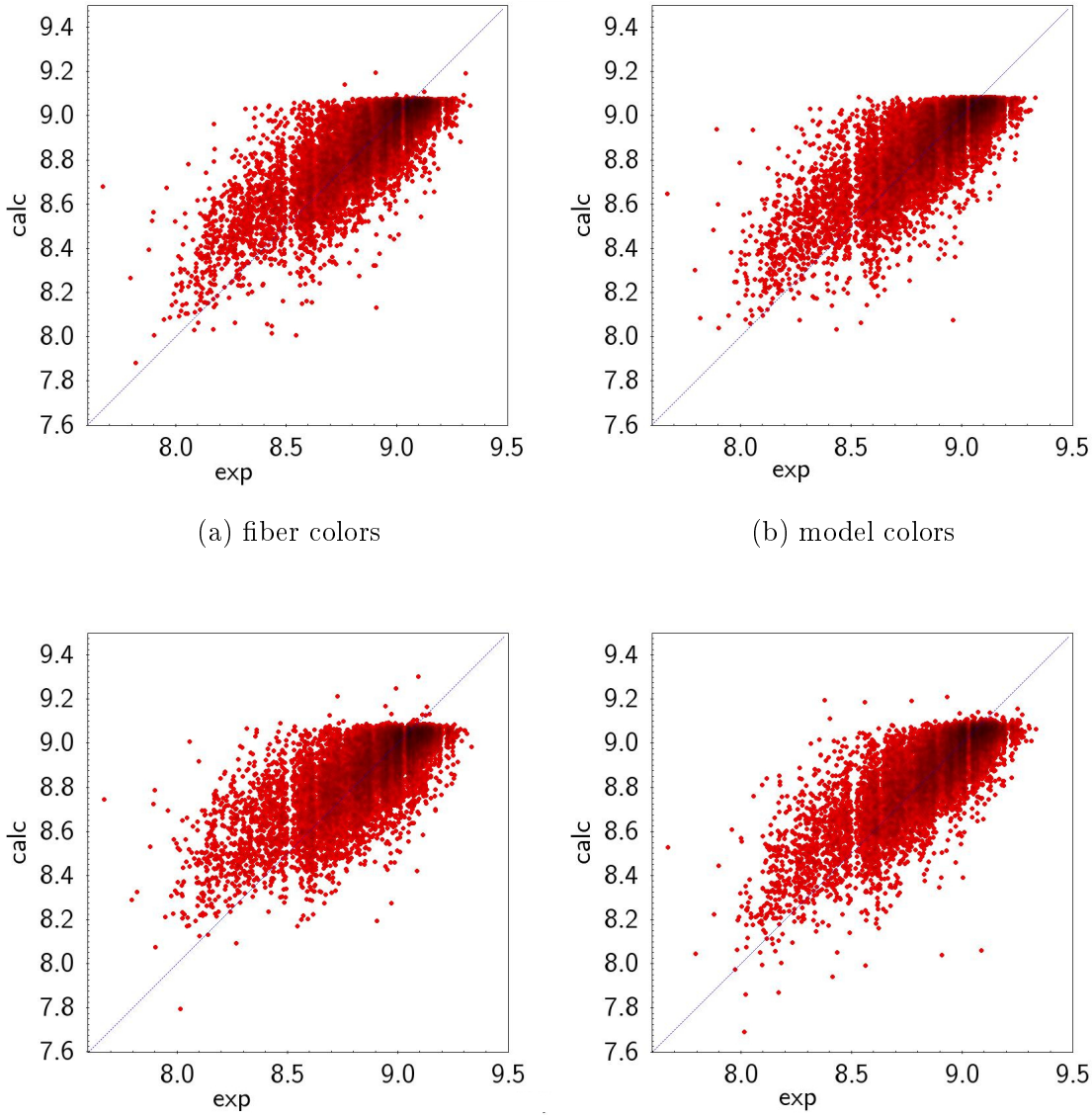
In this final section, we described the experiments performed to derive gas-phase metallicty estimates on the MPAUK photometric catalogue.

3.4.1 Optical+IR colors experiments

First, we performed some preliminary experiments on the whole MPAUK dataset, from which we randomly extracted two subsets for the training (32,587 galaxies) and the test phase (14,011 galaxies).
The magnitudes used as input features and the results for each experiment are listed in table 3.12, while figure 3.20 shows the scatter plots. Unlike with stellar metallicity, the values predicted by the network are consistent with the expected ones, even when using only SDSS optical colors as input features (the addition of IR colors does not improve significantly the outcome of the regression, at the cost of greatly increasing the time needed for training the neural network).
For each experiment, we noticed there were a few points that significantly deviate from the expected behavior, i.e. objects whose calculated metallicity value lied well outside the expected range spanned by the training set; so, for each regression we eliminated these outliers, and showed the scatter plots and the statistical indicators after the elimination of these points.

	Whole dataset		**Cleaned by outliers**	
Features	**|bias|**	σ	**|bias|**	σ
model colors (ugriz bands)	0.0013	0.1335	0.0016	0.1320
fiber colors (ugriz bands)	0.0003	0.1483	0.0010	0.1284
petrosian colors (ugriz bands)	0.0007	0.1730	0.0018	0.1443
fiber (ugriz bands) + IR colors	0.0004	0.1440	0.0018	0.1252

Table 3.12: Input features and results of the regression experiments performed on the MPAUK catalogue. We used the Setup Criteria #001 for all the experiments.



(a) fiber colors

(b) model colors

(c) petrosian colors

(d) fiber + IR colors

Figure 3.20: Scatter plots showing calculated Vs expected metallicity (the blue dotted line represents the ideal expected behavior), of the regression experiments performed on the MPAUK catalogue. We used the 4 optical colors derived from the fiber (a), model (b), petrosian (c) SDSS magnitudes, and the 4 colors derived from the fiber SDSS magnitudes + the 3 colors derived from the UKIDSS apermags (d) as input features.

3.4.2 Gas-phase metallicity - Comparison with Sanders et al., 2013

As mentioned in section 1.3.3, Sanders et al. (2013) (hereafter S13) improved the metallicity scatter in the LZ relation by adding color information.
Starting from the full MPA/JHU catalog, S13 performed a series of preliminary cuts on the data:

1. they only used galaxies included in the SDSS MAIN spectroscopic sample (r $<$ 17.77 mag after galactic redenning correction);
2. they limited the sample to galaxies in the redshift range $0.03 < z < 0.3$;
3. following Mannucci et al. (2010), they required that $(S/N)_{H_\alpha} > 25$ and $F_{H_\alpha}/F_{H_\beta} > 2.5$ to improve the signal-to-noise of all relevant emission lines;
4. they used only galaxies for which the fraction of the r-band flux within the SDSS fiber to the full Petrosian flux was > 0.05, to exclude galaxies where the SDSS spectroscopy includes a very little amount of the total flux and may not reflect the galaxy global properties;
5. they excluded AGNs following Kauffmann et al. (2003);
6. they only used galaxies with available K-corrections;
7. they required the metallicities to be in the range $(7 < Log(O/H) + 12 < 9.5)$.

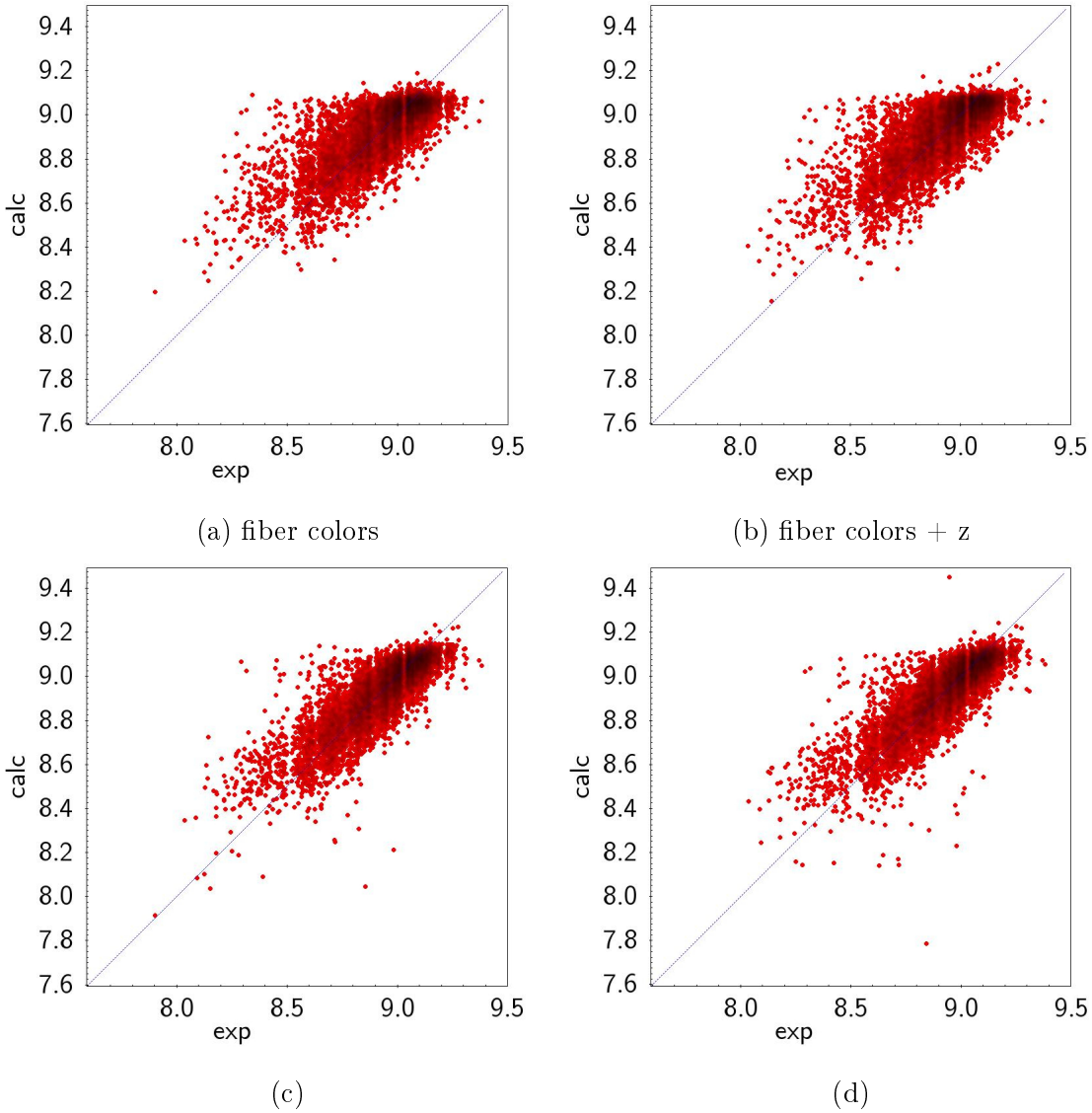
The metallicity dispersion found by S13 varies according to the colors and the metallicity diagnostics used, ranging from $\sigma_z \approx 0.078$ to $\sigma_z \approx 0.094$.
For comparison, we replied the cuts of their list which were based on or obtainable by photometric properties; so, we applied cuts 1-2-5 (actually, we used spectroscopic redshifts and AGN classification from the SDSS, but both could be obtained by photometric properties - Cavuoti et al., 2012, 2014), thus obtaining a catalogue of 31,162 objects, from which as usual we extracted two subsets for training and test (21,183 and 9,349 galaxies, respectively).

As before, for each experiment we show the results after the elimination of the few outliers; the input features and the results for each experiment are listed in table 3.13, while figure 3.21 shows the scatter plots.

Our results are generally consistent with the ones found by S13, with our best result ($\sigma = 0.0920$) obtained when using the four optical SDSS colors, the spectroscopic redshift and the r-band magnitude as input features (however, we point out again that we only partially applied their cuts on the data).

	Whole dataset		**Cleaned by outliers**	
Features	**|bias|**	σ	**|bias|**	σ
fiber colors	0.0004	0.1121	0.0004	0.1121
fiber colors + redshift	0.0006	0.1092	0.0010	0.1091
fiber colors + redshift + r-band fibermag	0.0004	0.0924	0.0005	0.0920
fiber colors + redshift + ugriz fibermags	0.00010	0.1011	0.0005	0.0935

Table 3.13: Input features and results of the regression experiments performed on the MPAUK catalogue, with the S01 cuts. We used the Setup Criteria #001 for all the experiments.



(a) fiber colors

(b) fiber colors + z

(c)

(d)

Figure 3.21: Scatter plots, showing calculated Vs expected metallicity (the blue dotted line represents the ideal expected behavior), of the regression experiments performed on the MPAUK catalogue with the S01 cuts. We used as input features: the 4 SDSS fiber colors (a), the 4 fiber colors + the spectroscopic redshift (b), the 4 fiber colors + the spectroscopic redshift + the SDSS r-band magnitude (c) and the 4 fiber colors + the spectroscopic redshift + the SDSS magnitudes in the 5 ugriz bands (d).

Chapter 4

Summary and discussion

In the previous chapters, we discussed galactic metallicity from a theoretical point of view, and briefly presented the most used methods to estimate it.
As summarized, the determination of metallicity is a quite complicated task, which involves the use of spectral emission or absorption features, according to whether one wants to study gas-phase or stellar metal abundance, respectively; reliable spectra or line-strength indices, however, are only available for nearby galaxies, and usually quite time-consuming to be obtained.
As well known, on the other hand, photometry generally presents various advantages over spectroscopy; photometric observations are easier to be carried out, can be extended to objects fainter than the spectroscopic limit, and are more efficient in terms of the number of objects observed per telescope time, which is a factor of increasing importance in the optic of exploiting ongoing and future large sky surveys.
For these reasons, it would be extremely convenient if we could develop ways to determine galactic metallicities through the observation of photometric properties only.

In this sense, we have presented the application of Machine Learning techniques to the problem of photometric metallicity determination, using a MLP neural network trained by a QNA learning rule.
We applied the method to two different catalogues, a stellar metallicity dataset of

galaxies extracted from the DR9 of the SDSS, and a gas-phase metallicity one from the SDSS DR7.

The first series of experiments performed to obtain accurate stellar metallicity estimates for the galaxies di not provide the expected results, since the network basically produced similar output independently of the expected values.
To better understand the reasons of this discrepancy, we used the BC03 code to generate a mock catalogue of galaxies, assuming a SSP modelization for the objects; thus, we obtained a dataset of artificial galaxies with different metallicities (six discrete classes) and ages for six redshift values.
We performed the same experiments on the mock-catalogue, obtaining better results, since the network yielded predictions consistent with the expected ones, and was able to correctly classify most objects. In particular, we derived an estimate of the RMS of the available colors, and used the highest-RMS ones (which are expected to be more sensitive to metallicity) as input features of the neural network; in this sense, the performance improvement when using these colors was evident.
However, when we tried to exploit the indications collected with these simulated data by using the highest-RMS colors as input features for the network with the real galaxies, the outcome was again unsatisfying.
The reason of the discrepancy between the results with simulated and real data still remains unclear; in the optic of future developments, it would be of great interest to test the network on a mock-catalogue generated by using a more complex model than a SSP, allowing the inclusion of additional effects like the attenuation by dust, to further test the prediction capability of the network.

With gas-phase metallicity, on the other hand, all the experiments revealed the capability of the network to provide reliable estimates.
The first tests were performed on the whole catalogue, using a combination of optical SDSS and IR UKIDSS colors as input features; in this case, the network provided metallicity values consistent with the expected ones, even if with a dispersion slightly worse than the one characterizing the simple LZ relation.
The obtained precision improved when we replicated some of the cuts used by S13 in their derivation of the LZC relation, limiting the dataset to objects belonging to the SDSS MAIN spectroscopic sample and with a redshift value in the range ($0.003 < z < 0.3$). The dispersion found in this case was smaller than the one obtained when using the LZ relation ($\sigma_z \approx 0.10$), and consistent with S13 results, even if still slightly greater than their best result; a further analysis, including the separation of the dataset in bins of different redshift or other photometric bands,

could be performed to improve the prediction performance of the network.

As a last remark, in this work as a first approximation we used spectroscopic values of the redshift when needed; as a future development, we plan to use photometric redshift estimates to verify the impact on the results.

Appendices

Appendix A

Neural network configuration

In this section, we give a brief recap of the main parameters for the configuration of the neural network used for the experiments performed in this work; more details can be found in the text for each particular case.

For each experiment, we divided the involved dataset in two subsets, used for the training and test phase of the algorithm respectively; generally, we used a 70%-30% random partitioning between the two subsets, but we specified their dimensions in each case.
In terms of the internal topology of the network, we alway used a MLP model with *two* hidden layers and the following features:

- **Input neurons:** N photometric features (reported in the text for each experiment, tables 3.1, 3.3, 3.5, 3.11, 3.12, 3.13);
- **Number of neurons in the hidden layers:** as a rule of thumb, we always used 2N+1 nodes in the first hidden layer, and N-1 in the second one;
- **Output layer:** one neuron (corresponding to the metallicity value) in the regression experiments, and six neurons (corresponding to the BC03 metallicity classes)in the classification ones;

	Setup Criteria #01	**Setup Criteria #001**
Max number of iterations	10000	10000
Restarts	60	60
Error threshold	0.00001	0.00001
Decay	0.01	0.001

Table A.1: Different parameters configurations of the neural network used for the experiments in the work.

The following parameters are defined as Setup Criteria for the training phases of the QNA (Brescia et al., 2013):

- **Max number of iterations:** the maximum number of iterations performed for each step of the Hessian approximation;
- **Restarts:** the number of restarts of the Hessian approximation from random positions;
- **Error threshold:** the minimum approximation error at each iteration step;
- **Decay:** It indicates the weight regularization decay. The term decay $\times$ ||network weights||^2 is added to the error function, where the network weights is the total number of weights in the network. When properly chosen, the generalization performances of the network are highly improved;

We used two configurations of Setup Criteria, as reported in table A.1; for each experiment, we specified in the text the criteria employed.

Appendix B

BC03 code - Mock catalogue creation

The Bruzual & Charlot 2003 EPS model allows the computation of the spectral evolution of stellar populations of ages between 1×10^5 and 2×10^{10} yr; the produced models present a spectral resolution of 3Å in the wavelength range $3200 - 9500$ Å, but can be extended at a lower resolution to the wavelength range 91 Å - 160 μm (Bruzual and Charlot, 2003).

The code is provided with files describing the evolution of SSP models trough time. Specifically, for each choice of the IMF (Chabrier or Salpeter; see Chabrier, 2003a,b; Salpeter, 1955) there are six files, one for each of the six possible metallicity values; as detailed in the documentation provided with the code, the files are identified by the conventional filename:

$$\text{bc2003}\_hr/lr\_mXX\_IMF\_\text{ssp.ised}$$

where

- hr/lr: indicates a high or low spectral resolution model, respectively;
- mXX: metallicity class, it can assume 6 different values:

  - m22: Z=0.0001;
  - m32: Z=0.0004;
  - m42: Z=0.004;
  - m52: Z=0.008;
  - m62: Z=0.02;
  - m72: Z=0.05;

- IMF: Chabrier or Salpeter IMF;

and each of these files describes the spectral evolution of that particular SSP model, by providing its SED at 221 unequally spaced time steps.
Therefore, to create the catalogue at a given redshift, for each metallicity class we derived the magnitudes in the 13 photometric bands (ugrizUBVRIJKH) for SSPs with age values taken from the default ones; specifically, for z=0.00, of the 221 default age values we eliminated the smaller ones (Age$<$ 0.1 Gyr), which left us with the 106 age values reported in table B.1.
To create the catalogues at the other redshift values, we increasingly reduced the total number of ages used, by removing those smaller than the corresponding Light Travel Time (LTT) at that redshift in the standard cosmological model assumed (since the age values indicate the age of the galaxy *today*, for increasing redshift we can only observe older galaxies); in table B.2, we indicate the lower age used for each redshift.

Using these age values for the various metallicity classes and the Chabrier IMF (we chose this IMF following Li et al. (2007); Li and Han (2008)), the user can reproduce our catalogue by using the programs **zmag** and **cm_evolution** to obtain the desired magnitudes:

- zmag provides the absolute magnitudes (observer or rest frame, with evolving or non-evolving spectrum) of a galaxy once the user specifies its metallicity, age, redshift and the required photometric band;

- cm_evolution provides a file with the magnitudes of the galaxy for different redshift values (from $z = 0$ to $z = z_{formation}$), once the user specifies the metallicity of the SSP model, the age of the galaxy today and the cosmological parameters used.

Alternatively, we provide a modified version of the zmag program; after specifying the desired metallicity value of the model, the program reads a list of ages and redshift values from an input file, and produces an output file listing the observer-frame

Age (Gyr)	**Age (Gyr)**	**Age (Gyr)**	**Age (Gyr)**	**Age (Gyr)**
0.10152	1.27805	4.75	10.25	15.75
0.11391	1.43400	5.00	10.50	16.0
0.12780	1.60898	5.25	10.75	16.25
0.14340	1.68	5.50	11.00	16.5
0.16090	1.70	5.75	11.25	16.75
0.18053	1.80	6.00	11.50	17.0
0.20256	1.90	6.25	11.75	17.25
0.22727	2.00	6.50	12.00	17.5
0.25500	2.10	6.75	12.25	17.75
0.28612	2.20	7.00	12.50	18.0
0.32103	2.30	7.25	12.75	18.25
0.36020	2.40	7.50	13.00	18.50
0.40415	2.50	7.75	13.25	18.75
0.45347	2.60	8.00	13.50	19.00
0.50880	2.75	8.25	13.75	19.25
0.57088	3.00	8.50	14.00	19.50
0.64054	3.25	8.75	14.25	19.75
0.71870	3.50	9.00	14.50	20.00
0.80640	3.75	9.25	14.75	//
0.90479	4.00	9.50	15.00	//
1.01519	4.25	9.75	15.25	//
1.13907	4.50	10.0	15.50	//

Table B.1: Age values (extracted from the 221 standard ones) used in the creation of the mock catalogue at redshift z=0.00.

Redshift	**Lower age value used (Gyr)**
z=0.10	1.4340
z=0.20	2.6
z=0.30	3.5
z=0.40	4.5
z=0.50	5.25
z=0.60	6

Table B.2: Lower age values used for the creation of the mock catalogue at different redshift values.

Color	RMS	Color	RMS	Color	RMS
J-H	2.480	g-J	1.368	V-r	1.031
z-H	2.293	B-H	1.330	g-z	1.030
z-J	2.268	I-z	1.329	B-z	1.026
z-K	2.137	B-K	1.328	B-g	1.006
I-H	2.059	i-z	1.323	V-R	1.001
I-J	2.028	B-J	1.309	V-I	1.000
i-H	1.981	i-I	1.308	R-I	0.994
I-K	1.962	u-K	1.285	B-I	0.989
i-J	1.949	u-H	1.284	B-r	0.982
i-K	1.901	u-J	1.269	g-I	0.982
R-H	1.763	u-B	1.163	B-R	0.980
R-J	1.730	H-K	1.143	r-I	0.972
R-K	1.716	R-z	1.124	g-r	0.971
r-H	1.675	u-g	1.117	g-R	0.970
r-J	1.642	r-z	1.082	B-i	0.970
r-K	1.639	V-z	1.070	B-V	0.966
J-K	1.620	u-z	1.069	g-i	0.956
V-H	1.553	u-V	1.064	V-i	0.955
V-K	1.533	u-r	1.060	g-V	0.938
V-J	1.524	u-R	1.053	r-R	0.923
g-H	1.390	u-I	1.049	r-i	0.881
g-K	1.385	u-i	1.039	R-i	0.856

(a) z=0.10

Color	RMS	Color	RMS	Color	RMS
J-H	11.089	V-J	1.284	B-R	0.990
J-K	2.701	B-K	1.278	u-i	0.989
z-H	2.228	u-H	1.203	u-I	0.985
z-K	2.011	u-K	1.196	B-i	0.984
I-H	1.974	B-g	1.189	B-I	0.979
i-H	1.879	I-z	1.184	V-z	0.978
I-K	1.835	B-J	1.184	g-z	0.974
z-J	1.808	g-J	1.183	r-R	0.972
i-K	1.762	u-J	1.127	g-V	0.959
R-H	1.684	i-z	1.081	r-i	0.946
I-J	1.648	H-K	1.061	g-R	0.945
R-K	1.610	u-g	1.034	g-i	0.945
r-H	1.605	B-V	1.021	g-I	0.942
i-J	1.578	R-z	1.013	g-r	0.940
r-K	1.546	u-V	1.009	R-i	0.930
v-H	1.458	r-z	1.005	V-i	0.929
R-J	1.441	B-z	1.003	r-I	0.927
V-K	1.422	u-z	0.999	V-R	0.9255
r-J	1.387	u-B	0.996	V-I	0.922
g-H	1.304	u-r	0.994	R-I	0.904
B-H	1.292	u-R	0.993	V-r	0.902
g-K	1.288	B-r	0.990	i-I	0.806

(b) z=0.20

Table B.3: The 66 available colors in the mock catalogue at redshift z=0.10 (a) and z=0.20 (b), ordered by their RMS.

evolving absolute magnitudes in the ugrizUBVRIJKH photometric bands.

After creating the files, we applied the method described in section 3.3.2 to derive the RMS of 66 available colors for each redshift; in tables B.3 to B.5 we report the obtained RMS.

Color	RMS	Color	RMS	Color	RMS
J-H	3.977	u-H	1.222	B-I	1.012
J-K	2.616	u-K	1.214	B-z	1.006
z-H	2.106	R-J	1.200	g-r	0.991
z-K	1.911	r-J	1.167	g-i	0.990
I-H	1.862	B-V	1.106	R-i	0.990
i-H	1.781	B-J	1.096	g-R	0.990
I-K	1.737	V-J	1.091	g-I	0.987
i-K	1.676	u-J	1.087	g-z	0.983
R-H	1.617	g-J	1.084	r-i	0.983
R-K	1.549	u-g	1.080	r-R	0.973
r-H	1.536	u-V	1.076	r-I	0.972
r-K	1.484	g-V	1.067	R-I	0.972
z-J	1.452	u-B	1.050	r-z	0.956
V-H	1.372	H-K	1.043	R-z	0.950
V-K	1.345	u-r	1.040	V-z	0.935
I-J	1.310	u-R	1.037	V-i	0.934
g-H	1.291	u-i	1.033	V-I	0.934
B-H	1.289	u-I	1.032	V-R	0.911
g-K	1.276	u-z	1.029	i-I	0.902
B-K	1.275	B-r	1.026	i-z	0.895
i-J	1.268	B-R	1.020	I-z	0.889
B-g	1.262	B-i	1.016	V-r	0.884

(a) z=0.30

Color	RMS	Color	RMS	Color	RMS
J-H	2.547	B-H	1.188	r-J	0.984
J-K	2.252	u-B	1.167	B-I	0.981
z-H	1.772	z-J	1.143	B-i	0.980
z-K	1.724	u-r	1.117	g-z	0.976
I-H	1.595	u-R	1.096	B-z	0.970
I-K	1.575	u-J	1.095	V-J	0.969
i-H	1.534	B-V	1.094	u-g	0.968
i-K	1.522	u-I	1.075	V-r	0.929
H-K	1.417	u-i	1.075	B-g	0.923
R-K	1.395	u-z	1.069	R-I	0.919
R-H	1.394	I-J	1.051	V-I	0.917
r-K	1.333	i-J	1.045	V-R	0.915
r-H	1.327	g-r	1.041	V-z	0.909
u-V	1.245	B-r	1.018	V-i	0.907
u-K	1.241	g-R	1.013	r-I	0.905
u-H	1.236	g-J	1.011	i-z	0.898
V-K	1.228	i-I	1.006	R-z	0.897
g-K	1.221	R-J	1.004	r-z	0.893
V-H	1.218	B-J	1.003	R-i	0.886
g-H	1.213	B-R	0.999	r-i	0.882
B-K	1.197	g-I	0.990	r-R	0.876
g-V	1.190	g-i	0.988	I-z	0.830

(b) z=0.40

Table B.4: The 66 available colors in the mock catalogue at redshift z=0.30 (a) and z=0.40 (b), ordered by their RMS.

Color	RMS	Color	RMS	Color	RMS
H-K	3.044	B-K	1.165	B-i	0.983
J-K	2.129	g-H	1.155	g-I	0.983
J-H	2.099	u-I	1.152	V-J	0.982
z-K	1.636	u-i	1.140	B-I	0.980
z-H	1.553	B-H	1.136	g-z	0.979
I-K	1.494	u-V	1.127	B-z	0.978
i-K	1.445	u-r	1.125	R-J	0.965
I-H	1.422	V-r	1.115	B-g	0.965
u-K	1.398	u-R	1.114	V-i	0.949
i-H	1.378	g-r	1.103	V-z	0.949
u-H	1.372	g-V	1.088	V-I	0.947
u-J	1.316	z-J	1.072	r-J	0.945
R-K	1.301	B-r	1.065	I-z	0.944
R-H	1.247	B-V	1.037	i-z	0.938
r-K	1.223	g-R	1.034	i-I	0.923
V-K	1.204	I-J	1.027	R-z	0.890
u-B	1.201	B-R	1.020	r-z	0.885
u-z	1.195	i-J	1.015	r-R	0.873
g-K	1.187	V-R	1.004	r-I	0.865
r-H	1.178	g-J	1.001	R-I	0.860
u-g	1.174	B-J	0.997	r-i	0.857
V-H	1.167	g-i	0.987	R-i	0.842

(a) z=0.50

Color	RMS	Color	RMS	Color	RMS
H-K	6.607	u-r	1.143	B-z	1.028
J-K	2.308	g-H	1.139	B-I	1.028
J-H	1.789	V-R	1.134	g-I	1.026
z-K	1.688	B-H	1.132	g-z	1.025
I-K	1.541	g-r	1.129	B-V	1.020
i-K	1.480	u-R	1.121	i-J	1.016
V-r	1.401	B-r	1.106	I-z	1.016
z-H	1.400	u-i	1.099	g-V	1.014
R-K	1.339	u-I	1.086	R-J	0.993
u-K	1.330	g-R	1.079	r-J	0.993
I-H	1.312	B-R	1.071	r-R	0.992
V-K	1.283	u-z	1.053	u-H	0.964
r-K	1.270	V-i	1.051	i-z	0.963
i-H	1.270	z-J	1.050	r-z	0.962
u-B	1.268	B-J	1.040	r-i	0.960
g-K	1.233	V-J	1.040	R-z	0.947
u-g	1.232	I-J	1.040	r-I	0.946
B-K	1.216	g-i	1.039	R-i	0.929
R-H	1.181	B-i	1.038	B-g	0.922
V-H	1.168	g-J	1.035	R-I	0.911
u-V	1.166	V-I	1.031	u-J	0.854
r-H	1.143	V-z	1.029	i-I	0.843

(b) z=0.60

Table B.5: The 66 available colors in the mock catalogue at redshift z=0.50 (a) and z=0.60 (b), ordered by their RMS.

Acknowledgements

Bibliography

Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., et al. (2009). The seventh data release of the sloan digital sky survey. *The Astrophysical Journal Supplement Series*, 182(2):543.

Ahn, C. P., Alexandroff, R., Prieto, C. A., et al. (2012). The ninth data release of the sloan digital sky survey: first spectroscopic data from the sdss-iii baryon oscillation spectroscopic survey. *The Astrophysical Journal Supplement Series*, 203(2):21.

Binney, J. and Tremaine, S. (2011). *Galactic dynamics*. Princeton university press.

Brescia, M. (2012). New trends in e-science: machine learning and knowledge discovery in databases. *Horizons in Computer Science Research, Volume 7, p. 1-73.*, 7:1–73.

Brescia, M., Cavuoti, S., D'Abrusco, R., et al. (2013). Photometric redshifts for quasars in multi-band surveys. *The Astrophysical Journal*, 772(2):140.

Brescia, M., Cavuoti, S., Longo, G., et al. (2014). Dameware: A web cyberinfrastructure for astrophysical data mining. *Publications of the Astronomical Society of the Pacific*, 126(942):783–797.

Brescia, M., Cavuoti, S., Paolillo, M., et al. (2012). The detection of globular clusters in galaxies as a data mining problem. *Monthly Notices of the Royal Astronomical Society*, 421(2):1155–1165.

Brescia, M. and Longo, G. (2013). Astroinformatics, data mining and the future of astronomical research. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 720:92–94.

Broyden, C. G. (1970). The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics*, 6(1):76–90.

Bruzual, G. and Charlot, S. (2003). Stellar population synthesis at the resolution of 2003. *Monthly Notices of the Royal Astronomical Society*, 344(4):1000–1028.

Casali, M., Adamson, A., de Oliveira, C. A., et al. (2007). The ukirt wide-field camera. *Astronomy & Astrophysics*, 467(2):777–784.

Cassisi, S. and Salaris, M. (2013). *Old stellar populations: how to study the fossil record of galaxy formation*. John Wiley & Sons.

Cavuoti, S. (2013). Data-rich astronomy: mining synoptic sky surveys. *PhD Thesis*.

Cavuoti, S., Brescia, M., D'Abrusco, R., et al. (2014). Photometric classification of emission line galaxies with machine-learning methods. *Monthly Notices of the Royal Astronomical Society*, 437(1):968–975.

Cavuoti, S., Brescia, M., Longo, G., et al. (2012). Photometric redshifts with the quasi newton algorithm (mlpqna) results in the phat1 contest. *Astronomy & Astrophysics*, 546:A13.

Chabrier, G. (2003a). The galactic disk mass function: reconciliation of the hubble space telescope and nearby determinations. *The Astrophysical Journal Letters*, 586(2):L133.

Chabrier, G. (2003b). Galactic stellar and substellar initial mass function1. *Publications of the Astronomical Society of the Pacific*, 115(809):763–795.

Charlot, S. and Longhetti, M. (2001). Nebular emission from star-forming galaxies. *Monthly Notices of the Royal Astronomical Society*, 323(4):887–903.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314.

Djorgovski, S. G., Mahabal, A., Drake, A., et al. (2013). Sky surveys. In *Planets, Stars and Stellar Systems*, pages 223–281. Springer.

Eisenstein, D. J., Weinberg, D. H., Agol, E., et al. (2011). Sdss-iii: Massive spectroscopic surveys of the distant universe, the milky way, and extra-solar planetary systems. *The Astronomical Journal*, 142(3):72.

Erb, D. K., Shapley, A. E., Pettini, M., et al. (2006). The mass-metallicity relation at z2. *The Astrophysical Journal*, 644(2):813.

Fletcher, R. (1970). A new approach to variable metric algorithms. *The computer journal*, 13(3):317–322.

Fosbury, R., Villar-Martín, M., Humphrey, A., et al. (2003). Massive star formation in a gravitationally lensed h ii galaxy at z= 3.357. *The Astrophysical Journal*, 596(2):797.

Fukugita, M., Ichikawa, T., Gunn, J., et al. (1996). The sloan digital sky survey photometric system. *The Astronomical Journal*, 111:1748.

Garnett, D. R. (2002). The luminosity-metallicity relation, effective yields, and metal loss in spiral and irregular galaxies. *The Astrophysical Journal*, 581(2):1019.

Goldfarb, D. (1970). A family of variable-metric methods derived by variational means. *Mathematics of computation*, 24(109):23–26.

Gunn, J., Carr, M., Rockosi, C., et al. (1998). The sloan digital sky survey photometric camera. *The Astronomical Journal*, 116(6):3040.

Gunn, J. E., Siegmund, W. A., Mannery, E. J., et al. (2006). The 2.5 m telescope of the sloan digital sky survey. *The Astronomical Journal*, 131(4):2332.

Hambly, N., Collins, R., Cross, N., et al. (2008). The wfcam science archive. *Monthly Notices of the Royal Astronomical Society*, 384(2):637–662.

Hewett, P. C., Warren, S. J., Leggett, S. K., et al. (2006). The ukirt infrared deep sky survey zy jhk photometric system: passbands and synthetic colours. *Monthly Notices of the Royal Astronomical Society*, 367(2):454–468.

Hey, A. J., Tansley, S., Tolle, K. M., et al. (2009). *The fourth paradigm: data-intensive scientific discovery*, volume 1. Microsoft Research Redmond, WA.

Hodgkin, S., Irwin, M., Hewett, P., et al. (2009). The ukirt wide field camera zyjhk photometric system: calibration from 2mass. *Monthly Notices of the Royal Astronomical Society*, 394(2):675–692.

Izotov, Y. I., Stasińska, G., Meynet, G., et al. (2006). The chemical composition of metal-poor emission-line galaxies in the data release 3 of the sloan digital sky survey. *Astronomy & Astrophysics*, 448(3):955–970.

Kauffmann, G., Heckman, T. M., Tremonti, C., et al. (2003). The host galaxies of active galactic nuclei. *Monthly Notices of the Royal Astronomical Society*, 346(4):1055–1077.

Kennicutt Jr, R. C. (1998). The global schmidt law in star-forming galaxies. *The Astrophysical Journal*, 498(2):541.

Kewley, L. J. and Dopita, M. (2002). Using strong lines to estimate abundances in extragalactic h ii regions and starburst galaxies. *The Astrophysical Journal Supplement Series*, 142(1):35.

Kewley, L. J. and Ellison, S. L. (2008). Metallicity calibrations and the mass-metallicity relation for star-forming galaxies. *The Astrophysical Journal*, 681(2):1183.

Kobulnicky, H. A., Kennicutt Jr, R. C., and Pizagno, J. L. (1999). On measuring nebular chemical abundances in distant galaxies using global emission-line spectra. *The Astrophysical Journal*, 514(2):544.

Lawrence, A., Warren, S., Almaini, O., et al. (2007). The ukirt infrared deep sky survey (ukidss). *Monthly Notices of the Royal Astronomical Society*, 379(4):1599–1617.

Lequeux, J., Peimbert, M., Rayo, J., et al. (1979). Chemical composition and evolution of irregular and blue compact galaxies. *Astronomy and Astrophysics*, 80:155–166.

Li, Z. and Han, Z. (2008). Colour pairs for constraining the age and metallicity of stellar populations. *Monthly Notices of the Royal Astronomical Society*, 385(3):1270–1278.

Li, Z., Han, Z., and Zhang, F. (2007). Potential of colors for determining age and metallicity of stellar populations. *Astronomy & Astrophysics*, 464(3):853–857.

Mannucci, F., Cresci, G., Maiolino, R., et al. (2010). A fundamental relation between mass, star formation rate and metallicity in local and high-redshift galaxies. *Monthly Notices of the Royal Astronomical Society*, 408(4):2115–2127.

Maraston, C. (2003). Stellar population models. In *Extragalactic Globular Cluster Systems*, pages 237–248. Springer.

Maraston, C., Pforr, J., Henriques, B. M., et al. (2013). Stellar masses of sdss-iii/boss galaxies at z 0.5 and constraints to galaxy formation models. *Monthly Notices of the Royal Astronomical Society*, page stt1424.

Maraston, C., Strömbäck, G., Thomas, D., Wake, D., and Nichol, R. (2009). Modelling the colour evolution of luminous red galaxies–improvements with empirical stellar spectra. *Monthly Notices of the Royal Astronomical Society: Letters*, 394(1):L107–L111.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

McGaugh, S. S. (1991). H ii region abundances-model oxygen line ratios. *The Astrophysical Journal*, 380:140–150.

McGaugh, S. S. and De Blok, W. (1997). Gas mass fractions and the evolution of spiral galaxies. *The Astrophysical Journal*, 481(2):689.

Peletier, R. F. (2013). Stellar populations. *Secular Evolution of Galaxies*, 1:353.

Pettini, M. and Pagel, B. E. (2004). [o iii]/[n ii] as an abundance indicator at high redshift. *Monthly Notices of the Royal Astronomical Society*, 348(3):L59–L63.

Pettini, M., Shapley, A. E., Steidel, C. C., et al. (2001). The rest-frame optical spectra of lyman break galaxies: star formation, extinction, abundances, and kinematics. *The Astrophysical Journal*, 554(2):981.

Pilyugin, L. (2001). On the oxygen abundance determination in hii regions.-high-metallicity regions. *Astronomy & Astrophysics*, 369(2):594–604.

Pilyugin, L. S. and Thuan, T. X. (2005). Oxygen abundance determination in h ii regions: The strong line intensities-abundance calibration revisited. *The Astrophysical Journal*, 631(1):231.

Salpeter, E. E. (1955). The luminosity function and stellar evolution. *The Astrophysical Journal*, 121:161.

Sanders, N. E., Levesque, E. M., and Soderberg, A. M. (2013). Using colors to improve photometric metallicity estimates for galaxies. *The Astrophysical Journal*, 775(2):125.

Schiavon, R. P. (2007). Population synthesis in the blue. iv. accurate model predictions for lick indices and ubv colors in single stellar populations. *The Astrophysical Journal Supplement Series*, 171(1):146.

Shanno, D. F. (1970). Conditioning of quasi-newton methods for function minimization. *Mathematics of computation*, 24(111):647–656.

Smee, S. A., Gunn, J. E., Uomoto, A., et al. (2013). The multi-object, fiber-fed spectrographs for the sloan digital sky survey and the baryon oscillation spectroscopic survey. *The Astronomical Journal*, 146(2):32.

Smith, J. A., Tucker, D. L., Kent, S., et al. (2002). The ugriz standard-star system. *The Astronomical Journal*, 123(4):2121.

Stasinska, G. (2004). Abundance determinations in h n regions and planetary nebulae. *Cosmochemistry: The Melting Pot of the Elements*, page 115.

Stasińska, G. (2005). Biases in abundance derivations for metal-rich nebulae. *Astronomy & Astrophysics*, 434(2):507–520.

Thomas, D., Maraston, C., and Bender, R. (2003). Stellar population models of lick indices with variable element abundance ratios. *Monthly Notices of the Royal Astronomical Society*, 339(3):897–911.

Tremonti, C. A., Heckman, T. M., Kauffmann, G., et al. (2004). The origin of the mass-metallicity relation: insights from 53,000 star-forming galaxies in the sloan digital sky survey. *The Astrophysical Journal*, 613(2):898.

Van den Bergh, S. (1962). The frequency of stars with different metal abundances. *The Astronomical Journal*, 67:486–490.

Worthey, G. (1994). Comprehensive stellar population models and the disentanglement of age and metallicity effects. *The Astrophysical Journal Supplement Series*, 95:107–149.

Worthey, G., Faber, S., Gonzalez, J. J., et al. (1994). Old stellar populations. 5: Absorption feature indices for the complete lick/ids sample of stars. *The Astrophysical Journal Supplement Series*, 94:687–722.

York, D. G., Adelman, J., Anderson Jr, J. E., et al. (2000). The sloan digital sky survey: Technical summary. *The Astronomical Journal*, 120(3):1579.

Zaritsky, D., Kennicutt Jr, R. C., and Huchra, J. P. (1994). H ii regions and the abundance properties of spiral galaxies. *The Astrophysical Journal*, 420:87–109.

List of Figures

List of Tables