Dedication

To my beloved Alessia,
because she has been a dream come true.
Without her I would have been much worse.

Acknowledgements

There are lots of people I would like to thank for a huge variety of reasons.
Firstly, I would like to thank my Supervisor, Prof. Giuseppe Longo. I could not have imagined having a better advisor and mentor for my PhD, and without his common-sense, knowledge, perceptiveness and enthusiasm I would never have enjoyed so much my work... our work, I better say, because I felt like working with a fellow more than with a Professor. Thank you again!

I would like to thank all the rest of the colleagues, students and friends gravitating around the Laboratory of Astrophysics (the former ”Wild Bunch”... a long list, don't expect I'll thank you one by one!), with whom I spent pleasantly long mornings and short but lively afternoons. A special acknowledgement goes to the brilliant researchers who have put a lot of effort in showing to me how a young scientist and astronomer should think and act, namely Betty e Maurizio. Thank you!

Thanks to Alessandro, Pat, Diego (after all the bothers and the phone interview at least you will get a citation to the article...), Marco and Antonino with whom I shared more than a publication: friendship and deep respect. Thanks also to my few friends here at the Department: we have been following the same steps through a decade of fun and friendship: I hope we won't get lost, wherever our future will take us.

Finally, I have to say a big ”Grazie” to my family: Mum, Dad and Sara, for bearing my wild side, and most importantly, to Alessia, for... for... for everything. And I can't leave out our dog Picabo, who jeopardized the whole damn' thesis chewing my pen-drive few days ago...

Raffaele D'Abrusco
Naples
November 27, 2007

Epigraph

Il pessimismo della ragione e l'ottimismo della volontá

A. Gramsci

Science! true daughter of Old Time thou art!
Who alterest all things with thy peering eyes.
Why preyest thou thus upon the poet's heart,
Vulture, whose wings are dull realities?
How should he love thee? or deem thee wise?
Who wouldst not leave him in his wandering
To seek for treasure in the jeweled skies,
Albeit he soared with an undaunted wing?
Hast thou not dragged Diana from her car?
And driven the Hamadryad from the wood
To seek a shelter in some happier star?
Has thou not torn the Naiad from her flood,
The Elfin from the green grass, and from me
The summer dream beneath the tamarind tree?

To Science, E. A. Poe

The starry night above me and the moral law within me

E. Kant

If you aren't rich, you should always look useful

L. F. Celine

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Weia! Waga!
Woge, du Welle,
walle zur Wiege!

Das Rheingold, Woglinde

1.1 Scientific justification

Observational cosmology, due to the large data-set produced by a new and more powerful generation of dedicated instruments, is undergoing a true paradigm shift in both objectives and methodologies. My primary scientific interests are in observational cosmology This vast field of research, due to the large data-set produced by a new and more powerful generation of dedicated instruments, is undergoing a true paradigm shift in both objectives and methodologies. Data sets in the Terascale or even in the Petascale (of both real and simulated data) are becoming the rule rather than the exception and, while on the one end they allow to tackle old and new problems with unprecedented accuracy and precision, on the other one their scientific exploitation poses challenging and still largely unresolved methodological and computational problems. In this emerging scenario, data mining plays a crucial role in digging out of massive and increasingly complex data sets those significant information (patterns, trends, etc.) which are at the very heart of most open problems in cosmology. These technical issues are hard challenges for the current cosmology and extragalactic astronomy, and stem from both data volume and data complexity since, given the high dimensionality of the parameter space and the high degree of degeneration among the parameters themselves, the extraction of knowledge becomes a highly non-trivial task. For the above reasons, during my undergraduate and PhD studies, my research interest focused both on the science which can be done using multi-wavelength survey data and on the understanding and implementation of new data mining methodologies which can help in the study of the large scale structure. It needs to be stressed that the work presented in the following pages finds its natural collocation in the ongoing efforts to build an International Virtual Observatory infrastructure.

1.2 About this thesis

The structure of this thesis is the following: a general review of the problems and opportunities raised by the ongoing shift in methodology and availability of data in astronomy and cosmology are discussed in the chapter 2. My contributions to the field and the description of the results achieved during my Ph.D. are contained in the chapters 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 of this thesis. Each topic is structured as follows: an introductory chapter outlining some of the most relevant background and the papers which have been published (or submitted for publication) in scientific journals. A strict separation between methodological and applicative work has been applied throughout this thesis in order to describe as clearly as possible the amount of work necessary to achieve both technical and scientific results and the problems of different nature encountered and addressed during the development of the techniques and their applications to astronomical topics. The last chapter contains the conclusions of this thesis, namely a brief summary of the main results I have obtained during my Ph.D. programme and a prospect of the future developments of my research.

1.3 Summary or arguments

During my Ph.D., I tackled several arguments strictly intertwined, which can be broadly grouped in two large areas: algorithm development and observational cosmology. The first part led me to develop data-mining methods for the characterization of the distribution of sources in the astronomical high-dimensional parameter spaces and for the retrieval of information therein contained through clustering and dimensionality reduction. The second part of my Ph.D. programme focused mainly on two different topics involving the application of the aforementioned methods:

- Photometric redshifts estimation finalized to cosmological studies. I developed an original algorithm for reliable photometric redshift classification based on a supervised application of a particular model of neural network (Multi Layer Perceptron). The process consists of two steps involving two independently trained sets of neural networks (NN). In the first step of the process, NNs are used to extract photometrically homogeneous classes of galaxies. In the second step, a second set of NNs is run on these photometrically homogeneous samples and trained on the corresponding "knowledge base" (consisting of galaxies for which a spectroscopic redshift is available). The method has been used to calculate photometric redshifts for the SDSS DR5 Galaxy sample (D'Abrusco et al. 2007), with an accuracy of ~ 0.019, which represents one of the best results ever obtained. Moreover, this method leads to a remarkably gaussian error distribution over the whole redshift range. This last consideration is important for the statistical and cosmological applications of the photometric redshift catalogues. I am currently working on the exploitation of these data for the derivation of statistical observables, such as the luminosity function and the 3-dimensional correlation function of galaxies from SDSS DR5 dataset, and also on the extension of this method to higher redshift. A brief introduction to redshift definition, photometric redshift methods and principles of neural networks is found in chapter 3, while the description of my original work can be found in chapter 4. Three of the main current applications of the photometric redshifts catalogue produced during my Ph.D., I have collaborated to are described in chapters 6, 8 and 10 respectively, each one proceeded by a chapter containing an introduction to the specific subject and brief review of the available literature (5,7,9).

- Efficient and unbiased candidate quasars selection. During the second part of my PhD programme, while involved in the scientific exploitation of the photometric redshifts catalogues, I have also worked at the development of an original technique for the selection of candidate quasars, especially during a period spent at the Institute of Astronomy of the University of Cambridge. From the methodological point of view, the method I have implemented is based on unsupervised clustering in the colour space carried out by a combination of two different dimensionality reduction and clustering algorithms respectively, namely the Probabilistic Principal Surfaces (PPS) and the Negative Entropy Clustering (NEC). The method exploits the knowledge on the position in the parameter space of spectroscopically confirmed quasars to characterize the distribution of photometric sources and extract possible candidates. This selection algorithm, applied to the combined SDSS and UKIDSS survey data, has provided a significant improvement respect to the alternative approaches available in the literature, and has the advantage of being easily generalizable to other wave bands using a different base of knowledge. A general preamble regarding the observational featured of Active Galactic Nuclei and the description of the main approaches found in the literature for the candidate quasars selection is contained in chapter 11, while my original work is described in 12.

Chapter 2

A new paradigm in astronomy

The path is clear,
though no eyes can see
the course laid down
long before.

Firth of Fifth, Genesis

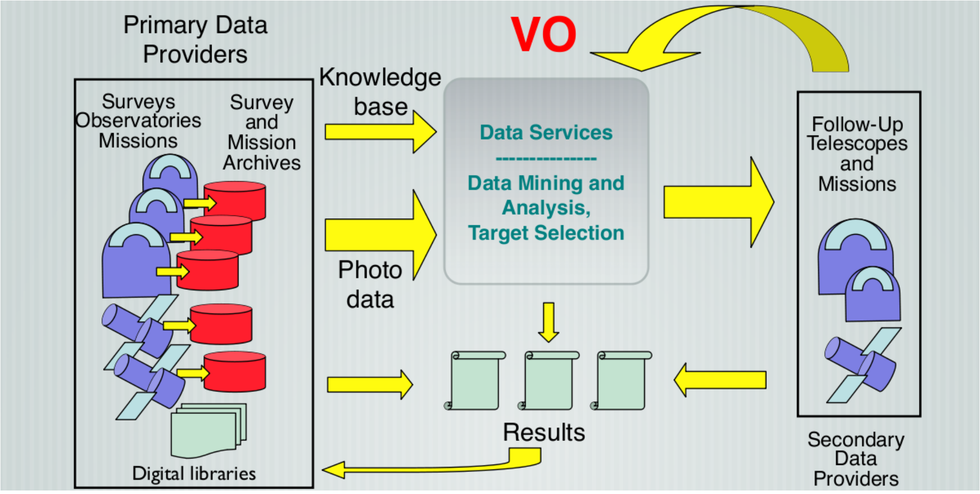
2.1 The Virtual Observatory

Until the first half of the 20th century, the observational bases of our knowledge of the Universe were built almost exclusively upon information extracted from the electromagnetic emission of celestial bodies, and in particular from the emissions spanning the region of the electromagnetic spectrum from $\sim 3300\AA$ to $\sim 8000\AA$, corresponding to the visible wavelengths often thermally emitted by objects with temperature between 3500 K° and 10000 K°. In the specific case of astronomy the elementary units of light, photons, are very useful probes for obtaining insights of the structure and properties of the sources for several reasons. First of all, photons are produced in a variety of physical processes and therefore can be used to investigate a large sample of physical conditions and environments. The characteristics of the emitted photons are strictly correlated to the conditions of the emitting objects: nature and dynamical state of the emitting particles (atoms, molecules or dust grains), temperature, pressure and presence of magnetic fields all leave unique fingerprints on the properties of the light reaching the observers. Photons interact very weakly with the scarce interstellar matter, and since the absorbed fraction of light depends on the wavelength λ of the radiation, the dimming can be easily corrected for. Even if the optical band has been the preferred window of human investigation of the Universe for historical, technological and evolutionary reasons, the possibility of observing the Universe in a wide range of different wavelengths is a quite recent achievement of observational astronomy, allowed by the huge developments of technology only in the last 50 years. It is not surprising that still today a general and efficient tool designed to combine and extract information from these heterogeneous data is missing. Only in the last few years, thanks to the huge growth of computational power

and the widespread diffusion of internet, a new tool, the so-called *International Virtual Observatory* (IVO), has been created to address this issue. The Virtual Observatory is an international astronomical community-based initiative, aimed at allowing global electronic access to the available astronomical data archives of space and ground-based observatories, sky survey databases. It also aims to enable data analysis techniques through a coordinating entity that provides common standards, wide-network bandwidth, and state-of-the-art analysis tools (Walton et al. 2006). IVO will make possible in the near future to have powerful and expensive new observing facilities at wavelengths from the radio to the X-ray and gamma-ray regions, through the federation of a vast new array of current and planned astronomical data sets at all wavelengths, together with advanced instrumentation techniques. These very large databases are archived and made accessible in a systematic and uniform manner to realize the full potential of the new observing facilities. In a more technical fashion, the VO aims at providing the framework for global access to the various data archives by facilitating the standardization of archiving and data-mining protocols, and will take advantage of state-of-the-art advances in data-handling software in astronomy and in other fields.
The birth of the Virtual Observatory will also urge researchers to turn to the GRID paradigm of distributed computing and resources to solve complex, front-line research problems. In order to implement this new paradigm, one has to join existing astronomical data centres and archives into an interoperating and single unit. This new astronomical data resource will form a Virtual Observatory (VO) so that astronomers can explore the digital Universe in the new archives across the entire spectrum. Similarly to how a real observatory consists of telescopes, each with a collection of unique astronomical instruments, the VO consists of a collection of data centres each with unique collections of astronomical data, software systems, and processing capabilities, which will enable new science.

2.2 Astronomical parameter space

The notion of parameter space of astronomical data can be introduced as follows. A n-dimensional space P^N whose axis are defined as observables, i.e. astronomical quantities derived from any kind of astronomical measurements, is given as a subsample of the space R^N, such that $P^N \subset R^N$. In this scheme, every astronomical observation O is associated to more observables quantities (for example, a measured flux $f_A(t)$ is associated at least at the number t representing the time when the the flux has been measured, and to the number (or interval of numbers) A standing for the band within which the flux has been observed), so that each observation can be linked to a point $\mathrm{o} \in R^m \subset R^N$, where usually $m \ll N$. In this sense, any astronomical observation is an


Primary Data Providers
VO
Surveys Observatories Missions
Survey and Mission Archives
Knowledge base
Data Services

Data Mining and Analysis, Target Selection
Photo data
Follow-Up Telescopes and Missions
Results
Digital libraries
Secondary Data Providers

Figure 2.1: *Simplified representation of the Virtual Observatory scheme.*

incomplete information defining a $(N-m)$-dimensional manifold inside P^N, and, for example, the position of any object will define a O^{N-2} manifold embedded in P^N, since it can be identified on the celestial sphere by two coordinates. A complete understanding of the observational aspects of the Universe could be achieved through a uniform and dense sampling of the parameter space P^N, while the current status of observations permits only to cover such space only sparsely. Some regions, as the region of optical range spanned by wavelength between $\sim 3300\mathring{A}$ and $\sim 8000\mathring{A}$, angular resolution between $0.8''$ and $2''$, integration time ranging from 10 to 3600 seconds, magnitudes between 14.0 and 20.0, are very well sampled, while observations in other zones of this complex parameter space are too sparse and underpopulated.

The introduction of the astronomical parameter space offers the possibility to observe the history of astronomical discoveries from a new standpoint. All advances in astronomy can be interpreted and can be accounted for by one of the following processes:

- exploration of regions of the astronomical parameter space mostly unknown thanks to the development of new observational techniques or measurement methods which provide data of completely new nature or covering still unexplored intervals of already known observables.

- search for correlations between apparently disjointed regions of the parameter space (these regions are actually connected when all dimensions of the parameter space are considered, but since only low dimensionality projected subspace R^m of P^N are accessible, they seem to be detached.)

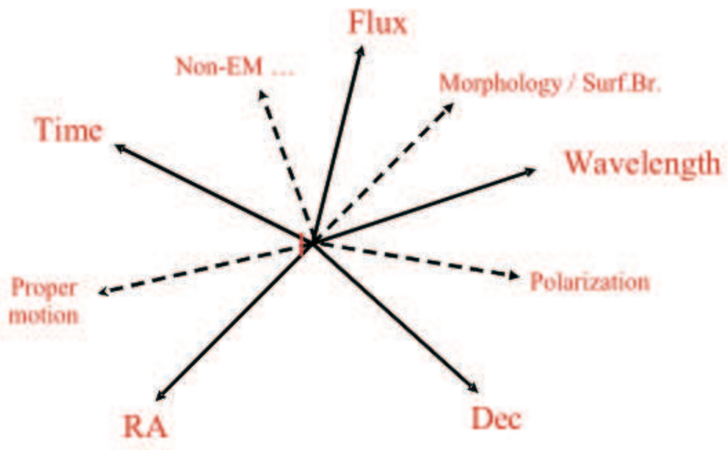


Figure 2.2: *Schematic representation of astronomical parameter space, where axes associated to different measurable quantities (right ascension, declination, flux, time, etc.) are shown.*

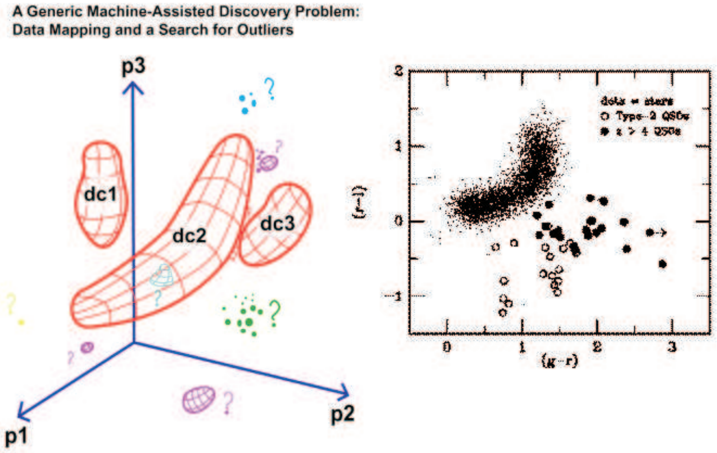


Figure 2.3: *On the left, the displacement of simulated points representing astronomical observations in a 3-dimensional parameter space and a possible clustering of these points are shown. In the figure on the right a real case is displayed, showing how candidates quasars (filled and open circles) are placed far from the region occupied by stars (dots) in a colour-colour plane (i.e, a particular 2-dimensional slice of the astronomical parameter space). This issue will become more evident in the chapters 11,12.*

On the other hand, the introduction of this spatial description of astronomical discoveries has deeply changed also the way observations are decided and planned: one of the leading consideration taken into account when preparing observations is that meeting new and potentially interesting phenomena is much easier when sampling scarcely populated regions of the parameter space. The analysis and visualization of a high dimensionality space are challenging and represent very heavy tasks from a computational point of view. The tools currently available do not permit to visualize more than $4/5$ parameters at the same time once that data of different nature have been difficulty gathered by matching positionally the sources considered. For this reason, the development of new and more powerful methods to automatically find and combine heterogeneous observations of the same objects or the same region of the sky in order to get as much as possible information about the largest manifold R^m of the parameter space, and to detect correlations between a high number of observables is a compelling need in order to fully exploit the potentiality of this new approach to astronomical discovery. The Virtual Observatory, as explained in paragraph 2.1, will answer many of these questions.

2.3 A VO powered approach to the Large Scale Structure of the Universe

My involvement in the development of new techniques for astronomical data mining and exploration to be used inside the VO, has been triggered by and focused on the problem of the accurate determination of the distribution of galaxies in the Universe (see chapter 3) and of their classification in physical types (see chapter 11). These apparently different tasks can be both tackled through the characterization of the galaxy distribution on the surface of a manifold embedded inside the general astronomical parameter space, where the position of each point is assigned by a physical position specified by three numbers (usually α, δ and the redshift z) and a label (which can be associated to a discrete or continuum variable) indicating at which class the source belongs. The dependence of the classification on the observational parameters chosen to define the set of ”essential qualities” allowing to discriminate between physically distinct sources and then used to assign each object to a particular class, will be further discussed in a more general fashion in chapter 11. Here we will concentrate on the general philosophy driving this approach to the map-making of the Universe. Even if the same name of this paragraph suggests an obvious analogy between the creation of a map of the distribution of galaxies in the Universe and the main goal of cartography (i. e. the production of detailed and as accurate as possible representations of the surface of the earth), a deeper examination clearly indicates that an even more profound connection and similarity occur with biological taxonomy. The effort of producing maps of the Universe is very similar to the task faced by taxonomists when trying to construct the classification of living creatures, because both endeavours can be ultimately reduced to the search for proximity and similarity in a huge populations of multiform individuals. The most complex aspect of both taxonomies is indeed the choice of the optimal sets

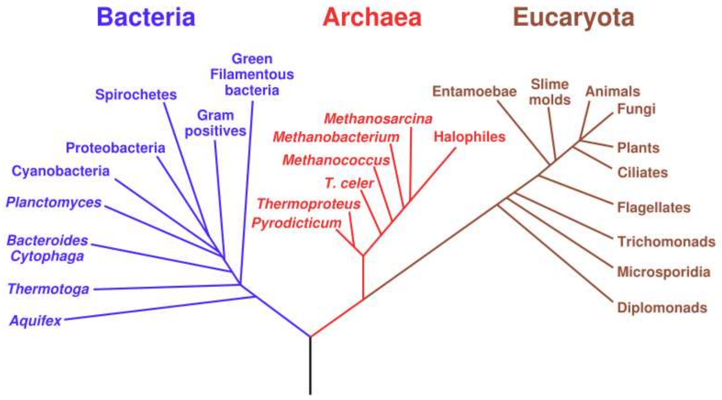


Figure 2.4: *A phylogenetic tree of living things, based on RNA data showing the separation of* bacteria*,* archaea *and* eukaryotes*. Genetic evidences suggest that* eukaryotes *evolved from the union of some* bacteria *and* archaea*, one becoming the nucleus and the other the main cell. This is the example of a phylogenetic hypothesis supported by genetic observations, where the driving mechanism of the evolution is the physical combination of two different organisms.*

of distinguishing characters that can be used to determine the representative individuals of each class and, as a consequence, the very nature of the whole classification. For biological classification, the possibility to interact with the individuals and populations greatly helps and shows the direction to follow in order to choose effectively the optimal set of parameters, whereas in the astronomical case this approach is impossible since no direct interaction with the population of objects under scrutiny is possible. For this reason, the number and nature of parameters that can be used is somehow limited by the same nature of astronomy. A deeper similarity between these two apparently distant fields is embodied by the possibility of a phylogenetic interpretation of the taxonomy. Phylogenetic systematics (also phylogenetics) is the field of biology aimed at the identification and comprehension of the evolutionary relationships among the many different kinds of life on Earth, both living and dead. According to phylogenetics, the differentiation of organisms during the history of life, i.e. the evolution of living beings, is regarded as a branching process whereby populations have been altered over time and have had the opportunity to speciate into separate branches, hybridize together again or terminate by extinction, so that similar classes of organisms at the present time, in first approximation, can be considered deriving by a common biological ancestor. In this scenario, the current biological taxonomy is a consequence of the evolution of the species, and can be used to reconstruct at least partially the evolutionary steps leading to the current distinct populations and classes of different organisms (see an example of phylogenetic tree in figure 2.4).

Phylogenetics based on the only observation of the current biological taxonomy has

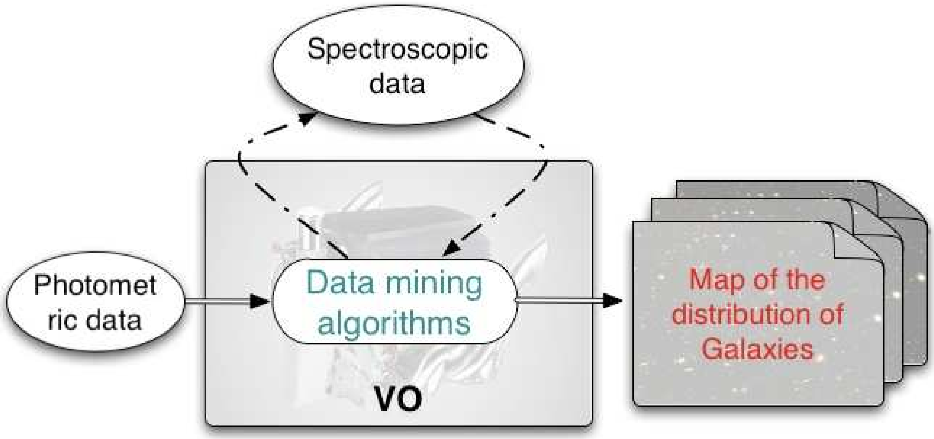


Figure 2.5: *Illustration of a VO powered approach to the study of the distribution of galaxies in the Universe, based on the exploitation of mixed photometric and spectroscopic survey data. Data mining and information extraction algorithms exploit the large availability of photometric data to determine the three dimensional position and classification of galaxies through a back-reaction process involving a spectroscopic sample used as a benchmark for the optimization of the performances of the techniques. Examples of applications of this philosophy are showed in chapters 4 and 12.*

proved only partly correct, because of the discovery of mechanisms which affect the evolution of life but can not be traced back to the current classification or, in the worst case, can mislead the interpretation of the observed similarities and common properties among the contemporary species (for example, evolutionary convergence and variations of external characters caused by environmental pressure or random genetic mutation). This consideration has led to the necessity of developing a phylogenetic systematics based not only on the current appearance of the species (the phenotype), but founded on the genetic characteristics of the organisms and on the comparison of DNA sequences belonging to different extant or extinct populations. Anyway, the shift from a static attitude in taxonomy towards a dynamical and evolutionary interpretation of current classification has represented a breakthrough in biology and genetics; this advance in the comprehension of the evolution of organisms has been supported by the availability of remarkable evidences of the past steps of the evolution and current species, mainly fossil records and genetic data, which have become indispensable for a correct construction of a modern phylogenetic tree.

A similar revolution in the comprehension of the relations between the current population of galaxies and their classification schemes, and the evolutionary steps which older populations of galaxies (i.e., at higher redshifts) have undergone during the history of the Universe and how different ”generations” of galaxies are related to each other, is expected to occur in the field of astronomy. Astronomy is ready to the big step from taxonomy (or phenotypical classification) to phylogenetics (a classification based on the actual nature of galaxies instead of their appearance, and for this reason, useful to identify evolutionary relations and mechanisms). This advance will be possible as soon as a reliable physical classification of galaxies will become available together with large samples of galaxies in various epochs (i.e. at different z), and a deeper comprehension of the main mechanisms regulating the evolution of galaxies[1] (see table 2.1). Many classifications of galaxies in the past focused their attention on the morphology, typically as seen in the blue, characterized by the prominence of spirals, bars, rings, etc. (see paragraph 11.3 for a description of Hubble's classification). Even if these details represent interesting phenomena, they usually involve only a small fraction of the total mass and do not carry much information of the basic physical properties of the galaxy. For physical classification of galaxies, therefore, we intend a taxonomy based on the fundamental properties, such as the total mass, density, energy and angular momentum, and the nearest observables thereof, like luminosities in different bands, the spectral energy distribution, colours and surface brightness of galaxies. The understanding of how the observables above are connected to the appearance and to the physical parameters which actually shape up the properties of a galaxy is a challenge which has already been tackled by astronomers by is still far from being successfully completed. A heuristically motivated example of phylogenetic interpretation of the current morphological classification of galaxies (figure 2.6) has been proposed by Djorgovski in (Djorgovski 1992).

The Virtual Observatory infrastructure can provide the natural environment for the creation of a phylogenetics of the observed populations of galaxies and investigate the physical mechanisms of the evolution and formation of the large scale structure of the Universe. The availability of large samples of galaxies is a key requirement to develop an accurate map of the Universe and consequently generate a physical taxonomy and understanding of the connections between different types of galaxies. The capability of VO to gather, federate and provide to the astronomical community a coherent arrangement of the overwhelming mass of data produced by the multi-band surveys of the entire sky reveals greatly useful because allowing a new approach to the extragalactic astronomy based on the massive application of information retrieval algorithms on refined datasets obtained thanks to the application of data mining techniques on raw large datasets. The most convenient type of information which can be inferred by astronomical sources is photometry, because photometric observations require lower integration

[1] It is worth recalling that astronomy has an advantage over biology in this case, since the properties of galaxies which have lived in past epochs are still accessible, and pictures of the Universe at different ages can be taken simply observing objects at different redshifts

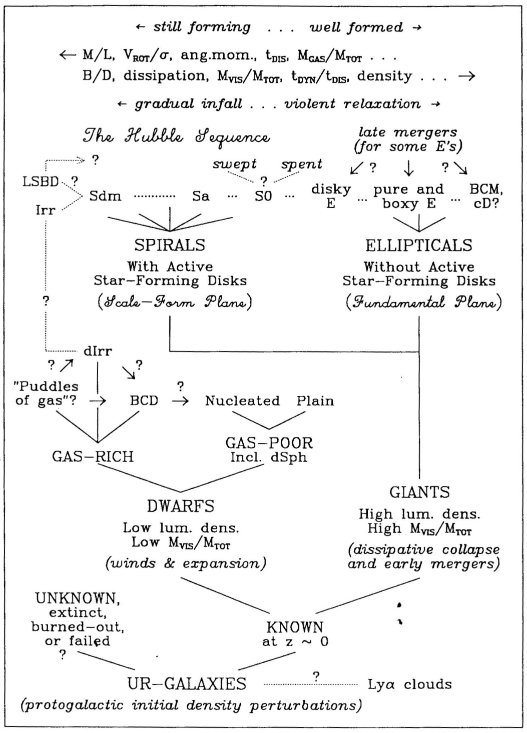

← still forming . . . well formed →
← M/L, V_{ROT}/σ, ang.mom., t_{DIS}, M_{GAS}/M_{TOT} . . .
B/D, dissipation, M_{VIS}/M_{TOT}, t_{DYN}/t_{DIS}, density . . . →
← gradual infall . . . violent relaxation →
The Hubble Sequence
late mergers (for some E's)
swept
spent
?
LSBD
Irr
Sdm
Sa
S0
disky E
pure and boxy E
BCM, cD?
SPIRALS
With Active Star-Forming Disks
(Scale-Form Plane)
ELLIPTICALS
Without Active Star-Forming Disks
(Fundamental Plane)
dIrr
"Puddles of gas"?
BCD
Nucleated
Plain
GAS-RICH
GAS-POOR
Incl. dSph
DWARFS
Low lum. dens.
Low M_{VIS}/M_{TOT}
(winds & expansion)
GIANTS
High lum. dens.
High M_{VIS}/M_{TOT}
(dissipative collapse and early mergers)
UNKNOWN, extinct, burned-out, or failed
KNOWN at z ~ 0
UR-GALAXIES
Lyα clouds
(protogalactic initial density perturbations)

Figure 2.6: *A modern view of galaxy taxonomy with a phylogenetics interpretation of the origins and evolution of morphological classification of galaxies.*

time for a given signal to noise ratio and reach fainter sources than spectroscopic ones, let alone the capability of acquiring data for a very large number of sources at the same time. For this reason, a map of the large scale structure of the Universe should be inferred primarily by photometric data in order to assuring more reliable statistics and allowing to reach regions of the Universe at high redshift (i.e. older populations of galaxies). On the other hand, the characterization of galactic sources in terms of photometric data only is also useful as a benchmark for the data mining and information extraction algorithms, since the manifold of the astronomical parameter space determined by colours is usually much simpler than the full dimensional region spanned by galaxies. In this sense, this parameters are used to test the performances of the techniques which will be applied more extensively on a larger set of parameters carrying a larger amount of information on the physical properties of galaxies than the simple photometry. Spectroscopic data are nonetheless useful, because in both tasks the mapping of

General taxonomy		Biological taxonomy		Galaxy taxonomy
Classes	⟷	Species	⟷	Galaxy types
+		+		+
Evolution mechanisms	⟷	Fossils and genetic laws	⟷	Galaxies at different z and physical mechanisms
↓		↓		↓
Phylogenetics of populations	⟷	Phylogenetics of species	⟷	?

Table 2.1: *Schematic table of comparative characteristics of taxonomy, evolutionary mechanisms and corresponding phylogenetics in a generic case, for biological classification and galaxy classification.*

the distribution of galaxies can be split into (redshift determination and classification of galaxies), spectroscopic priors (the spectroscopic z and spectroscopic classification for a significant subsample of the photometric galaxies) can be used to check and improve the accuracy of the results. This process of extraction of the information from photometric datasets is illustrated in figure 2.5. During my PhD, I have worked on the development of this approach to observational cosmology, encouraged by the belief that a new epoch of discoveries and advances will shortly improve out comprehension of the properties and evolution of galaxies in the Universe.

Chapter 3

Photometric redshifts

Ripple in still water,
When there is no pebble tossed,
Nor wind to blow.

Ripple, The Grateful Dead

3.1 Redshift: the definition

When a source of radiation moves toward or away from an observer, the observed frequencies of photons will be different from their emitted frequencies. Quantitatively, if v_r is the radial velocity of a source that emits photons that have frequency ν_0 in the rest frame of the source, then these photons will be detected at frequency:

$$\nu = (1 - \beta)\gamma\nu_0 \tag{3.1}$$

where c is the velocity of light. To the lowest order in $\frac{v}{c}$, $\gamma = 1$ and the frequency shift is:

$$\Delta\nu \equiv \nu - \nu_0 = -\beta\nu_0 \tag{3.2}$$

In terms of wavelengths $\lambda = \frac{c}{\nu}$, it is necessary to this order in $\frac{v}{c}$ that $\frac{\Delta\lambda}{\lambda_0} = -\frac{\Delta\nu}{\nu_0}$ so that:

$$\Delta\lambda = \frac{v_r}{c}\lambda_0 \tag{3.3}$$

Thus $\Delta\lambda$ for a receding source is positive, and the observed spectrum is redshifted relative to its rest wavelength; on the other hand, for an approaching source, $\Delta\lambda$ is negative and the observed spectrum is blue-shifted. In general, the redshift of an object is defined to be:

$$z \equiv \frac{\Delta\lambda}{\lambda_0} \tag{3.4}$$

In the early part of the twentieth century, Slipher, Hubble and others made the first measurements of the redshifts and blue shifts of galaxies beyond the Milky Way. While the interpretation of the observed red and blue shifts in terms of Doppler effect was immediate, much less obvious was the later discovery by Lundmark and Hubble of a rough correlation between the increasing redshifts and the increasing distance of galaxies (Lundmark 1925). Theorists almost immediately realized that these observations could be explained by a different mechanism for producing redshifts. In fact, Hubble's law of the correlation between redshifts and distances is required by models of cosmology derived from general relativity that have a metric expansion of space. As a result, photons propagating through the expanding space are stretched, creating the cosmological redshift. This cosmological effect differs from the simple Doppler effect because the velocity boost (i.e. the Lorentz transformation) between the source and observer is not due to classical momentum and energy transfer, but the photons emitted by the receding source increase their wavelength as the space-time through which the emitter is travelling expands. This effect is prescribed by the current cosmological model as an observable manifestation of the time-dependent cosmic scale factor (a) in the following way:

$$1 + z = \frac{a_0}{a_t} \tag{3.5}$$

where a_t is the scale factor measured at time t before present time, while a_0 is the current value of the same scale factor. This type of redshift is called cosmological redshift or Hubble redshift. It needs to be stressed that, according to the cosmological interpretation of redshift, the galaxy located at redshift z is not receding simply by means of a physical velocity in the direction away from the observer; instead, the intervening space between the observer and the emitting galaxy is stretching, so accounting for the large-scale isotropy of the effect demanded by the cosmological principle. For redshifts of $z < 0.1$, the effects of space-time expansion are minimal and observed redshifts dominated by the peculiar motions of the galaxies relative to one another that cause additional Doppler redshifts and blue shifts, while at higher redshifts the cosmological component is almost always dominant respect to proper motion component. From a mathematical viewpoint, the statements "distant galaxies are receding" and "the space between galaxies is expanding" are related by a change of coordinate systems. Expressing this change in a correct mathematical framework requires working with the formalism of the Friedmann-Robertson-Walker metric in the theory of the general relativity.

3.2 Spectroscopic measurement of redshift

The spectrum of light emitted by a single source can be measured. To determine the redshift of the source, features in the spectrum such as absorption lines, emission lines or other variations in light intensity are searched for. If found, these features can be compared with known features in the spectrum of various chemical compounds as measured in the local rest frame, i.e. inside a laboratory located on the earth. If the same pattern of intervals between emission and absorption lines is seen in an observed spectrum from a distant source but occurring at shifted wavelengths respect to an observed reference spectrum, this pattern can be associated to the same chemical element whose rest frame spectrum have been identified and measured in the laboratory. If the same spectral line is identified in both spectra but at different wavelengths, the redshift of the astronomical source can be derived. Determining the redshift of an object in this way requires a whole range of spectral features observed and identified in both astronomical and laboratory template spectra and cross-correlating them, so that the value of redshift z_{best} which best accommodates the shift in wavelength of the identified spectral features can be associated with the actual redshift of the source z.

3.3 Photometric measurement of redshift

The technique of photometric redshifts has a relatively long history. Its first applications date back to the sixties, but recently the photometric redshifts technique has experienced a burst of interest due to the fact that many deep multicolour photometric surveys have been carried out, providing photometric information on large numbers of objects inaccessible to spectroscopic observations both because they are to faint for the current spectroscopic limits or just because too numerous to be observed with the available multiple object spectrographs. Photometric redshifts provide an estimate of the redshift of galaxies (or AGNs, quasars) using only large/medium band photometry instead of spectroscopy. The efficiency of the method relies in the identification of spectral breaks, i.e. strong spectral features, still recognizable after the integration of the Spectral Energy Distribution (SED) below the filter's transmission function. The precision of this estimate is worse than the spectroscopic redshift one, depending on the filters set and on the photometric accuracy, but for many cosmological and extragalactic applications the photometric redshift represents a sufficient information. Thus, the photometric redshift tool is nowadays extensively applied and has rapidly become a crucial tool of observational cosmology. Two methods for photometric redshift estimation are largely used in the literature, namely the empirical training set method (Connolly et al. 1995, Brunner et al. 1997, D'Abrusco et al. 2007), and the SED fitting (Lanzetta et al. 1996, Fernández-Soto et al. 1999). A brief introduction to both these techniques is presented below.

3.3.1 Empirical training set methods

The empirical training set method constructs a direct empirical correlation between colours of the sources and their redshifts. The essence of this approach is to derive a function between redshift and photometric data by using a large and representative training set of galaxies for which both photometry and redshift are known (the so-called base of knowledge), and then use this function to estimate the redshifts of objects for which only photometric information are known, and no estimation of redshift is available. Several different tools have been used to determine the shape of the function: linear or non-linear fitting (Brunner et al. 1997, Wang et al. 1998, Budavári et al. 2005), support vector machines (Wadadekar 2005); artificial neural network (Firth et al. 2003, Ball and Loveday 2004, Collister and Lahav 2004, Vanzella et al. 2004, Li et al. 2006); instance-based learning (Csabai et al. 2003, Ball et al. 2007), nearest-neighbour estimations (Csabai et al. 2003). All these methods have their pro's and con's which will be better described, at least in part, in what follows.

3.3.2 SED fitting method

The SED fitting procedure is based on the overall fit of the shape of the spectrum of the source whose redshift is under scrutiny, and on the detection and isolation of strong spectral properties. In order to obtain more secure results, the filter set must be chosen in order to bracket some of these features, as the 4000 break or the Lyman break at 912. The observed photometric SEDs are compared to those obtained from a set of reference spectra using the same photometric system. These template SEDs can be either observed or synthetic, i.e. derived by the observation of an assorted sample of real galaxies or produced by simulating the stellar population and emission properties of different kind of galaxies. The photometric redshift $z_{\rm phot}$ of a given source corresponds to the best fit of its photometric SED by the set of template spectra, in general through a χ^2 minimization procedure. The observed SED of a given galaxy is compared to a set of template spectra:

$$\chi^2(z) = \sum_{i=1}^{N_{filters}} \left[\frac{F_{obs,i} - bF_{temp,i}(z)}{\sigma_i} \right]^2 \tag{3.6}$$

where $F_{obs,i}$, $F_{temp,i}$ and σ_i are the observed and template fluxes and their uncertainty in filter i, respectively, and b is a normalization constant. A combination of this method with the Bayesian marginalization introducing an a priori probability was proposed by (Benítez 2000), which demonstrated that in this case the dispersion of $z_{\rm phot}$ can be significantly improved. Alternatively, the photometric redshift estimate can be safely improved introducing the Bayesian inference when prior information is not related to the photometric properties of sources. Examples of such priors that could be combined with the $z_{\rm phot}$ technique are the morphology or the clues inferred from gravitational lensing modelling. The major advantages of the SED fitting technique are its simplicity and the fact that it does not require any spectroscopic sample and it can be extended to objects located at different redshift that the sample of galaxies whose spectra have formed the SEDs template set. At the same time, this is also its weak point, since the method needs a fiducial choice of spectral templates with the underlying hypothesis that these SEDs are valid for all objects.

3.4 Artificial neural networks: general remarks

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurones, and this mechanism is adopted by ANNs as well. Neural networks, with their remarkable ability to derive meaning from complicated or imprecise data, can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques.

A trained neural network can be thought of as an "expert" in the category of information it has been given to analyze, and can provide estimates regarding one or more parameters associated to the sample on which it has been trained. Two of the main advantages of using a neural network for pattern recognition and/or classification tasks are reported below:

- Adaptive learning, i.e. the ability to learn how to do tasks based on the data given for training or initial experience accumulated by the neural network;
- Self-organization, i.e. the capability of the ANN to preserve and modify its own organization or the internal representation of the information it has received during learning time.

3.5 Models of artificial neural networks

ANNs are simple mathematical models defining a function $f : X \rightarrow Y$. Each type of ANN model corresponds to a class of such functions. The word network in the term

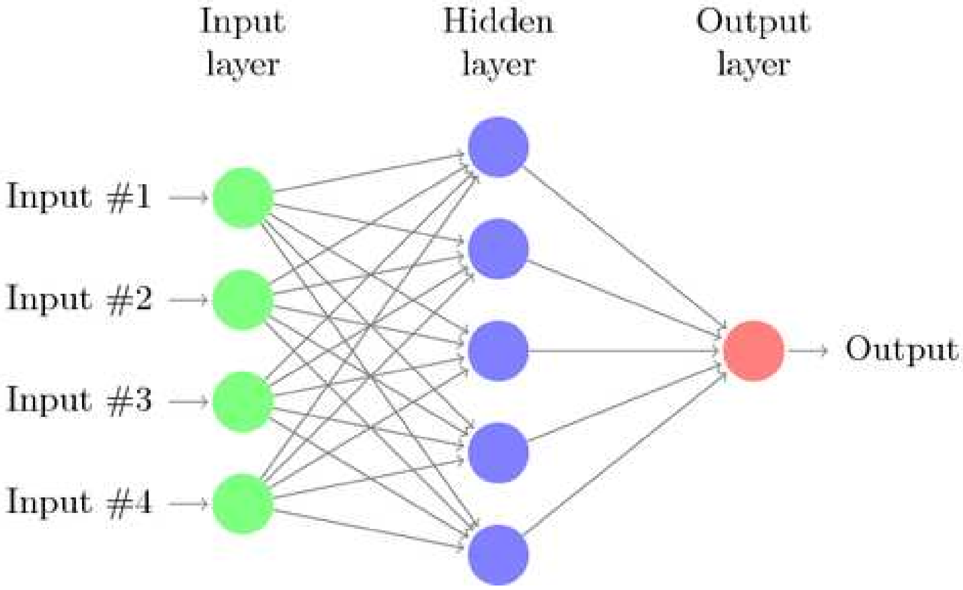


Figure 3.1: *Simple model of an artificial neural network. The input, hidden and output layers and the connections between different neurons are sketched.*

'artificial neural network' arises because the function $f(x)$ is defined as a composition of other functions $g_i(x)$, which can further be defined as a composition of other functions. This web of dependencies between variables can be conveniently represented as a network structure. The simplest architecture of an artificial neural network is a stacking of different layers, each composed by a variable number of neurons .Three different types of layer, namely the input, hidden and output layers, are can be distinguished according to their position within the ANN: the input layer is composed by neurons receiving as input external signals, the hidden layers receive their inputs from the input or other hidden layers and, after processing the signal, they pass it to another layer of neurons, and finally the output layer supply the final processed signal to an external device. A schematic representation of this template neural network architecture can be seen in picture 3.1. A widely used but simple type of composition is the non-linear weighted sum, where $f(x) = K\big(\sum_i w_i g_i(x)\big)$, where K is some predefined function, such as the hyperbolic tangent. The collection of functions g_i is indicated as a vector g $= (g_1, g_2, \dots, g_n)$. The figure 3.2 depicts such a decomposition of the function f with dependencies between variables indicated by arrows. These can be interpreted in a probabilistic frame as follows: the random variable $F = f(G)$ depends upon the random variable $G = g(H)$, which depends upon $H = h(X)$, which depends upon the random variable X. The same process can be seen from a functional point view, in these terms: the input x is transformed into a 3-dimensional vector h, which is then transformed into a 2-dimensional vector g, which is finally transformed into f. In either case, for this particularly simple network architecture, the components of individual layers are independent of each other (e.g., the components of g are independent of each other, given their input h). Networks such as the one described above are commonly called feedforward, because their graph is a directed acyclic graph, while networks with cycles are commonly called recurrent.

3.5.1 Learning process

However interesting the functions defining the dependencies between variables in a given neural network architecture may be in themselves, the most interesting feature in ANN is its capability of learning. Given a specific task to solve, and a class of functions F, learning means using a set of observations (hereafter base of knowledge), in order to find $f^\star \in F$ which solves the task in an optimal sense. This entails defining a cost function $C : F \rightarrow \mathbb{R}$ such that, for the optimal solution $f^\star$, $C(f^\star) \leq C(f) \forall f \in F$ or, literally, that no solution has a cost less than the cost of the optimal solution. The cost function C is an important concept in learning, because it is a measure of how far away from the optimal solution to the problem encountered, the current solution provided by the ANN is. Learning algorithms search through the solution space in order to find a function that has the smallest possible cost. For applications where the solution is dependent on some data, the cost must necessarily be a function of the observations, otherwise there would not be any correlation between the model of the ANN and the data. As a simple example, consider the problem of finding the model f which minimizes $C = E\left[(f(x) - y)^2\right]$, for data pairs (x, y) drawn from some distribution $\mathcal{D}$. In practical situations N samples are considered from $\mathcal{D}$ and thus, for the above example, $\hat{C} = \frac{1}{N} \sum_{i=1}^{N} (f(x_i) - y_i)^2$ is minimized. In other terms, the cost is minimized over a sample of the data rather than the true data distribution. When $N \rightarrow \infty$, some forms of online learning must be used where the cost is partially minimized as soon as each new example is seen. While online learning is often used when $\mathcal{D}$ is fixed, it is most useful in the case where the distribution changes slowly over time. In neural network methods, some form of online learning is frequently also used for finite datasets.

3.5.2 Cost function

While it is possible to arbitrarily define some ad hoc cost function, frequently a particular cost will be used either because it has desirable properties (such as convexity) or because it arises naturally from a particular formulation of the problem (i.e., In a probabilistic formulation the posterior probability of the model can be used as an inverse cost).

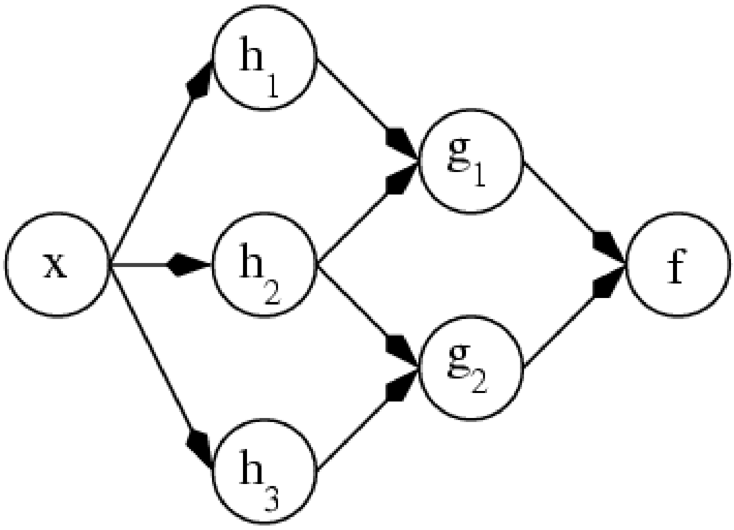


Figure 3.2: *Artificial neural network dependencies graph*

3.5.3 Learning tasks

The choice of a particular cost function depends on the task the ANN is asked to perform. Three learning paradigms are used in modelling of neural networks: supervised learning, unsupervised learning and reinforcement learning. The main characteristics of each of these paradigms are summarized in the list below:

- Supervised learning. In supervised learning, a set of example pairs (x, y) with $x \in X, y \in Y$ is given and the aim of the learning process is to find a function f in the allowed class of functions that matches the examples, i.e. which represents the mapping function inferred by the data. In this case, the cost function is related to the mismatch between the initial mapping and the data and it implicitly contains prior knowledge about the problem domain. A commonly used cost is the mean-squared error which tries to minimize the average error between the network's output $f(x)$ and the target value y over all the example pairs. When this cost is minimized using gradient descent for the class of neural networks called Multi-Layer Perceptrons, the so-called backpropagation algorithm for training neural networks is obtained. Tasks that fall within the paradigm of supervised learning are pattern recognition (also known as classification) and regression (also known as function approximation).

- Unsupervised learning. In unsupervised learning, once the data x is given, the cost function to be minimized can be any function of the data x and the output of the network f. The cost function is dependent on the task assigned to the ANN and on the *a priori* assumptions (the implicit properties of the model, its parameters and the observed variables). Tasks that fall within the paradigm of unsupervised learning are in general estimation problems; the applications include clustering, the estimation of statistical distributions, compression and filtering.

- Reinforcement learning. In reinforcement learning, data x is usually not given but generated by an agent's interactions with the environment. At each point in time t, the agent performs an action y_t and the environment generates an observation x_t and an instantaneous cost c_t, according to some (usually unknown) dynamics. The aim is to discover a policy for selecting actions that minimizes some measure of a long-term cost, i.e. the expected cumulative cost. The environment's dynamics and the long-term cost for each policy are usually unknown, but can be estimated. More formally, the environment is modelled as a Markov decision process (MDP) with states $\{s_1, \ldots, s_n\} \in S$ and actions $\{a_1, \ldots, a_m\} \in A$ with the following probability distributions: the instantaneous cost distribution $P(c_t, |, s_t)$, the observation distribution $P(x_t, |, s_t)$ and the transition $P(s_t + 1, |, s_t, a_t)$, while a policy is defined as conditional distribution over actions given the observations. Taken together, the two define a Markov chain. The aim is to discover the policy that minimizes the cost, i.e. the MC for which the cost is minimal. ANNs are frequently used in reinforcement learning as part of the overall algorithm.

3.5.4 Learning algorithms

Training a neural network model essentially means selecting one model from the set of allowed models (or, in a Bayesian framework, determining a distribution over the set of allowed models) that minimizes the cost criterion. There are numerous algorithms available for training neural network models; most of them can be viewed as a straightforward application of optimization theory and statistical estimation. Most of the algorithms used in training artificial neural networks are employing some form of gradient descent technique. This is done by simply taking the derivative of the cost function with respect to the network parameters and then changing those parameters in a gradient-related direction. Evolutionary methods, simulated annealing, and Expectation-Maximization (hereafter EM) and non-parametric methods are among other commonly used methods for training neural networks.

3.6 Redshift surveys

Mapping out the large scale distribution of individual galaxies is a step toward understanding the origin and evolution of the large scale structure of the Universe. With the recent advent of automated telescopes and improvements in spectroscopes and photometric redshift estimation techniques, larger and larger volumes of the Universe have been covered and mapped combining redshifts with angular positions, so providing a 3-dimensional map of the distribution of visible matter. The first redshift survey was the CfA Redshift Survey (Huchra et al. 1988), started in 1977 with the initial data collection completed in 1982. More recently, the 2dF Galaxy Redshift Survey (Sadler et al. 2002) determined the large-scale structure of one section of the Universe, measuring z-values for over 220,000 galaxies (see figure 3.3). Another remarkably large mixed (photometric and spectroscopic) ongoing survey is the Sloan Digital Sky Survey (SDSS) which has obtained (both spectroscopic and photometric) estimates for about 100 million objects (Adelman-McCarthy et al. 2006); in particular, SDSS has provided reliable spectroscopic measurements of redshifts for about 10^6 galaxies out to 0.5 and has detected quasars beyond z = 6. Among the many ongoing spectroscopic survey projects aimed at exploring regions of the Universe at higher redshift than the above mentioned SDSS and characteristically covering smaller area of the sky but comparable sampled volumes thanks to the peculiar geometry of the Universe, DEEP2 deserves a specific citation (Davis et al. 2001). The DEEP2 Redshift Survey exploits the Keck telescopes with the "DEIMOS" spectrograph and represents a follow-up to the pilot programme DEEP1. DEEP2 is designed to observe faint galaxies with redshifts 0.7 and above, and it has been therefore planned to provide a complement at higher redshifts of SDSS and 2dF surveys, which both cover much larger fields of view and have explored carefully the nearby Universe.

3.7 Applications of photometric redshifts to Cosmology

Todays most pressing cosmological questions demand the construction of galaxy surveys of unprecedented depth and volume. Such questions include whether some different form of dark energy is causing the accelerating rate of cosmic expansion driven by Einsteins cosmological constant, what are the properties of this dark energy, and whether competing models of inflation may be discriminated by accurate measurements of the shape of the primordial power spectrum of mass fluctuations. As already stated, galaxy surveys delineate the large-scale structure of the Universe and thereby provide a powerful and independent constraint on the cosmological models. The currently favoured concordance model in which $\approx 70\%$ of the energy density of todays Universe is resident in a relatively unclustered form known as dark energy is evidenced by a combination of observations of the Cosmic Microwave Background CMB) (Spergel et al. 2003) with those of galaxy clustering (Percival et al. 2001). These independent datasets are required to break the degeneracy between model parameters and render a unique cosmology. According to standard cosmological theory, if the linear regime

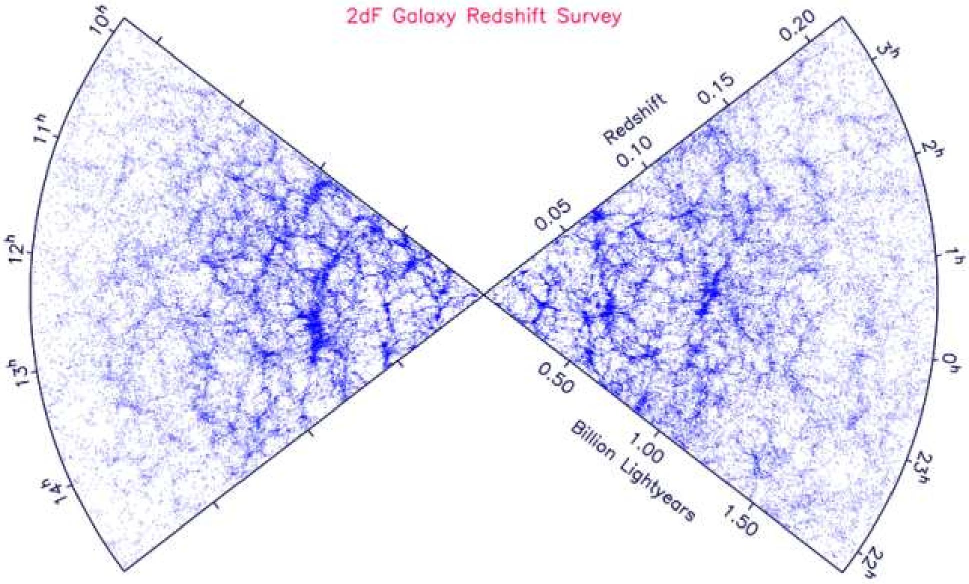


Figure 3.3: *Cone plot of the distribution in redshift space of the 2dF survey galaxies.*

clustering power spectrum is measured with sufficient precision then it will no longer appear smooth and monotonic: specific features and modulations will become apparent. Two such attributes are predicted: firstly, a series of acoustic oscillations sinusoidal modulations in power as a function of scale imprinted in the baryonic component before recombination (Hu and Sugiyama 1996) and secondly, a turnover a broad maximum in clustering power on large scales originating from the radiation-dominated epoch. These features encode characteristic cosmological scales that can be extracted from the observations, greatly improving constraints upon cosmological models (Seo and Eisenstein 2003). Moreover, other currently-unknown modulations in power (e.g. signatures of inflation) may be discovered when the clustering pattern is examined with sufficiently high precision. Very recently, the acoustic signature has been convincingly identified for the first time in the clustering pattern of Luminous Red Galaxies in the Sloan Digital Sky Survey. The 2dF Galaxy Redshift Survey has produced consistent measurements, thus confirming the previous hints on the existence of the acoustic signature in the clustering pattern.

The current challenge is to make more accurate measurements at different redshifts, using these features to further constrain the cosmological parameters, in particular the dark energy model. At low redshift the available volume is limited, so that the effect of cosmic variance is significant so that it is virtually impossible obtaining an unbiased picture of the spatial distribution of structures and galaxies. For these reasons, shallow spectroscopic surveys are insensitive to clustering modes on very large scales and are hampered by non-linear growth of structure on small scales (see for example figure 3.3). Large-scale surveys at higher redshifts are, as a consequence, required to map greater cosmic volumes, whose exploration will allow to trace clustering modes with longer wavelengths and will reveal the pattern of linear clustering to significantly smaller scales. The high redshift spectroscopic surveys currently in progress (e.g. DEEP2 (Davis et al. 2003)), cover solid angles of $\approx 10 \ \mathrm{deg}^2$, which are insufficient for detecting the predicted features in the clustering power spectrum. Such projects are fundamentally limited by existing instrumentation, making use of spectrographs with relatively small fields of view ($\approx 10^\circ \div 20^\circ$) and restricted (albeit impressive) multi-object capabilities. In this context, the role that photometric redshift catalogues derived from deep imaging surveys can play in addressing the scientific goals outlined above is of great importance. Extensive imaging surveys (covering $\sim 10,000 \ \mathrm{deg}^2$ to reasonable depths $(r \approx 22)$) have been recently completed and expansions to wider areas are ongoing (e.g. SDSS); one of the results of these scientific enterprises will be the exploitation of the implied redshift distribution maps over cosmic distances to $z \approx 1$ with sufficient number density in order to measure the clustering of galaxies: these measurement are now limited by cosmic variance rather than by shot noise like in the past. Future deeper imaging surveys like PanSTARRS (Kaiser et al. 2000) and the CTIO Dark Energy Survey are being planned to address a host of scientific questions. Such surveys will also provide powerful measurements of features in the galaxy clustering pattern, contingent of the fact that methods for the calculation of photometric redshifts will provide accurate measurement and reliable estimation of the uncertainty affecting those estimates. The utility of photometric redshifts derived from broadband galaxy colours rather than from spectra has been well-established so far by analysing the results obtained on the already available surveys. It needs to be noticed that the blurring of large-scale structure in the radial direction due to the photometric redshift error degrades measurements of the clustering pattern. However, on physical scales larger than that implied by the redshift error, the information is preserved. Moreover, on smaller scales the tangential information always survives, and the vast area which may be readily covered by an imaging survey can potentially provide more independent structure modes on a given scale than those yielded by a fully spectroscopic survey of a smaller solid angle, implying very competitive cosmological constraints. Photometric redshifts have already been used to construct volume-limited samples of low-redshift galaxies and measure their angular clustering properties (Budavári et al. 2003, Cooray et al. 2001). The cosmological parameter constraints resulting from future photometric redshift imaging surveys have been simulated also by (Blake and Bridle 2005).

Published as: D'Abrusco, R., Staiano, A., Longo, G., Brescia, M., Paolillo, M., De Filippis, E., Tagliaferri, R. – *"Mining the SDSS archive. I. Photometric redshifts in the nearby universe."* 2007, Astrophysical Journal, vol. 663, issue 2, 752

Chapter 4

Photometric redshifts in the nearby universe

Abstract

In this paper we present a supervised neural network approach to the determination of photometric redshifts. The method, even though of general validity, was fine tuned to match the characteristics of the Sloan Digital Sky Survey (SDSS) and as base of 'a priori' knowledge, it exploits the rich wealth of spectroscopic redshifts provided by this unique survey. In order to train, validate and test the networks, we used two galaxy samples drawn from the SDSS spectroscopic dataset, namely: the General Galaxy sample (GG) and the Luminous Red Galaxies subsample (LRG). Due to the uneven distribution of measured redshifts in the SDSS spectroscopic subsample, the method consists of a two steps approach. In the first step, objects are classified in nearby ($z < 0.25$) and distant ($0.25 < z < 0.50$), with an accuracy estimated in 97.52%. In the second step two different networks are separately trained on objects belonging to the two redshift ranges. Using a standard Multi Layer Perceptron operated in a Bayesian framework, the optimal architectures were found to require 1 hidden layer of 24 (24) and 24 (25) neurons for the GG (LRG) sample. The presence of systematic deviations was then corrected by interpolating the resulting redshifts. The final results on the GG dataset give a robust $\sigma_z \simeq 0.0208$ over the redshift range $[0.01, 0.48]$ and $\sigma_z \simeq 0.0197$ and $\sigma_z \simeq 0.0238$ for the nearby and distant samples respectively. For the LRG subsample we find instead a robust $\sigma_z \simeq 0.0164$ over the whole range, and $\sigma_z \simeq 0.0160$, $\sigma_z \simeq 0.0183$ for the nearby and distant samples respectively. After training, the networks have been applied to all objects in the SDSS Table GALAXY matching the same selection criteria adopted to build the base of knowledge, and photometric redshifts for ca. 30 million galaxies having $z < 0.5$ were derived. A second catalogue containing photometric redshifts for the LRG subsample was also produced. Both catalogues can be downloaded at the URL: http://people.na.infn.it/ astroneural/SDSSredshifts.htm .

4.1 Introduction

After the pioneering work by the Belgian astronomer Vandererkhoven, who in the late thirties used prism-objective spectra to derive redshift estimates from the continuum shape and its macroscopic features (notably the Balmer break at ~ 4000 Å), (Baum 1962)

was the first to test experimentally the idea that redshift could be obtained from multiband aperture photometry by sampling at different wavelengths the galaxy spectral energy distribution (hereafter SED). After a period of relative lack of interest, the 'photometric redshifts' technique was resurrected in the eighties (Butchins 1981), when it became clear that it could prove useful in two similar but methodologically very different fields of application:
i) as a method to evaluate distances when spectroscopic estimates become impossible due to either poor signal-to-noise ratio or to instrumental systematics, or to the fact that the objects under study are beyond the spectroscopic limit (Bolzonella et al. 2002);
ii) as an economical way to obtain, at a relatively low price in terms of observing and computing time, redshift estimates for large samples of objects.
The latter field of application has been widely explored in the last few years, when the huge data wealth produced by a new generation of digital surveys, consisting in accurate multiband photometric data for tens and even hundreds of millions of extragalactic objects, has become available. Photometric redshifts are of much lower accuracy then spectroscopic ones but even so, if available in large number and for statistically well controlled samples of objects, they still provide a powerful tool to derive a 3-D map of the universe. A map which is crucial for a variety of applications among which we shall quote just a few: to study large scale structure (Brodwin et al. 2006); to constrain the cosmological constants and models ((Blake and Bridle 2005) and references therein, (Budavári et al. 2003) and (Tegmark et al. 2006)); to map matter distribution using weak lensing ((Edmondson et al. 2006) and references therein).
In this paper we present a new application of neural networks to the problem of photometric redshift determination and use the method to produce two catalogues of photometric redshifts: one for $\sim$ 30 million objects extracted form the SDSS-DR5 main GALAXY dataset and a second one for a Luminous Red Galaxies sample.
The paper is structured as it follows. In the Sections 4.2 and 6.3, we shortly summarize the various methods for the determination of photometric redshifts, and the theory behind the adopted model of neural network. In § 4.4, we describe both the photometric data set extracted from the SDSS and the base of knowledge used for the training and test and, in § 4.5 we discuss the method and present the results of the experiments. It needs to be stressed that even though finely tailored to the characteristics of the SDSS data, the method is general and can be easily applied to any other set provided that a large enough base of knowledge is available.
As stressed by several authors, photometric redshift samples are useful if the structure of the errors is well understood; in § 4.7 we therefore present a discussion of both systematic and random errors and propose a possible strategy to correct for systematic errors (§ 4.6). In § 4.8 we shortly describe the two catalogues. Finally, in § 4.9, we discuss the results and present our conclusions.

This paper is the first in a series of three. In the second one we shall present the catalogue of structures extracted in the nearby sample using an unsupervised clustering algorithm working on the three dimensional data set produced from the SDSS data. In paper III we shall complement the information contained in the above quoted catalogues by discussing the statistical clustering of objects in the photometric parameter space.

4.2 Photometric redshifts

Without entering into too much detail, photometric redshifts methods can be broadly grouped in a few families: template fitting, hybrid and empirical methods.

Template fitting methods are based on fitting a library of template Spectral Energy Distributions (SEDs) to the observed data, and differ mainly in how these SEDs are derived and in how they are fitted to the data. SEDs may either be derived from population synthesis models (Bruzual A. and Charlot 1993) or from the spectra of real objects (Coleman et al. 1980) carefully selected in order to ensure a sufficient coverage of the parameter space (mainly in terms of morphological types and/or luminosity classes). Both approaches (synthetic and empirical) have had their pro's and con's widely discussed in the literature, (cf. (Koo 1999), but see also (Fernández-Soto et al. 2002), (Massarotti, Iovino and Buzzoni 2001),(Massarotti, Iovino, Buzzoni and Valls-Gabaud 2001) and (Csabai et al. 2003)). Synthetic spectra, for instance, sample an 'a priori' defined grid of mixtures of stellar populations and may either include unrealistic combinations of parameters, or exclude some unknown cases. On the other end, empirical templates are necessarily derived from nearby and bright galaxies and may therefore be not representative of the spectral properties of galaxies falling in other redshift or luminosity ranges. Ongoing attempts to derive a very large and fairly exhaustive set of empirical templates using the SDSS spectroscopic dataset are in progress and will surely prove useful in a nearby future.

Hybrid SED fitting methods making use of a combination of both observed and theoretically predicted SEDs have been proposed with mixed results by several authors (Bolzonella et al. 2000, Padmanabhan et al. 2005).

The last family of methods, id est the empirical ones, can be applied only to 'mixed surveys', id est to datasets where accurate and multiband photometric data for a large number of objects are supplemented by spectroscopic redshifts for a smaller but still significant subsample of the same objects. These spectroscopic data are used to constrain the fit of an interpolating function mapping the photometric parameter space and differ mainly in the way such interpolation is performed. As it has been pointed out by many authors (Connolly et al. 1995, Csabai et al. 2003), in these methods the main uncertainty comes from the fact that the fitting function is just an approximation of the complex relation existing between the colours and the redshift of a galaxy and by the fact that as soon as the redshift range and/or the size of the parameter space increase, a single interpolating function is bound to fail. Attempts to overcome this problem have been proposed by several authors. For instance, (Brunner et al. 1999), divided the redshift and colour range in several intervals in order to optimize the interpolation. (Csabai et al. 2003) used instead an improved nearest neighbor method consisting in finding, for each galaxy in the photometric sample, the galaxy in the training set which has the smallest distance in the parameter space and then attributing the same redshift to the two objects.

More recently, several attempts to interpolate the a priori knowledge provided by the spectroscopic redshifts have been made using statistical pattern recognition techniques such as neural networks (Tagliaferri et al. 2002, Vanzella et al. 2004, Firth et al. 2003) and Support Vector Machines (Wadadekar 2005), with results which will be discussed more in detail in what follows.

It has to be stressed that since the base of knowledge is purely empirical (*i.e.* spectroscopically measured redshifts), these methods cannot be effectively applied to objects fainter than the spectroscopic limit. To partially overcome this problem, noticeable attempts have been made to build a 'synthetic' base of knowledge using spectral synthesis models, but it is apparent that, in this case, the uncertainties of the SED fitting and empirical methods add up.

In any case, it is by now well established that when a significant base of knowledge is available, empirical methods outperform template fitting ones and that the use of the latter should be confined to those case where a sui base of knowledge is missing.

4.3 The Multi Layer Perceptron

Neural Networks (hereafter NNs) have long been known to be excellent tools for interpolating data and for extracting patterns and trends and since few years they have also dug their way into the astronomical community for a variety of applications (see the reviews (Tagliaferri 2003a, Tagliaferri 2003b) and references therein) ranging from star-galaxy separation, spectral classification (Winter et al. 2004) and photometric redshifts evaluation (Tagliaferri et al. 2002, Firth et al. 2003). In practice a neural network is a tool which takes a set of input values (input neurons), applies a non-linear (and unknown) transformation and returns an output. The optimization of the output is performed by using a set of examples for which the output value is known a priori. NNs exist in many different models and architectures but since the relatively low complexity of astronomical data does not pose special constrains to any step of the method which will be

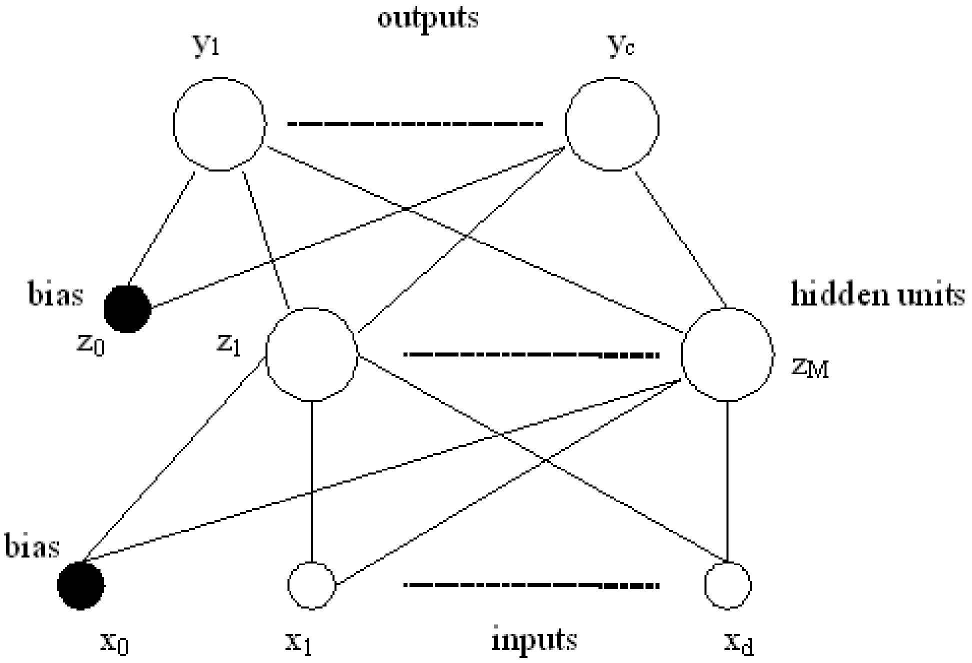


Figure 4.1: *A schematic representation of the Multi Layer Perceptron architecture.*

discussed below we used a very simple neural model known as *Multi-Layer Perceptron or MLP* which is probably the most widely used architecture for practical applications of neural networks.

In most cases an MLP consists of two layers of adaptive weights with full connectivity between inputs and intermediate (namely, hidden) units, and between hidden units and outputs (see Fig. 4.1).

Note, however, that an alternative convention is sometimes also found in literature which counts layers of units rather than layers of weights, and regards the input as separate units. According to this convention the network showed in Fig. 4.1 would be called three-layer network. However, since the layers of adaptive weights are those which really matter in determining the properties of the network function, we refer to the former convention.

4.3.1 MLP: the flux of the computation

The MLP realizes a complex nonlinear mapping from the input to the output space. Let us denote the N input values to the network by $\mathbf{x} = \{x_1, x_2, \ldots, x_d\}$. The first layer of the network forms a linear combinations of these inputs to give a set of intermediate activation variables $a_j^{(1)}$

$$a_j^{(1)} = \sum_{i=1}^{d} w_{ji}^{(1)} x_i + b_j^{(1)}, j = 1, \ldots, M \tag{4.1}$$

with one variable $a_j^{(1)}$ associated with each of the M hidden units. Here $w_{ji}^{(1)}$ represents the elements of the first-layer weight matrix and $b_j^{(1)}$ are the biases parameters associated with the hidden units. The variables $a_j^{(1)}$ are then transformed by the nonlinear activation functions of the hidden layer. Here we restrict attention to $tanh$ activation functions. The outputs of the hidden units are then given by

$$z_j = \tanh(a_j^{(1)}), j = 1, \ldots, M \tag{4.2}$$

The z_j are then transformed by the second layer of weights and biases to give the second-layer activation values $a_k^{(2)}$

$$a_k^{(2)} = \sum_{j=1}^{M} w_{kj}^{(1)} z_j + b_k^{(2)}, k = 1, \ldots, c \tag{4.3}$$

where c is the number of output units. Finally, these values are passed through the output-unit activation function to give output values y_k, where $k = 1 \ldots, c$. Depending on the nature of the problem under consideration we have:

- for regression problems: a linear activation function, i.e. $y_k = a_k^{(2)}$;
- for classification problems: a logistic sigmoidal activation functions applied to each of the output independently, i.e.:

$$y_k = \frac{1}{1 + \exp(-a_k^{(2)})},$$

4.3.2 MLP Training Phase

The basic learning algorithm for MLPs is the so called *backpropagation* and is based on the error-correction learning rule. In essence, backpropagation consists of two passes through the different layers of the network: a forward pass and a backward pass. In the forward pass an input vector is applied to the input nodes of the network, and its effect propagates through the network layer by layer. Finally, a set of outputs is produced as the actual response of the network. During the backward pass, on the other hand, the weights are all adjusted in accordance with the error-correction rule. Specifically, the actual response of the network is subtracted from a desired (target) response (which we denote as a vector $\mathbf{t} = \{t_1, t_2, \ldots, t_c\}$) to produce an error signal. This error signal is then propagated backward through the network. There are several choices for the form of the error signal to produce and this choice still depends on the nature of the problem, in particular:

- for regression problems we adopted the sum-of-squares error function:

$$E = \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{c}\{y_k(\mathbf{x}^n; \mathbf{w}) - t_k^n\}^2;$$

- for classification problems we used the cross-entropy error function:

$$E = -\sum_{n}\sum_{k=1}^{c}\{t_k^n \ln y_k^n + (1 - t_k^n)\ln(1 - y_k^n)\}.$$

The weights are adjusted to make the actual response of the network move closer to the desired response in a statistical sense. In this work we adopted a computational more efficient variant of the backpropagation algorithm, namely the quasi-newton method. Furthermore, we employed a weight-decay regularization technique in order to limit the effect of the overfitting of the neural model to the training data, therefore the form of the error function is:

$$\tilde{E} = E + \nu\frac{1}{2}\sum_{i} w_i^2,$$

where the sum runs over all the weight and biases. The ν controls the extents to which the penalty term $\frac{1}{2}\sum_i w_i^2$ influences the form of the solution.

It must be stressed that the universal approximation theorem (Haykin 1999) states that the two layers architecture is capable of universal approximation and a considerable number of papers have appeared in the literature discussing this property (cf. (Bishop 1995) and reference therein). An important corollary of this result is that, in the context of a classification problem, networks with sigmoidal nonlinearities and two layer of weights can approximate any decision boundary to arbitrary accuracy. Thus, such networks also provide universal non-linear discriminant functions. More generally, the capability of such networks to approximate general smooth functions allows them to model posterior probabilities of class membership. Since two layers of weights suffice to implement any arbitrary function, one would need special problem conditions (R.O. Duda and Stork 2001) or requirements to recommend the use of more than two layers. Furthermore, it is found empirically that networks with multiple hidden layers are more prone to getting caught in undesirable local minima. Astronomical data do not seem to require such level of complexity and therefore it is enough to use just a double weights layer, i.e a single hidden layer.

As it was just mentioned, it is also possible to train NNs in a Bayesian framework, which allows to find the more efficient among a population of NNs differing in the hyperparameters controlling the learning of the network (Bishop 1995), in the number of hidden nodes, etc. The most important hyperparameters being the so called α and β. α is related to the weights of the network and allows to estimate the relative importance of the different inputs and the selection of the input parameters which are more relevant to a given task (*Automatic Relevance Determination*; (Bishop 1995)). In fact, a larger value for a component of α implies a less meaningful corresponding weight. β is instead related to the variance of the noise (a smaller value corresponding to a larger value of the noise) and therefore to a lower reliability of the network. The implementation of a Bayesian framework requires several steps: initialization of weights and hyperparameters; training the network via a non linear optimization algorithm in order to minimize the total error function. Every few cycles of the algorithm, the hyperparameters are re-estimated and eventually the cycles are reiterated.

4.4 The data and the 'base of knowledge'

The Sloan Digital Sky Survey (hereafter SDSS) is an ongoing survey to image approximately π sterad of the sky in five photometric bands (u, g, r, i, z) and it is also the only survey so far to be complemented by spectroscopic data for $\sim 10^6$ objects (cf. the SDSS web pages at http://www.sdss.org/ for further details). The existence of such spectroscopic subset (hereafter SpS), together with the accurate characterization of biases and errors renders the SDSS an unique and ideal playing ground on which to train and test most photometric redshifts methods.

Several criteria may be adopted in extracting galaxy data from the SDSS database (Yasuda et al. 2001). We preferred, however, to adopt the standard SDSS criterion and use the GALAXY table membership. The data used in this work were therefore extracted from the SDSS catalogues. More in particular, the spectroscopic subsample (hereafter SpS), used for training and testing purposes, was extracted from the Data Release 4 (hereafter DR4; cf. (Adelman-McCarthy et al. 2006)) . While this work was in progress the Data Release 5 (DR5) was made publicly available. Thus, the photometric data used to produce the final catalogues were derived from the latter data. We wish to stress that this extension of the dataset was made possible by the fact that the properties of the DR5 are the same of the DR4 except for a wider sky coverage.

In this paper we made use of two different bases of knowledge extracted from the SpS of the DR4:

- The *General Galaxy Sample or GG sample*: composed of $445,933$ objects with $z < 0.5$ matching the following selection criteria: dereddened magnitude in r band, $r <$

21; $mode = 1$ which corresponds to primary objects only in the case of deblended sources.

- The *Luminous Red Galaxies sample or LRG sample*: composed of $97,475$ Luminous Red Galaxies candidates having spectroscopic redshift < 0.5. The SDSS spectroscopic survey (Eisenstein et al. 2001) was planned in order to favour the observation of the so called Luminous Red Galaxies or LRGs which are expected to represent a more homogeneous population of luminous elliptical galaxies which can be effectively used to trace the large scale structures (Eisenstein et al. 2001). We therefore extracted from the SDSS-DR4 all objects matching the above listed criteria and, furthermore, flagged as $primTarget =' TARGET\_GALAXY\_RED'$.

LRGs are of high cosmological relevance since they are both very luminous (and therefore allow to map the universe out to large distances), and clearly related to the cosmic structures (being preferably found in clusters). Furthermore, their spectral energy distribution is rather uniform, with a strong break at 4000 Å produced by the superposition of a large number of metal lines (Schneider et al. 1983, Eisenstein et al. 2003). LRGs are therefore an ideal target to test the validity of photometric redshift algorithms (see for instance: (Hamilton 1985), (Gladders and Yee 2000), (Eisenstein et al. 2001), (Willis et al. 2001) and (Padmanabhan et al. 2005)). The selection of LRG objects was performed using the same criteria extensively described in (Padmanabhan et al. 2005) and, given the rather lengthy procedure, we refer to that paper for a detailed description of the cuts introduced in the parameter space.

Since it is well known that photometric redshift estimates depend on the morphological type, age, metallicity, dust, etc. it has to be expected that if some morphological parameters are taken into account besides than magnitudes or colours alone, estimates of photometric redshifts should become more accurate. Such an effect was for instance found by (Tagliaferri et al. 2002, Vanzella et al. 2004).

In order to be conservative and also because it is not always simple to understand which parameters might carry relevant information, for each object we extracted from the SDSS database not only the photometric data but also the additional parameters listed in Table 4.1.

These parameters are of two types: those which we call 'features' (marked as F in Table 4.1), are parameters which potentially may carry some useful information capable to improve the accuracy of photometric redshifts, while those named 'labels' (marked as L) can be used to better understand the biases and the characteristics of the 'base of knowledge'.

For what magnitudes are concerned, and at a difference with other groups who used the $modelMag$, we used the so called dereddened magnitudes (*dered*), corrected for the

N	Parameter		F/L
	objID	SDSS identification code	–
	ra	right ascention (J2000)	–
	dec	declination (J2000)	–
1	petroR50$_i$	50 $\%$ of Petr. rad. in the i-th band, $i = u, g, r, i, z$	F
2	petroR90$_i$	90 $\%$ of Petr. rad. in the i-th band, $i = u, g, r, i, z$	F
3	dered$_i$	dered. mag. in the $i - th$ band, $i = u, g, r, i, z$	F
4	lnLDeV$_r$	log likelihood for De Vaucouleurs profile, r band	F
5	lnLExp$_r$	log likelihood for exponential profile, r band	F
6	lnLStar$_r$	log likelihood for PSF profile, r band	F
z	spectroscopic redshift		L
specClass	spectral classification index		L

Table 4.1: *List of the parameters extracted from the SDSS database and used in the experiments. Column 1: running number for features only. Column 2: SDSS code. Column 3: short explanation. Column 4: type of parameter, either feature (F) or label (L).*

best available estimate of the SDSS photometric zero-points:

$$\Delta(u, g, r, i, z) = (-0.042, 0.036, 0.015, 0.013, -0.002)$$

as reported in (Padmanabhan et al. 2005). It has to be stressed, however that such corrections are of little relevance for empirical methods since they affect equally all data sets.

Finally we must stress that we impose the condition that the objects had to be 'primary' ($mode = 1$) and detected in all five bands. The latter condition being required by the fact that all empirical methods suffer, one way or the other, from the presence of missing data and, to our knowledge, no clear cut method has been found to overcome this problem.

4.4.1 Features selection

In order to evaluate the significance of the additional features, our first set of experiments was performed along the same line as described in (Tagliaferri et al. 2002) using a Multi Layer Perceptron with 1 hidden layer and 24 neurons. In each experiment, the training, validation and test sets were constructed by randomly extracting from the overall dataset three subsets, respectively containing 60%, 20% and 20% of the total amount of galaxies.

On the sample, we run a total of $N + 1$ experiments. The first one was performed using all features, while the other N were performed taking away the $i - th$ feature with $i = 1, ..., N$. For each experiment, following (Csabai et al. 2003), we used the test set

Parameters	σ_3
all	0.0202
all but 1	0.0209
all but 2	0.0213
all but 4 & 5	0.214
all but 6	0.215
only magnitudes	0.0199

Table 4.2: *Results of the feature significance estimation. Column 1: features used. Features are numbered as in Table 4.1. Column 2: robust sigma of the residuals.*

to evaluate the robust variance σ_3 obtained by excluding all points whose dispersion is larger then 3σ (see § 4.7). The values are listed in Table 4.4.1.

As it can be seen, the most significant parameters are the magnitudes (or the colours). Other parameters affect only the third digit of the robust sigma and, due to the large increase in computing time during the training phase (which scales as N^2, where N is the number of input features) and to avoid loss of generality of higher redshifts, where additional features such as the Petrosian radii are either impossible to measure or affected by large errors, we preferred to drop all additional features and use only the magnitudes. The fact that on the contrary of what was found in (Vanzella et al. 2004) and (Tagliaferri et al. 2002) additional features do not play a significant role may be understood as a consequence of the fact that in this work the training set is much larger and more complete than in these earlier works and therefore the colour parameter space is (on average, but see below) better mapped.

4.5 The evaluation of photometric redshifts

One preliminary consideration: as it was first pointed out by (Connolly et al. 1995), when working in the near and intermediate redshift universe ($z < 1$), the most relevant broad band features are the Balmer break at 4000 Å and the shape of the continuum in the near UV. Near IR bands become relevant only at higher redshift and this is the main reason why we decided to concentrate on the near universe ($z < 0.5$), where the SDSS optical bands provide enough spectral coverage.

One additional reason comes from the redshift distribution of the objects in the SpS-DR4 shown in Fig. 4.2 (solid line). As it can be clearly seen, the histogram presents a clear discontinuity at $z \simeq 0.25$ (86% of the objects have $z < 0.25$ and only 14% are at a higher redshift) and in practice no objects are present for $z > 0.5$.

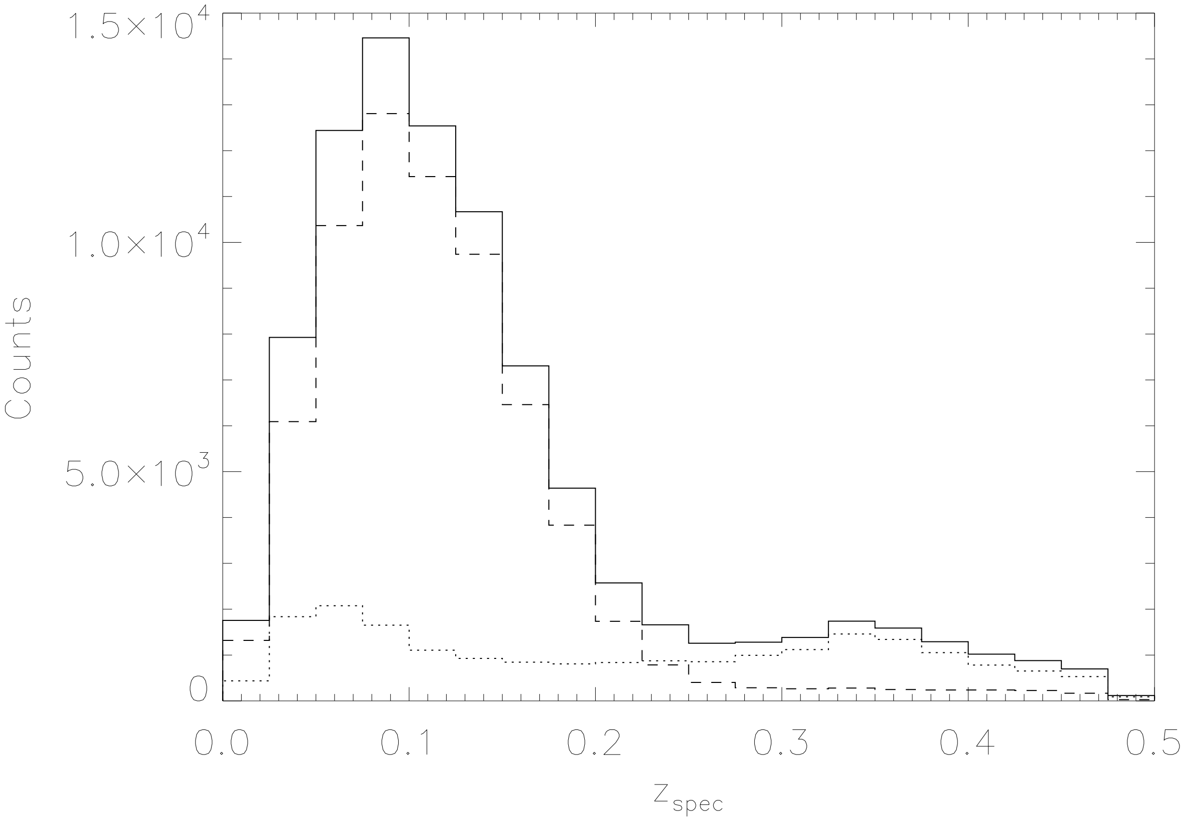


Figure 4.2: *Distribution of redshifts in the SpS sample. Solid line: GG sample. Dashed line: non-LRG sample. Dotted line: LRG sample (see text for details). Notice the sharp drop at $z \sim 0.25$.*

In Fig. 4.2 we also plot as dotted line the redshift distribution of the galaxies in the SpS data set which match the LRG photometric selection criteria. As it can be seen, within the tail at $z > 0.25$ only a very small fraction (11.4%) of the objects does not match the LRG selection criteria. In Fig. 4.3 we plot the redshift of objects belonging to the GG sample against their luminosity in r band: black dots represent those galaxies which have been *a posteriori* identified as LRG. As it is clearly seen, the overall distribution at redshift ≤ 0.25 drops dramatically at $r \sim 17.7$, due to the selection criteria of the spectroscopic SDSS survey. At higher redshift, namely $z > 0.25$, the galaxy distribution is dominated by LRGs with few contaminants and extends to much fainter luminosities. Nevertheless LRGs are systematically brighter then GG galaxies all over the redshift interval $z < 0.50$.

Such large inhomogeneity in the density and nature of training data, poses severe constraints on any empirical method since the different weights of samples extracted in the different redshift bins would lead either to over fitting in the densest region, or to

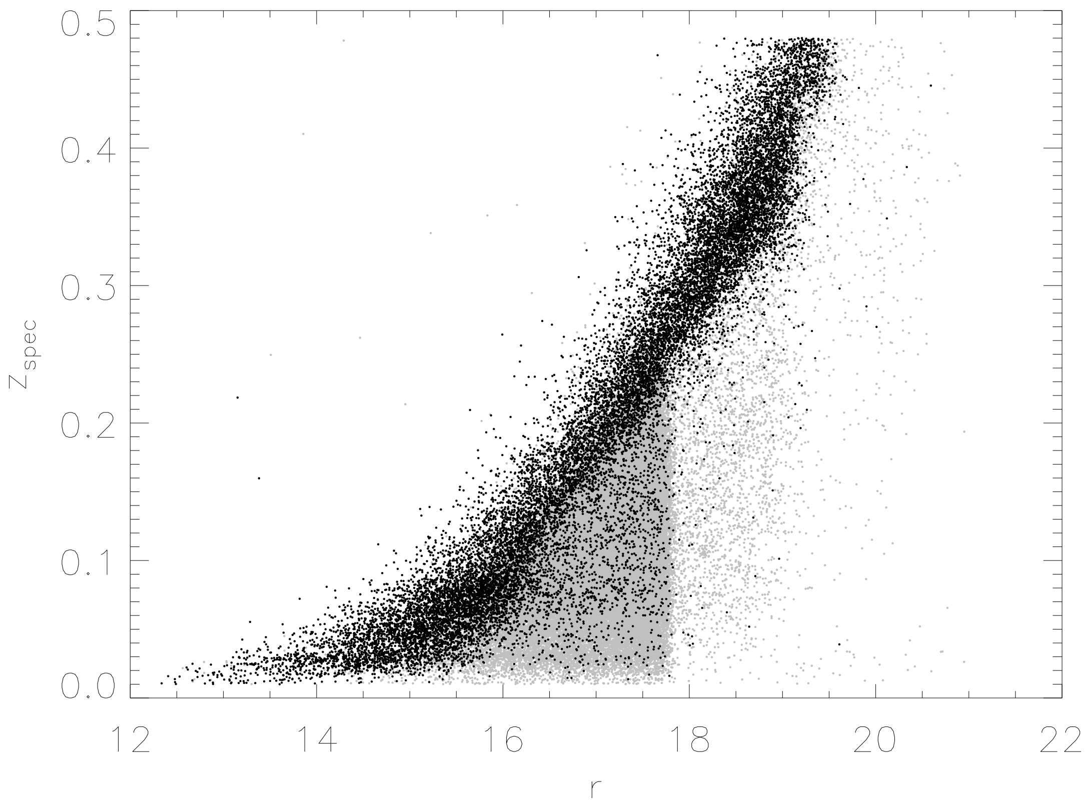


Figure 4.3: *Distribution of the objects in the GG sample versus the* r *magnitude (grey circles). We plot the LRG objects as black circles.*

the opposite effect in the less populated ones. Furthermore, the dominance of LRGs at $z > 0.25$ implies that in this redshift range the base of knowledge offers a poor coverage of the parameter space.

The first problem can be solved by taking into account the fact that, as shown in (Tagliaferri et al. 2002) and (Firth et al. 2003), NNs work properly even with scarcely populated training sets, and by building a training set which uniformly samples the parameter space or, in other words, which equally weights different clusters of points (notice that in this paper we use the word cluster in the statistical sense, id est to denote a statistically significant aggregation of points in the parameter space). In the present case the dominance of LRGs at high redshifts renders the parameter space heavily undersampled.

In fact, as it will be shown in Paper III, a more detailed analysis of the parameter space shows that at high redshift, the objects group into one very large structure containing more than 90% of the data points, plus several dozens of much smaller clusters .

	SDSS nearby	SDSS far
NN nearby	76498	1096
NN far	1135	11145

Table 4.3: *Confusion matrix for the "nearby-distant" test set.*

4.5.1 The nearby and intermediate redshifts samples

In order to tackle the above mentioned problems, we adopted a two steps approach: first we trained a network to recognize nearby (id est with $z < 0.25$) and distant ($z > 0.25$) objects, then we trained two separate networks to work in the two different redshift regimes. This approach ensures that the NNs achieve a good generalization capabilities in the nearby sample and leaves the biases mainly in the distant one. To perform the separation between nearby and distant objects, we extracted from the SDSS-4 SpS training, validation and test sets weighting, respectively, 60%, 20% and 20% of the total number of objects (449,370 galaxies). The resulting test set, therefore, consisted of 89,874 randomly extracted objects. Extensive testing (each experiment was done performing a separate random extraction of training, validation and test sets) on the network architecture lead to a MLP with 18 neurons in 1 hidden layer. This NN achieved the best performances after 110 epochs and the results are detailed, in the form of a confusion matrix, in Table 4.5.1.

As it can be seen, this first NN is capable to separate the two classes of objects with an efficiency of 97.52%, with slightly better performances in the nearby sample (98.59%) and slightly worse in the distant one (92.47%).

In Fig. 4.4 we plot against the redshift the percentage (calculated binning over the redshifts) for the objects in the test set which were misclassified (id est objects belonging to the nearby sample which were erroneously attributed to the distant one and viceversa). The distribution appears fairly constant from $z_{spec} \sim 0.05$ to $z_{spec} \sim 0.45$, while higher (but still negligible respect to the total number of objects in the sample) percentages are found at the extremes.

Notice that, when using photometric data alone, the absence of training data for $z > 0.5$, does not allow to evaluate the fraction of contaminants having $z > 0.5$ which are erroneously attributed to the distant sample. However, given the adopted cuts in magnitude, this number may be safely assumed to be negligible.

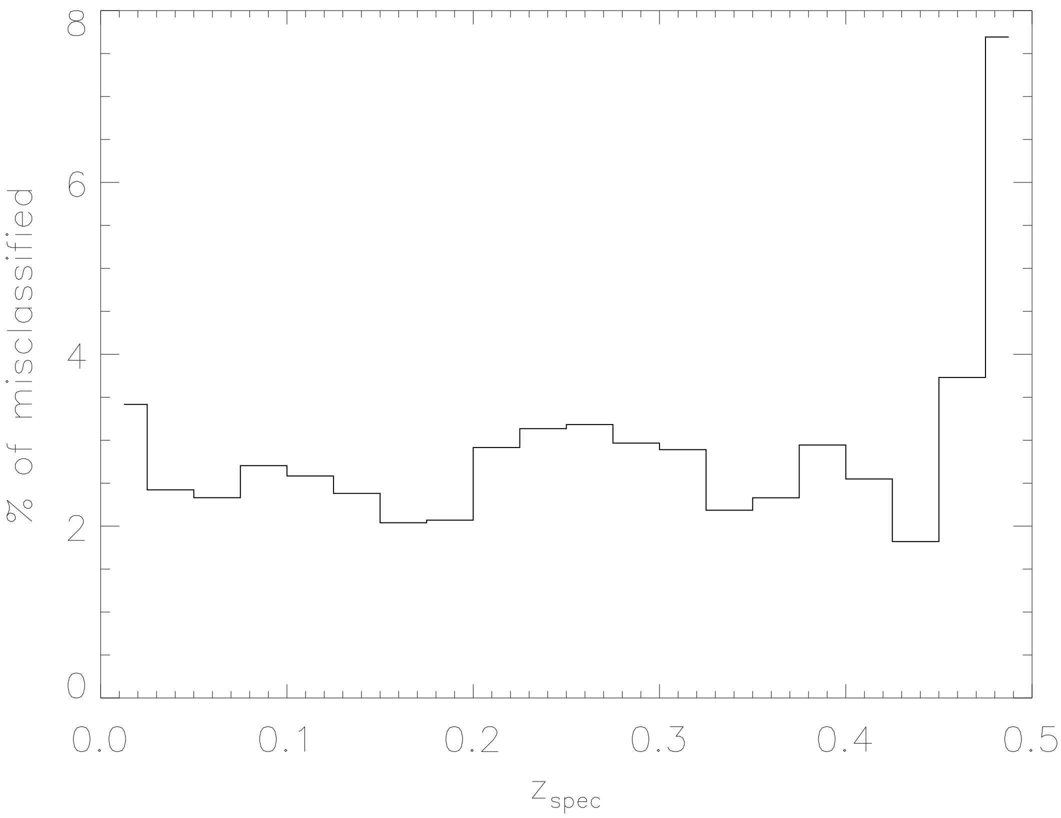


Figure 4.4: *Percentage distribution of misclassified objects of GG sample normalized to the total number of galaxies in each redshift bin.*

4.5.2 The photometric redshifts

Once the first network has separated the nearby and distant objects, we can proceed to the derivation of the photometric redshifts working separately in the two regimes. Since NNs are excellent at interpolating data but very poor in extrapolating them, in order to minimize the systematic errors at the extremes of the training redshift ranges we adopted the following procedure.

For the nearby sample we trained the network using objects with spectroscopic redshift in the range $[0.0, 0.27]$ and then considered the results to be reliable in the range $[0.01, 0.25]$. In the distant sample, instead, we trained the network over the range $[0.23, 0.50]$ and then considered the results to be reliable in the range $[0.25, 0.48]$.

In order to select the optimal NN architecture, extensive testing was made varying the network parameters and for each test the training, validation and test sets were randomly extracted from the SpS. The results of the Bayesian learning of the NNs were

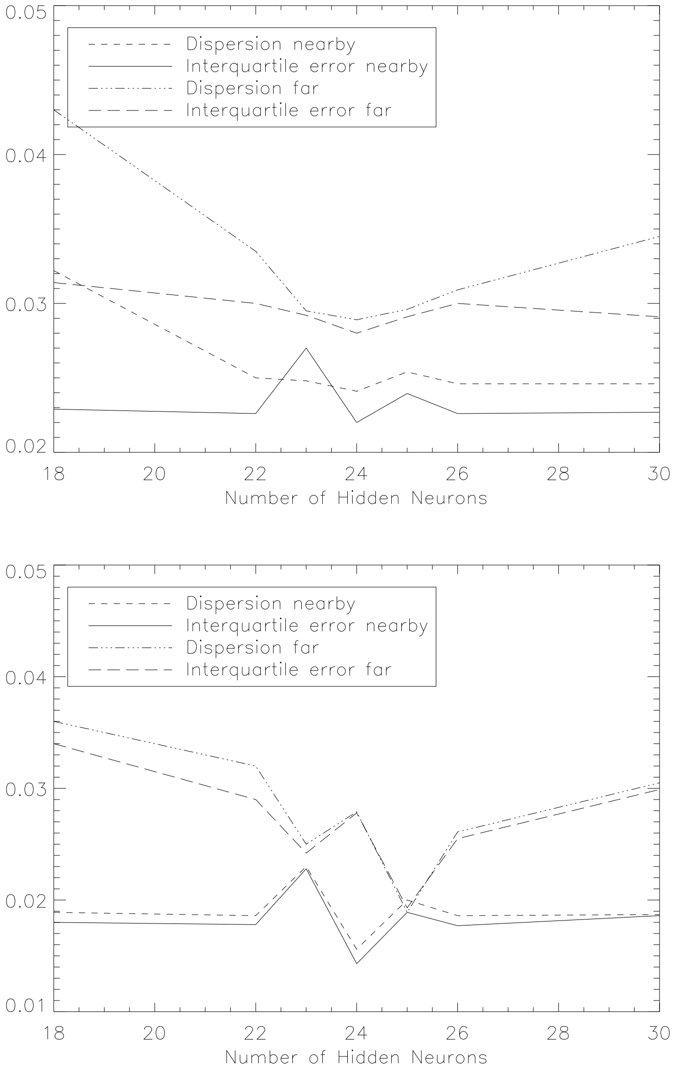


Figure 4.5: *Upper panel: GG sample, trend of the interquartile error and of the robust σ as a function of the number N of the neurons in the hidden layer. The nearby and distant samples are plotted separately. Lower panel: the same as above for the LRG sample.*

found to depend on the number of neurons in the hidden layer; for the GG (LRG) sample the performances were best when this parameter was set to 24 for the nearby sample and for the distant one (24 and 25 respectively for the LRG sample). In Fig. 4.5 we give the trends as a function of the number of hidden neurons, of the interquartile errors and robust dispersion obtained for the nearby and distant GG samples respectively.

For the GG sample, the best experiment, the robust variance turned out to be $\sigma_3 = 0.0208$ over the whole redshift range and 0.0197 and 0.0245 for the nearby and distant objects, respectively. For what the LRG sample is concerned, we obtained $\sigma_3 \simeq 0.0163$

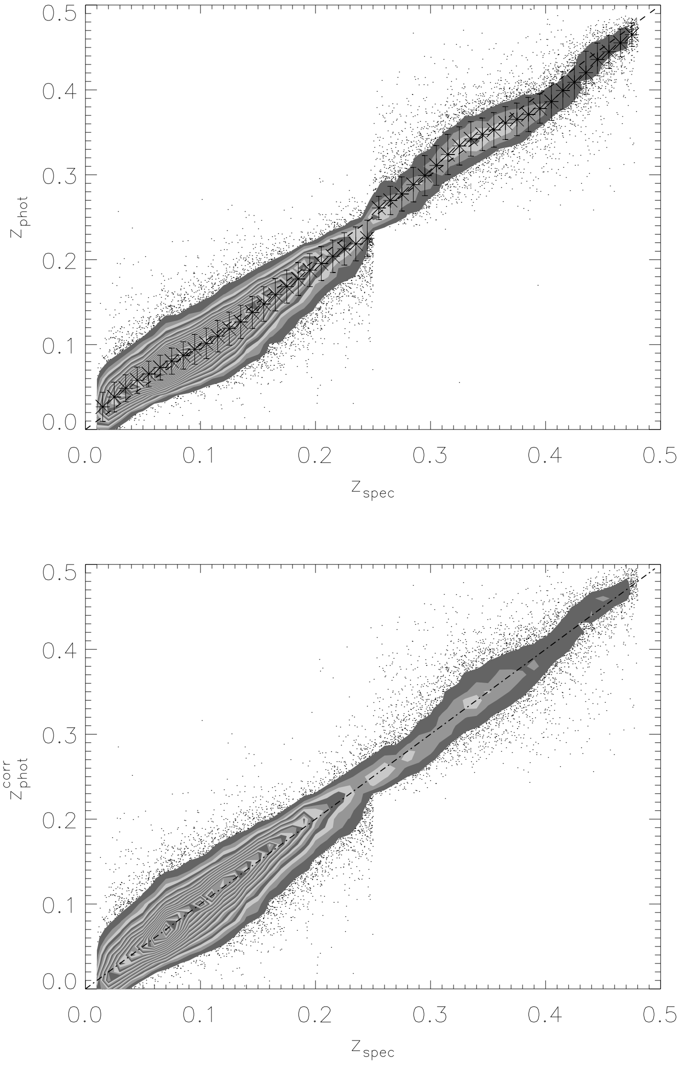


Figure 4.6: *Upper panel: photometric versus spectroscopic redshifts for the objects in the GG test set. The continuous lines are iso-density contours increasing with a step of 2 % of the maximum density. The crosses mark the average value of photometric redshifts in a specific spectroscopic redshift bin (see text), while the error bars give the robust variance σ_3. Lower panel: same as above after the correction for the systematic trends via interpolation (see text).*

over the whole range, and $\sigma_3 \simeq 0.0154$ and $\sigma_3 \simeq 0.0189$ for the nearby and distant samples, respectively. In the upper panels of Figs. 4.6 and 4.7 we plot the spectroscopic versus the photometric redshifts for the GG and the LRG samples, respectively. Due to the huge number of points which would make difficult to see the trends in the densest regions, we preferred to plot the data using isocontours (using a step of 0.02 times the maximum data point density).

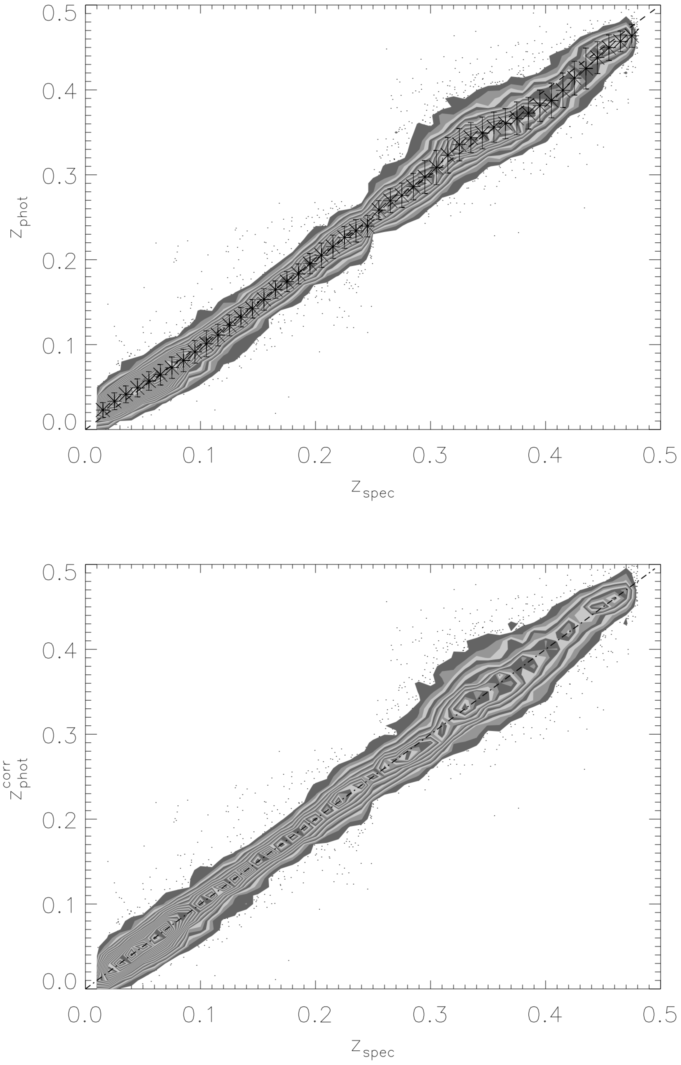


Figure 4.7: *Same as in Fig. 4.6 for the LRG sample.*

The mean value of the residuals are -0.0036 and -0.0029 for the GG and the LRG samples, respectively. These figures alone, however, are not very significant since systematic trends are clearly present in the data as it is shown in Fig. 4.8 and in Fig. 4.9, where we plot for each 0.05 redshift bin the average value of the photometric redshifts and the robust sigma of the residuals.

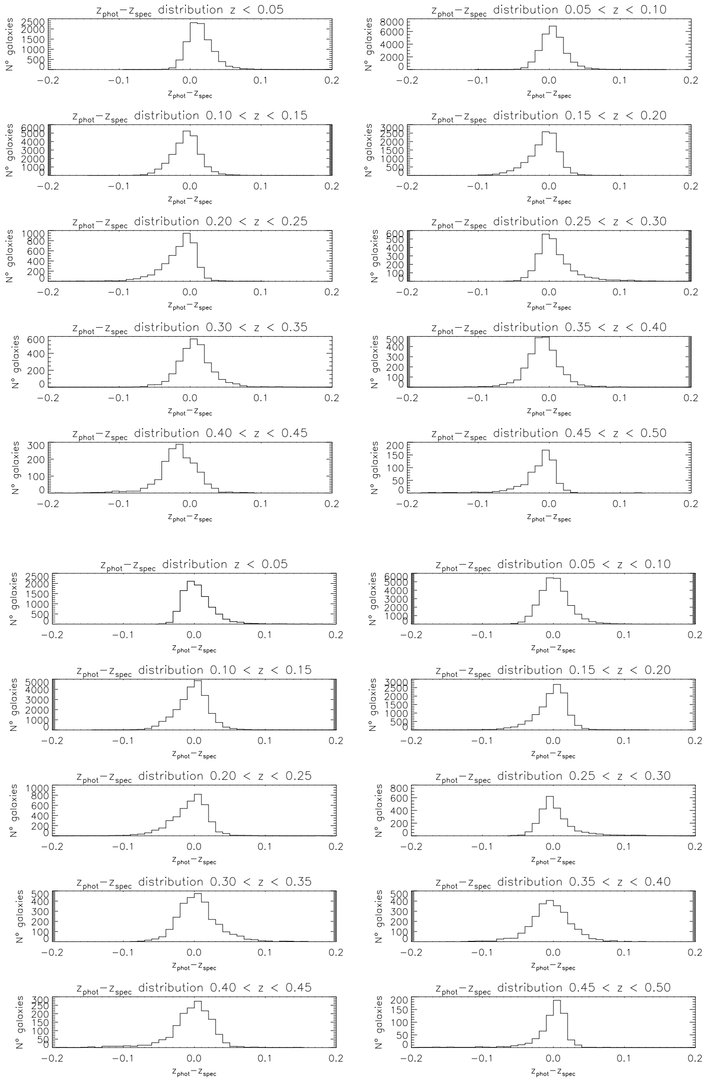


Figure 4.8: *Histograms of residuals for the GG sample in slices of redshift. Upper panels: before the correction. Lower panels: after the correction.*

4.6 Interpolative correction

The most significant deviations, as it could be expected (Connolly et al. 1995), are clearly visible in the nearby sample for $z < 0.1$ and in the distant sample at $z \sim 0.4$. The first feature is due to the fact that at low redshifts faint and nearby galaxies cannot be easily disentangled by luminous and more distant objects having the same colour. The second one is instead due to a degeneracy in the SDSS photometric system introduced by a small gap between the g and r bands. At $z \sim 0.4$, the Balmer break falls into this gap and its position becomes ill defined (Padmanabhan et al. 2005).

It needs to be stressed, however, that these trends represent a rather normal behavior

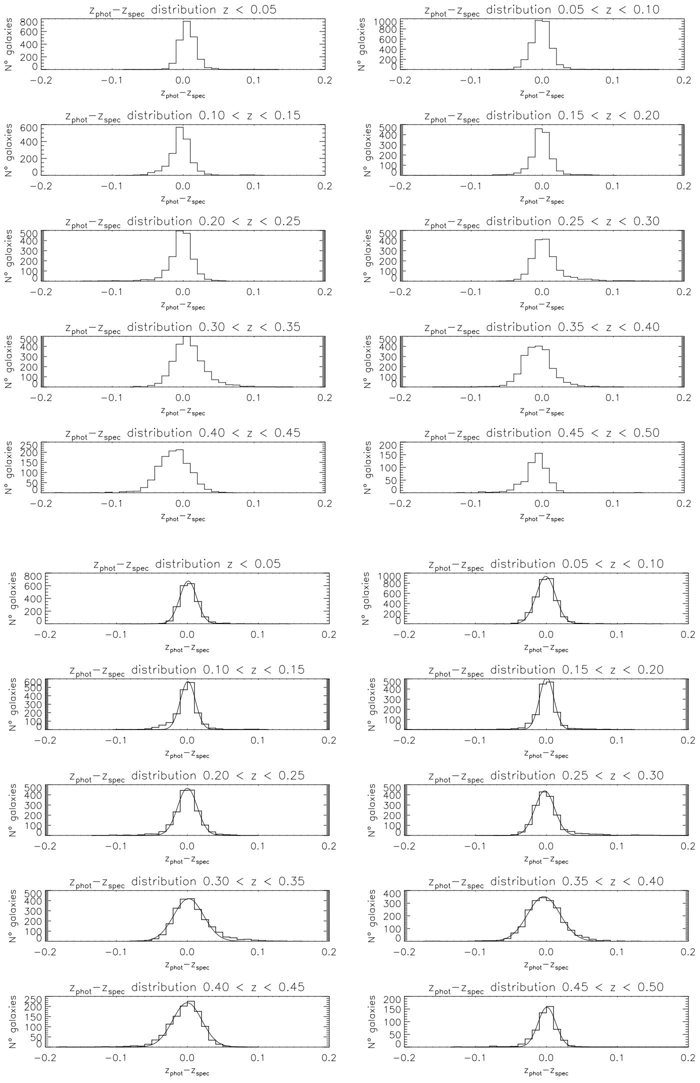


Figure 4.9: *Same as in previous figure but for the LRG sample.*

for empirical methods which has already been explicitly noted in (Tagliaferri et al. 2002) and (Vanzella et al. 2004) and is clearly visible (even when it is not explicitly mentioned) in almost all photometric redshifts data sets (Wadadekar 2005) available so far for the SDSS.

In order to minimize the effects of such systematic trends, but at the risk of a slight increase in the variance of the final catalogues we applied to both data sets an interpolative correction computed separately in the two redshift intervals. We used a χ^2 fitting to find, separately in each redshift regime, the polynomials which best fit the average points.These polynomials (of the fourth and fifth order, respectively) turned out to be. For the GG sample:

$$P_4 \quad [0.005, 1.570, -12.577, 78.948, -157.961] \tag{4.4}$$

$$P_5 \quad [12.15, -178.2, 1039.3, -2959.0, 4135.5, -2271.3] \tag{4.5}$$

and for the LRG sample:

$$P_4 \quad [0.011, 0.885, -1.820, 21.350, -53.159] \tag{4.6}$$

$$P_5 \quad [13.1, -192.5, 1123.3, -3207.2, 4504.5, -2491.6] \tag{4.7}$$

Thus, the correction to be applied is:

$$z_{phot}^{corr} = z_{phot} - (z_{phot}^{calc} - z_{spec}) \tag{4.8}$$

where $z_{phot}^{calc} = P_4(z_{spec})$ for near objects and $z_{phot}^{calc} = P_5(z_{spec})$ for the distant ones.
Obviously, when applying this method to objects for which we do not possess any spectroscopic estimate of redshift, it is impossible to perform the transformation Eq. (4.8) to correct NNs z_{phot} estimates for systematic trends and we are obliged to use an approximation. In other words, we replace the unknown z_{spec} with z_{phot} in the Eq. (4.8), obtaining the relation:

$$\tilde{z}_{phot}^{corr} = z_{phot} - (\tilde{z}_{phot}^{calc} - z_{phot}) \tag{4.9}$$

where $\tilde{z}_{phot}^{calc} = P_4(z_{phot})$ or $\tilde{z}_{phot}^{calc} = P_5(z_{phot})$ depending on the redshift range.
This is equivalent to assuming that the same NNs z_{phot} distribution represents, with good approximation, the underlying and unknown z_{spec} distribution. After this correction we obtain a robust variance $\sigma_3 = 0.0197$ for the GG sample and 0.0164 for the LRG samples, computed in both cases over the whole redshift range, and the resulting distributions for the two samples are shown in the lower panels of Figs. 4.6 and 4.7.

4.7 Discussion of the systematics and of the errors

As noticed by several authors (see for instance (Schneider et al. 2006, Padmanabhan et al. 2005)), while some tolerance can be accepted on the amplitude of the redshift error, much more critical are the uncertainties about the probability distribution of those errors. This aspect is crucial since (Padmanabhan et al. 2005) the observed redshift distribution is related to the true redshift distribution via a Fredholm equation which is ill defined and strongly dependent on the accuracy with which the noise can be modelled. In this respect, many recent studies on the impact of redshift uncertainties on various cosmological aspects are available: dark energy from supernovae studies and cluster number counts (Huterer et al. 2004); weak lensing (Bernstein and Jain 2004, Huterer

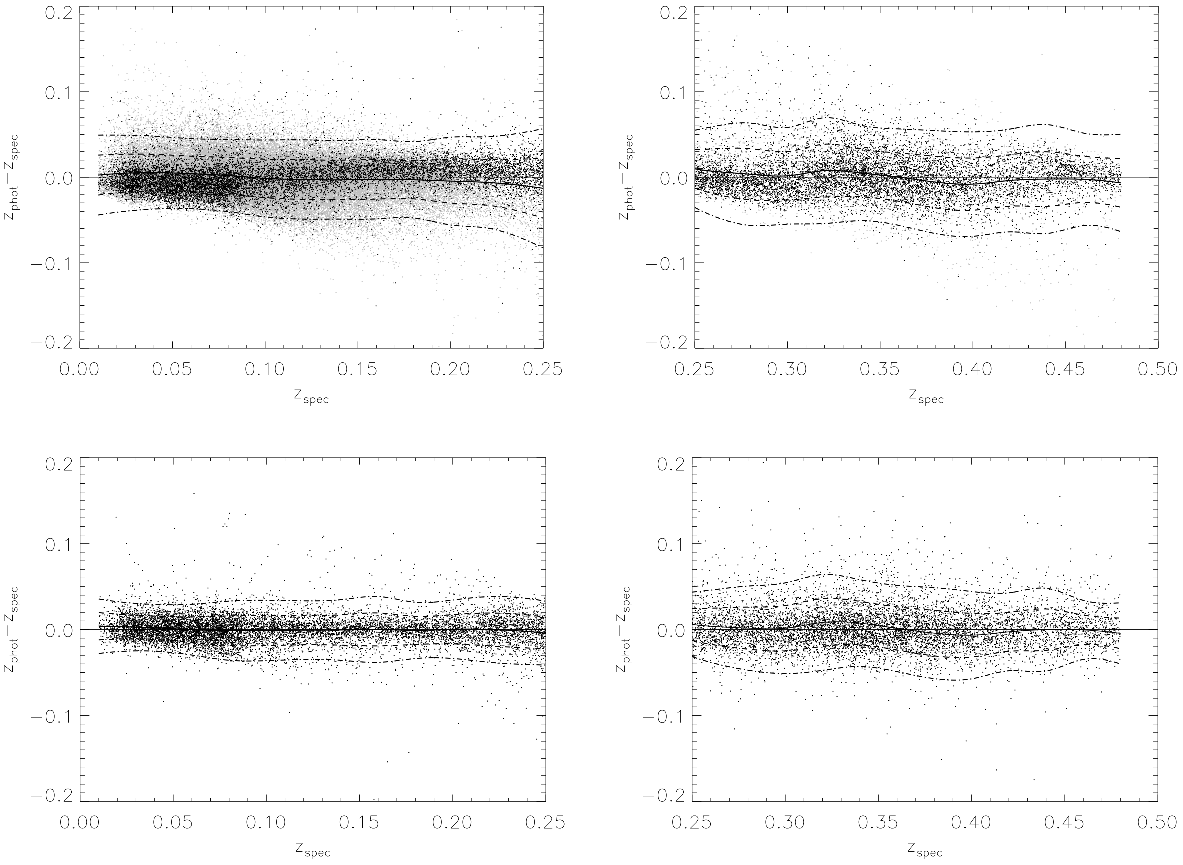


Figure 4.10: *Distribution of the residuals versus spectroscopic redshift after the correction for systematic trends. Upper panels: GG nearby and distant samples. Lower panels: LRG nearby and distant samples. The central line marks the average value of the residuals. The* $1\ \sigma$ *and* $2\ \sigma$ *confidence levels are also shown.*

et al. 2006, Ishak 2005, Ma et al. 2006); baryon oscillations (Zhan 2006, Zhan and Knox 2006). All these studies model the error distribution as Gaussian.

However, photometric redshift error distributions, due to spectral-type/redshift degeneracies, often have bimodal distributions, with one smaller peak separated from a larger peak by z of order unity (Benítez 2000, Fernández-Soto et al. 2002, Fernández-Soto et al. 2001), or more complex error distributions, as it can be seen in Fig. 4.8 within the GG sample.

In order to evaluate the robustness of the σ_r, several instances of the process were applied to different randomly selected training, validation and test sets and the robust sigma was found to vary only on the fourth significant digit. Small differences were found only in the identification of catastrophic objects, which however did not present any significant variation in their frequency.

The distribution of the residuals as a function of the spectroscopic redshift for the GG and LRG samples is shown in Fig. 4.10 separately for the near and distant objects. We have also studied the dependence of such residuals from the r-band luminosity of the galaxies in the two different magnitude ranges (cf. 4.5), ($r < 17.7$ and $r > 17.7$) and in the near and intermediate redshift bins, as shown in Fig. 4.12 and Fig. 4.13 for the GG and LRG galaxies respectively. Clear systematics are found only for near/faint and intermediate/luminous LRGs residuals: in the former case, the mean value of residual $z_{phot} - z_{spec}$ is systematically higher then 0, while in the latter it is constantly biased to negative values. Both cases can be addressed reminding that these galaxies occupy a poorly sampled volume in the parameter-space, and therefore the NN fails to reproduce the exact trend of spectroscopic redshift.

In Fig. 4.11 we show the same plot as in Fig. 4.6 but without isocontours and plotting as red dots the objects which "a posteriori" were labeled as members of the LRG sample. Interestingly enough, in the nearby sample the non-LRG and the LRG have robust variances of $\sigma_3 = 0.021$ and $\sigma_3 = 0.020$. Notice, however, that the LRG objects show a clear residual systematic trend. This behaviour can be explained by the fact that in the nearby sample the training set contains a large enough number of examples for both samples of objects and the network can therefore achieve a good generalization capability. In the distant sample the Non-LRG and LRG objects have instead robust variances given by: $\sigma_3 = 0.321$ and $\sigma_3 = 0.021$. Also in this case the observed behavior can be easily explained as due to the heavy bias toward the LRGs which form $\sim 88.5\%$ of the sample. It must be stressed that while the remaining 11.4% of the objects still constitute a fairly large sample of objects, the uneven distribution of the training data between the two groups of objects, overtrains the NN toward the LRG objects which therefore are much better traced.

This confirms what already found by several authors (Padmanabhan et al. 2005): the derivation of photometric redshifts requires besides than an accurate evaluation of the errors also the identification of an homogeneous sample of objects.

Objects not matching the 3σ criterion used for the robust variance are: 3.47% for the LRG sample and 3.18% for the GG sample. Before correction, the rejected points are $\simeq 2\%$ of the overall distribution for the GG sample and $\simeq 1.8\%$ for the LRG one.

As it was already mentioned, the SDSS data set has been extensively analyzed by several authors who have used different methods for photometric redshift determination. Unfortunately, a direct comparison is not always possible due to differences in either the data sets (different data releases have been used) or in the way errors were estimated. It must be stressed, however, that due to the fact that, above a minimum and reasonably low threshold, the NN performances are not affected much by the number of objects in the training set, the former factor can be safely neglected. So far, the most extensive works are those by (Csabai et al. 2003) and (Way and Srivastava 2006). In the

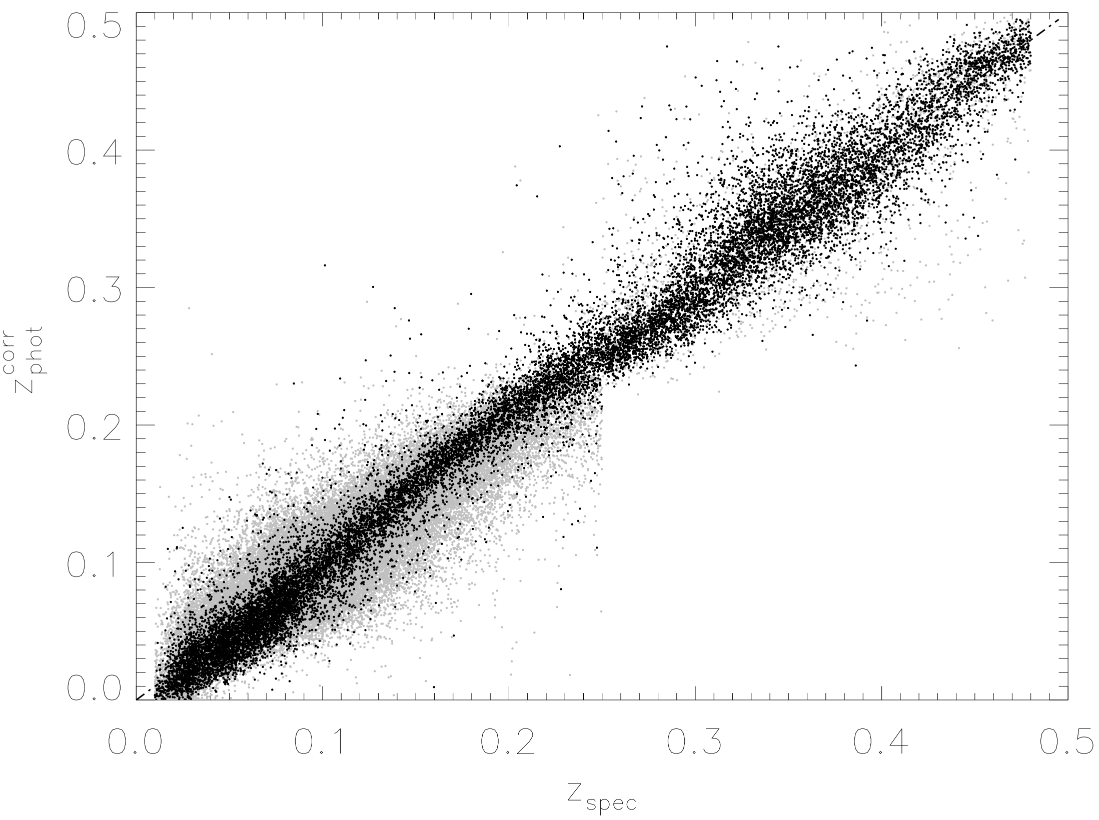


Figure 4.11: *Plot of the same data shown in the lower panel of Fig. 4.6, with the LRG and GG objects marked as black and grey circles respectively.*

former various methods were tested against the EDR data. With reference to their Table 3, and using the 'iterated' σ which almost coincides with the robust variance adopted here, we find that the best performances were obtained, among the SED fitting methods for the BC synthetic spectra ($\sigma_{it} \simeq 0.0621$ and $\sigma_{it} \simeq 0.0306$, for the GG and LRG samples respectively. This method, however leads to very clear systematic trends and to a large number of catastrophic outliers ($\sim 3.5\%$). Much better performances were attained by empirical methods and, in particular, by the interpolative one which leads to a $\sigma_{it} \simeq 0.0273$) with a fraction of catastrophic redshifts of only 2%. In (Way and Srivastava 2006) the authors made use of an Ensemble of NN (E) and Gaussian Process Regressions (GP). Their best results using the magnitudes only were 0.0205 and 0.0230 for the E and GP methods respectively, and at a difference with our method, their methods greatly benefits by the use of additional parameters such as the Petrosian radii, the concentration index and the shape parameter.

Two points are worth to be stressed. First of all, their selection criteria for the con-

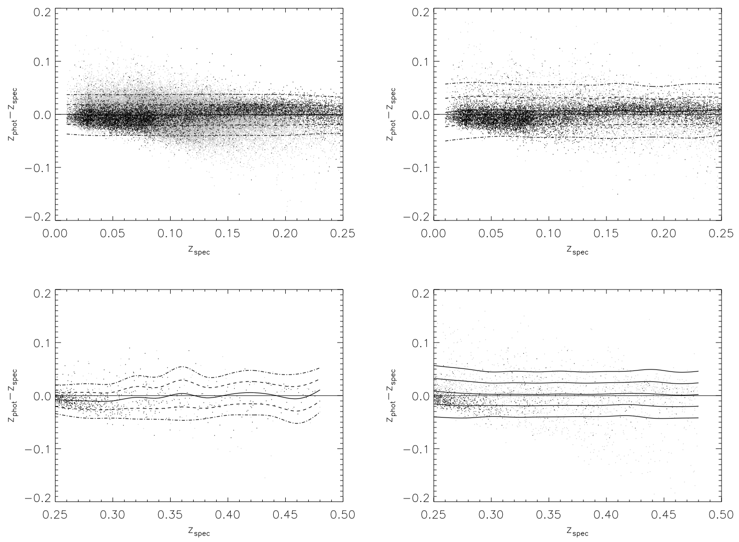


Figure 4.12: *Distribution of residuals for the GG sample divided in magnitude bins. Upper left panel: nearby sample, $r < 17.7$; upper right panel: nearby sample, $r > 17.7$; Lower left panel: distant sample, $r < 17.7$; lower right panel: $17.7 < r$.*

struction of the training set appear much more restrictive and it is not clear what performances could be achieved should such restriction be relaxed. Second, even though such 'ensemble' approach is very promising and is likely to be the most general one, it has to be stressed that the bagging procedure, used in (Way and Srivastava 2006) to combine the NNs, is known to be very effective only in those cases where the intrinsic variance of the adopted machine learning model is high. In this specific case, large number of training data and few input features, the NN result very stable and therefore other combining procedures, such as AdaBoost (Freund 1996), should be preferred (Dietterich 2002). This might also be the reason why when only the photometric parameters are used their method gives slightly worse performances than ours and instead leads to better results when the number of features is increased.

An additional machine learning approach, namely Support Vector Machines, was used by (Wadadekar 2005). In Table 4.4 we shortly summarize the main results of the above quoted papers.

4.7.1 Contamination by distant galaxies

The fact that our NN's are trained on a sample of galaxies with observed redshift $z_{spec} < 0.5$ introduces some contamination from objects which even though at $z > 0.5$ still have $r < 21$ and therefore match the photometric selection criteria.

The only possible way to avoid such an effect would be to use a knowledge base

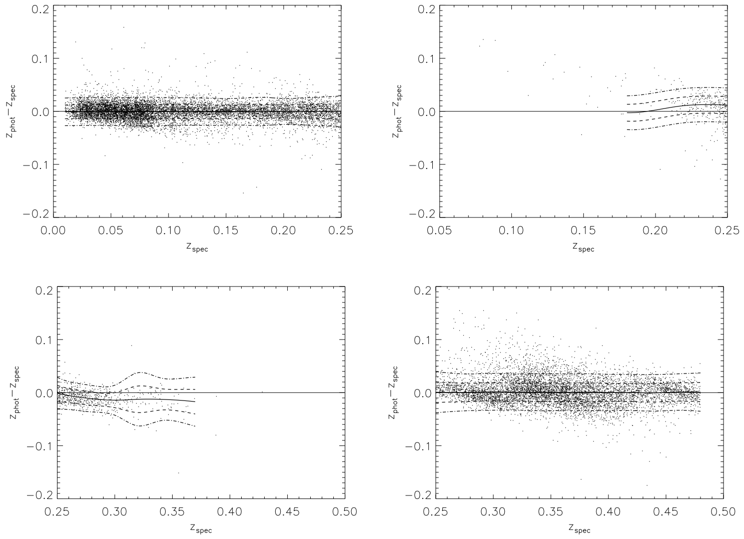


Figure 4.13: *Distribution of residuals for the LRG sample divided in magnitude bins. Upper left panel: nearby sample, $r < 17.7$; upper right panel: nearby sample, $17.7 < r$; Lower left panel: distant sample, $r < 17.7$; lower right panel: $17.7 < r$.*

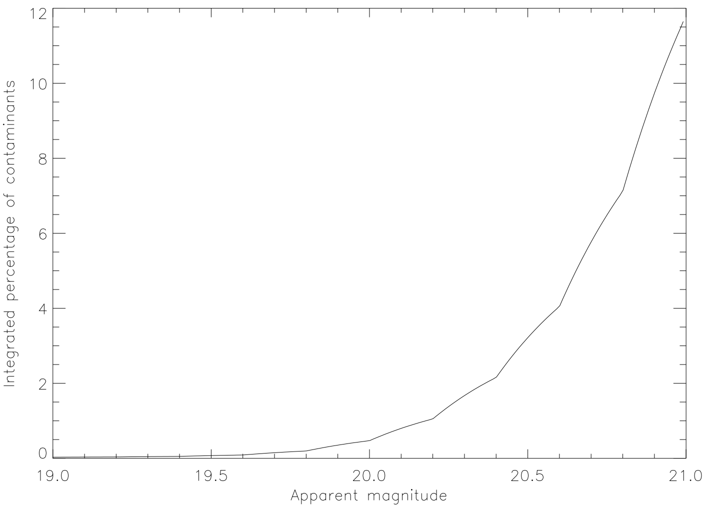


Figure 4.14: *Estimated distribution of contaminants as a function of the apparent r magnitude. The y-axis gives the expected fraction of objects at $z > 0.5$ which are erroneously evaluated by our procedure.*

covering in an uniform way all significant regions of the photometric parameter space down to the adopted magnitude limit. In the case of SDSS this is true for magnitudes brighter than 17.7 but is not true at fainter light levels where the only region uniformly covered by the spectroscopic subsample is that defined by the LRG selection criteria. A possible way out could be to extend the base of knowledge to fainter light level by

Reference	Method	Data	Δz	σ	Range
(Csabai et al. 2003)	SED fitting CWW	EDR		0.0621	
(Csabai et al. 2003)	SED fitting BC	EDR		0.0509	
(Csabai et al. 2003)	interpolative	EDR		0.0451	
(Csabai et al. 2003)	bayesian	EDR		0.0402	
(Csabai et al. 2003)	empirical, polynomial fit	EDR		0.0318	
(Csabai et al. 2003)	K-D tree	EDR		0.0254	
(Suchkov et al. 2005)	Class X	DR-2		0.0340	
(Way and Srivastava 2006)[a]	Gaussian Process	DR-3		0.0230	
(Way and Srivastava 2006)[a]	ensemble	DR-3		0.0205	
(Collister and Lahav 2004)	ANNz	EDR		0.0229	
(Wadadekar 2005)	SVM	DR-2		0.027	
(Wadadekar 2005)[a]	SVM	DR-2		0.024	
(Vanzella et al. 2004)	MLP ff	DR1	0.016	0.022	< 0.4
(Padmanabhan et al. 2005)	Template fitting and Hybrid	DR1-LRG	< 0.01	~ 0.035	< 0.55
this work before int.	MLP	DR5-GG	-0.0036	0.0197	0.01,0.25
this work before int.	MLP	DR5-GG	-0.0036	0.0245	0.25,0.48
this work after int.	MLP	DR5-GG			0.01,0.25
this work after int.	MLP	DR5-GG			0.25,0.48
this work before int.	MLP	DR5-LRG	-0.0029	0.0194	0.01,0.25
this work before int.	MLP	DR5-LRG	-0.0029	0.0205	0.25,0.48

Table 4.4: *Comparisons of various methods for the photometric redshift estimation applied to the SDSS data. Column 1: reference; Column 2: method (for the acronyms see text); Column 3: data set (EDR=Early Data Release; DR1 through DR5 the various SDSS data releases); Column 4: systematic offset; Column 5: standard deviation;s Column 6: redshift range over which the average error is estimated; ([a]): additional morphological and photometric parameters.*

including statistically significant and complete samples of spectroscopic redshifts from other and deeper surveys. The feasibility of using a third NN to classify (and eventually throw into a waste basket) objects having $z > 0.5$ is under study. At the moment, however, since we are interested in validating the method and in producing catalogues to be used for statistical applications, we shall estimate the number and the distribution in magnitude of such contaminants on statistical grounds only using the r-band luminosity function derived from SDSS data by (Blanton et al. 2003). This function, in fact, allows to derive for any given absolute magnitude the number of objects which even though at a redshift larger than 0.5 still match our apparent magnitude threshold and thus are misclassified. By integrating over the absolute magnitude and over the volume covered by the survey we obtain the curve in Fig. 4.14 which corresponds to a total number of contaminants of $\sim 3.74 \times 10^6$. It has to be noticed however that for magnitudes brighter than 20.5, the fraction of contaminants is less than 0.04 and drops below 0.01 for $r < 20$.

4.8 The catalogues

As mentioned above, the catalogues containing the photometric redshift parameters together with the parameters used for their derivation can be downloaded at the URL: http : //people.na.infn.it/ astroneural/SDSSredshifts.htm/. This data, for consistency with the SDSS survey, has been subdivided in several files, each corresponding to a different SDSS *stripe* of the observed sky. A *stripe* is defined by a line of constant survey latitude η, bounded on the north and south by the edges of the two strips (scans along a constant η value), and bounded on the east and west by lines of constant lambda. Because both strips and stripes are defined in "observed" space, they are rectangular areas which overlap as one approaches the poles (for more details see http : //www.sdss.org). The data for both GG and LRG samples have been extracted using the queries described in 4.4. The catalogues can be downloaded as 'FITS' files, containing the fundamental parameters used for redshift determination and the estimated photometric redshift for each individual source. In more details (in brackets SDSS database names of the parameters): unique SDSS identifier ('objID'), right ascension J2000 ('ra'), declination J2000 ('dec'), dereddened magnitudes ('dered_u', 'dered_g', 'dered_r', 'dered_i', 'dered_z'), the estimated value of photometric redshift before correction ('zphot') and after correction ('zphot_corr').

4.9 Conclusions

In the previous paragraphs we discussed a 'two steps' application of neural networks to the evaluation of photometric redshifts. Even though finely tailored on the characteristic of the SDSS, the method is completely general and can be easily applied to any other multiband data set provided that a suitable base of spectroscopic knowledge is available. As most other neural networks methods, several advantages are evident:

1. The NN can be easily re-trained if new data become available. Even though the training phase can be rather demanding in terms of computing time, once the NN has been trained, the derivation of redshifts is almost immediate (10^7 objects are processed on the fly on a normal laptop).

2. Even though it was not necessary in this specific case, all sorts of a priori knowledge can be taken into account.

On the other end, the method suffers of those limitations which are typical of all empirical methods based on interpolation. Most of all, the training set needs to ensure a complete and if possible uniform coverage of the parameter space.

Our method allowed to derive photometric redshifts for $z \lesssim 0.5$ with robust variances of $\sigma_3 = 0.0208$ for the GG sample ($\sigma_3 = 0.0197$ and $\sigma_3 = 0.0238$ for the nearby and distant sample respectively) and $\sigma_3 = 0.0164$ for the LRG sample ($\sigma_3 = 0.0160$ and $\sigma_3 = 0.0183$). This accuracy was reached adopting using a two-step approach allowing to build training sets which uniformly sample the parameter space of the overall population.

In the case of LRGs, the better accuracy and the close gaussianity of the residuals, are explained by the fact that this sample was selected based on the a priori assumption that they form a rather homogeneous population sharing the same SED. In other words, this result confirms what has long been known, id est the fact that when using empirical methods, it is crucial to define photometrically homogeneous populations of objects.

In the more general case it would be necessary to define photometrically homogeneous populations of objects in absence of a priori information and therefore relying only on the photometric data themselves. This task, as it has been shown for instance by (Suchkov et al. 2005, Bazell and Miller 2005) is a non trivial one, since the complexity of astronomical data and the level of degeneration is so high that most unsupervised clustering methods partition the photometric parameter space in far too many clusters, thus preventing the build-up a of a suitable base of knowledge. A possible way to solve this problem will be discussed in Paper III.

Chapter 5

Ultra High Energy Cosmic Rays anisotropy

Order and simplification are the first steps toward the mastery of a subject.

T. Mann

5.1 Introduction

Cosmic rays are high energy charged particles, originating in outer space, that travel at nearly the speed of light and strike the Earth from all directions. Most cosmic rays are the nuclei of atoms, ranging from the lightest to the heaviest elements in the periodic table. Cosmic rays also include high energy electrons, positrons, and other subatomic particles. The term ″cosmic rays″ usually refers to galactic cosmic rays, which originate in sources outside the solar system, distributed throughout our Milky Way galaxy. However, this term has also come to include other classes of energetic particles in space, including nuclei and electrons accelerated in association with energetic events on the Sun (called solar energetic particles), and particles accelerated in interplanetary space.
Cosmic rays were discovered in 1912 by Victor Hess, when he found that an electroscope discharged more rapidly as he ascended in a balloon. He rightly attributed this to a source of radiation entering the atmosphere from above, and in 1936 was awarded the Nobel prize for his discovery. For some time it was believed that the radiation was electromagnetic in nature (hence the name cosmic ″rays″), and some popular textbooks still incorrectly include cosmic rays as part of the electromagnetic spectrum. However, during the 30′s it was found that cosmic rays must be electrically charged because they are affected by the Earth′s magnetic field. From the 30′s to the 50′s, before man-made particle accelerators reached very high energies, cosmic rays served as a source of particles for high energy physics investigations, and led to the discovery of subatomic particles that included the positron and muon. Although these applications continue, since the dawn of the space age the main focus of cosmic ray research has been directed towards astrophysical investigations of where cosmic rays originate, how they get accelerated to such high velocities, what role they play in the evolution of the Galaxy, and what their composition tells us about matter from outside the solar system.

To measure cosmic rays directly, before they have been slowed down and broken up by the atmosphere, research is carried out by space or balloon borne instruments, using particle detectors similar to those used in nuclear and high energy physics experiments. The energy of cosmic rays is usually measured in units of MeV or GeV. Most galactic cosmic rays have energies between 100 MeV (corresponding to a velocity for protons of 43% of the speed of light) and 10 GeV (corresponding to 99.6% of the speed of light). Over a wide energy range the flux of particles with energy greater than E (measured in GeV), which is also called the integrated spectrum, is well represented by a broken power-law given approximately by the formula:

$$N(> E) = k(E + 1)^{-a} \tag{5.1}$$

where $k \sim 5000$ per m^2 per steradian per second and $a \sim 1.6$. Cosmic rays include essentially all of the elements in the periodic table; about 89% of the nuclei are hydrogen (protons), 10% helium, and about 1% heavier elements. The common heavier elements (such as carbon, oxygen, magnesium, silicon, and iron) are present in about the same relative abundances as in the solar system, but there are important differences in elemental and isotopic composition that provide information on the origin and history of galactic cosmic rays. For example there is a significant overabundance of the rare elements Li, Be, and B produced when heavier cosmic rays such as carbon, nitrogen, and oxygen fragment into lighter nuclei during collisions with the interstellar gas. The isotope $^{22}\mathrm{Ne}$ is also overabundant, showing that the nucleosynthesis of cosmic rays and solar system material have differed. Electrons constitute about 1% of galactic cosmic rays. At low energies, electrically charged cosmic rays are deflected by magnetic fields, and their arrival directions are randomized, making it impossible to tell from where they originated. However, cosmic rays in other regions of the Galaxy can be traced by the electromagnetic radiation they produce. Supernova remnants such as the Crab Nebula are known to be a source of cosmic rays from the radio synchrotron radiation emitted by cosmic ray electrons spiraling in the magnetic fields of the remnant. In addition, observations of high energy (10 - 1000) MeV gamma rays resulting from cosmic ray collisions with interstellar gas show that most cosmic rays are confined to the disk of the Galaxy, presumably by its magnetic field. Similar collisions of cosmic ray nuclei produce lighter nuclear fragments, including radioactive isotopes such as $^{10}\mathrm{Be}$, which has a half-life of 1.6 million years. The measured amount of $^{10}\mathrm{Be}$ in cosmic rays implies that, on average, cosmic rays spend about 10 million years in the Galaxy before escaping into inter-galactic space. When high energy cosmic rays undergo collisions with atoms of the upper atmosphere, they produce a cascade of ”secondary” particles that shower down through the atmosphere to the Earth’s surface. Secondary cosmic rays include pions (which quickly decay to produce muons, neutrinos and gamma rays), as well as electrons and positrons produced by muon decay and gamma ray interactions with atmospheric atoms. The number of particles reaching the Earth's surface is related to the energy of the cosmic ray that struck the upper atmosphere. Cosmic rays with energies beyond 10^{14} eV are studied with large "air shower" arrays of detectors distributed over many square kilometers that sample the particles produced. Most secondary cosmic rays reaching the Earth's surface are muons, with an average intensity of about 100 per m^2 per second.

5.2 Composition of cosmic rays

When a CR enters the Earth atmosphere it collides with a nucleus of an air atom, producing a roughly conical cascade of billions of elementary particles which reaches the ground in the form of a giant saucer travelling at nearly the speed of light. Unfortunately, because of the highly indirect method of measurement, extracting precise information from the Extensive Air Showers (hereafter EASs) has proved to be exceedingly difficult. The most fundamental problem is that the first generations of particles in the cascade are subject to large inherent fluctuations and consequently this limits the event-by-event energy resolution of the experiments. In addition, the center-of-mass energy of the first few cascade steps is well beyond any reached in collider experiments. Therefore, one needs to rely on hadronic interaction models that attempt to extrapolate, using different mixtures of theory and phenomenology, our understanding of particle physics. At present, the different approaches used to model the underlying physics of $p\bar{p}$ collisions show clear differences in multiplicity predictions which increase with rising energy (Anchordoqui et al. 1999, Ranft 1999, Alvarez-Muñiz et al. 2004). Therefore, distinguishing between a proton and a nucleus shower is extremely difficult at the highest energies (Knapp et al. 2003). Photon and hadron primaries can be distinguished by comparing the rate of vertical to inclined showers, a technique which exploits the attenuation of the electromagnetic shower component for large slant depths. Comparing the predicted rate to the rate observed by Haverah Park[1] for showers in the range $60^\circ < \alpha < 80^\circ$, the authors in (Ave et al. 2000) conclude that above 10^{19} eV, less than 48% of the primary CRs can be photons and above 4×10^{19} eV less than 50% can be photons. The longitudinal development has a well defined maximum, usually referred to as X_{max}, which increases with primary energy as more cascade generations are required

[1] Haverah Park was a research site operated by the Physics Department of the University of Leeds specifically for the detection of air showers produced by cosmic rays, and for 20 years it was home to one of the largest extensive air shower arrays in the world with an area of 12 square kilometres. The array was made up of water Cerenkov detectors housed in wooden huts and operated until 1987 when it was switched off. During its lifetime many thousands of extensive air showers were recorded and studied in this facility.

to degrade the secondary particle energies. Evaluating X_{max} is a fundamental part of many of the composition studies done by detecting air showers. For showers of a given total energy, heavier nuclei have smaller X_{max} because the shower is already subdivided into A nucleons when it enters the atmosphere. Specifically, the way the average depth of maximum $\langle X_{max}\rangle$ changes with energy depends on the primary composition and particle interactions according to

$$\langle X_{max}\rangle = D_e \log_e \left(\frac{E}{E_0}\right) \tag{5.2}$$

where D_e is the so-called elongation rate and E_0 is a characteristic energy that depends on the primary composition (Linsley and Watson 1981). Therefore, since $\langle X_{max}\rangle$ and D_e can be determined directly from the longitudinal shower profiles measured with a fluorescence detector, E_0 and thus the composition, can be extracted after estimating E from the total fluorescence yield. Indeed, the parameter often measured is D_{10}, the rate of change of $\langle X_{max}\rangle$ per decade of energy. Another important observable which can be related to primary energy and chemical composition is the total number of muons N_μ reaching ground level. For vertical proton showers, numerical simulations (Alvarez-Muñiz and Halzen 2001) indicate that the muon production is related to the energy of the primary via 5.3:

$$E = 1.64 \times 10^{18} \left(\frac{N_\mu^p}{10^7}\right)^{1.073} eV \tag{5.3}$$

Thus, modelling a shower produced by a nucleus with energy E_A as the collection of A proton showers, each with energy A^{-1} of the nucleus energy, leads to $N_\mu^A \propto A\left(\frac{E_A}{A}\right)^{0.93}$. Consequently, one expects a CR nucleus to produce about $A^{0.07}$ more muons than a proton. This implies that an iron nucleus produces a shower with around 30% more muons than a proton shower of the same energy. The analysis of the elongation rate and the spread in X_{max} at a given energy reported by the Flys Eye Collaboration suggests a change from an iron dominated composition at $10^{17.5}$ eV to a proton dominated composition near 10^{19} eV (Bird et al. 1993). Such behaviour of D_e is in agreement with an earlier analysis from Haverah Park (Walker and Watson 1982). However, the variation of the density of muons with energy reported by the Akeno Collaboration favours a composition that remains mixed over the 10^{18} - 10^{19} eV decade.

More recently, Flys Eye data were reanalyzed considering not only proton and iron components but a larger number of atomic mass hypotheses. Additionally, they adopted a different hadronic model that shifts the prediction of X_{max} for primary protons of 10^{18} eV from 730 g cm^{-2} to 751 g cm^{-2}. The difference, although apparently small, has a significant effect on the mass composition inferred from the data. The study indicates that at the highest energies ($10^{18.5}$ - 10^{19} eV and somewhat above) there is a significant

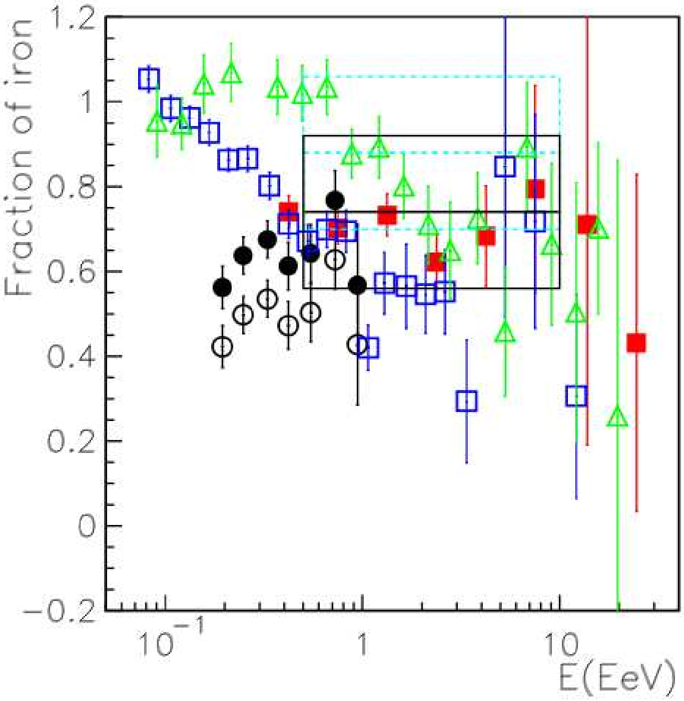


Figure 5.1: *Predicted fraction of iron nuclei in the CR beam at the top of the atmosphere from various experiments: Flys Eye (open triangles), AGASA A100 (filled squares), AGASA A1 (open squares). Open and filled circles represent different hadronic interaction event generators. The solid (dashed) line rectangle indicates the mean composition with the corresponding error estimated using the Volcano Ranch data and QGSJET98; the systematic shift in the fraction of iron induced by the hadronic event generator is 14%.*

fraction of primaries with charge greater than unity. This result is more in accord with the conclusions of the Akeno group than those of the Flys Eye group. Very recently, the Volcano Ranch data was re-analyzed taking into account a bi-modal proton-iron model (Dova et al. 2003). The best fit gives a mixture with $75 \pm 5\%$ of iron, with corresponding percentage of protons. A summary of the different bi-modal analyses is shown in figure 5.1. Within statistical errors and systematic uncertainties introduced by hadronic interaction models, the data seem to indicate that iron is the dominant component of CRs between $\sim 10^{17}$ eV and $\sim 10^{19}$ eV. Nonetheless, in view of the low statistics at the end of the spectrum and the wide variety of uncertainties in these experiments, one may conservatively say that this is not a closed issue.

5.2.1 Distribution of arrival directions of CR

The distribution of arrival directions is perhaps the most helpful observable in yielding clues about the CR origin. On the one hand, if cosmic rays cluster within a small angular region (Hayashida et al. 1996) or show directional alignment with powerful compact objects (Farrar and Biermann 1998), one might be able to associate them with isolated sources in the sky. On the other hand, if the distribution of arrival directions exhibits a large-scale anisotropy, this could indicate whether or not certain classes of sources are associated with large-scale structures (such as the Galactic plane or the Galactic halo). Cosmic ray air shower detectors which experience stable operation over a period of a year or more can have a uniform exposure in right ascension, α. A traditional technique to search for large-scale anisotropies is then to fit the right ascension distribution of events to a sine wave with period $\frac{2\pi}{m}$ (m^{th} harmonic) to determine the components (x, y) of the Rayleigh vector:

$$x = \frac{2}{N}\sum_{i=1}^{N}\cos(m\alpha_i) \tag{5.4}$$

$$y = \frac{2}{N}\sum_{i=1}^{N}\sin(m\alpha_i) \tag{5.5}$$

The m_{th} harmonic amplitude of N measurements of α_i is given by the Rayleigh vector length $R = (x^2 + y^2)^{\frac{1}{2}}$. The expected length of such a vector for values randomly sampled from a uniform phase distribution is $R_0 = \frac{2}{\sqrt{N}}$. The chance probability of obtaining an amplitude with length larger than that measured is $p(\geq R) = e^{-k_0}$, where $k_0 = \frac{R^2}{R_0^2}$. To give a specific example, a vector of length $k_0 \geq 6.6$ would be required to claim an observation whose probability of arising from random fluctuation was 0.0013 (a 3σ result). For example, AGASA has revealed a correlation of the arrival direction of the cosmic rays to the Galactic Plane (GP) at the 4σ level (AGASA Collaboration: N. Hayashida et al. 1998). The energy bin width which gives the maximum k_0-value corresponds to the region $10^{17.9}$ eV $10^{18.3}$ eV where $k_0 = 11.1$, yielding a chance probability of $p(\geq R^{AGASA}_{E\sim E_{eV}}) \approx 1.5 \times 10^5$. The GP excess, which is roughly 4% of the diffuse flux, is mostly concentrated in the direction of the Cygnus region, with a second spot towards the Galactic Center (GC). Evidence at the 3.2σ level for GP enhancement in a similar energy range has also been reported by the HiRes Collaboration (Bird et al. 1999). The existence of a point-like excess in the direction of the GC has been confirmed via independent analysis of data collected with SUGAR, thus providing a remarkable level of agreement among experiments which employ a variety of different techniques. At lower energies ($\sim$ PeV), the Rayleigh analysis shows no evidence of anisotropy (Antoni et al. 2004). Hence, the excess from the GP is very suggestive of neutrons as candidate primaries, because the directional signal requires relatively-stable neutral primaries, and time-dilated neutrons can reach the Earth from typical Galactic distances when the neutron energy exceeds 10^{18} eV. Arguably, if the Galactic messengers are neutrons, then those with energies below 10^{18} eV will decay in flight, providing a flux of cosmic antineutrinos above 1 TeV that should be observable at kilometer-scale neutrino telescopes. A measurement of the ν-flux will supply a strong confirmation of the GP neutron hypothesis. For the ultra high energy ($\sim 10^{19.6}$ eV) regime, all experiments to date have reported $k_0 \ll 6.6, \forall m < 5$ (Edge et al. 1978). This does not imply an isotropic distribution, but it merely means that available data are too sparse to claim a statistically significant measurement of anisotropy. In other words, there may exist anisotropies at a level too low to discern given existing statistics (Evans et al. 2003). The right harmonic analyses are completely blind to intensity variations which depend only on declination δ. Combining anisotropy searches in α over a range of declinations could dilute the results, since significant but out of phase Rayleigh vectors from different declination bands can cancel each other out. Moreover, the analysis methods that consider distributions in one celestial coordinate, while integrating away the second, have proved to be potentially misleading (Wdowczyk and Wolfendale 1979). An unambiguous interpretation of anisotropy data requires two ingredients: exposure to the full celestial sphere and analysis in terms of both celestial coordinates. In this direction, a recent study of the angular power spectrum of the distribution of arrival directions of CRs with energy $\geq 10^{19.6}$ eV, as seen by the AGASA and SUGAR experiments, shows no departures from either homogeneity or isotropy on an angular scale greater than $10°$. Finally, the recently analyzed HiRes data is also statistically consistent with an isotropic distribution (High Resolution Fly'S Eye Collaboration et al. 2004). The simplest interpretation of the existing data is that, beyond the ankle, a new population of extragalactic CRs emerges to dominate the more steeply falling Galactic population. Moreover, there are two extreme explanations for the near observed isotropy beyond $10^{19.6}$ eV: one is to argue a cosmological origin for these events, and the other is that we have nearby sources (say, within the Local Supercluster) with a tangled magnetic field in the Galaxy, and beyond, which bends the particle orbits, camouflaging the exact location of the sources. Although there seems to be a remarkable agreement among experiment on predictions about isotropy on large scale structure, this is certainly not the case when considering the two-point correlation function on a small angular scale. The analyses carried out by AGASA Collaboration seem to indicate that the pairing of events on the celestial sky could be occurring at higher than chance coincidence (Hayashida et al. 2000). Specifically, when showers with separation angle less than the angular resolution $\theta_{min} = 2.5°$ are paired up, AGASA finds five doublets and one triplet among the 58 events reported with mean energy above $10^{19.6}$ eV. The probability of observing these clusters by chance coincidence under an isotropic distribution was quoted as smaller than 1%. A third independent analysis, using the GoldbergWeiler formalism, confirmed the result reported by AGASA Collaboration and further showed that the chance probability is extremely sensitive to the angular binning. The world data set has also been studied: six doublets and two triplets out of 92 events with energies $\geq 10^{19.6}$ eV were found, with the chance probability being less than 1% in the restricted region within $\pm 10^{\circ}$ of the super-Galactic plane. The angular two-point correlation function of a combined data sample of AGASA ($E > 4.8 \times 10^{19}$ eV) and Yakutsk ($E > 2.4 \times 10^{19}$ eV) was analyzed (Tinyakov and Tkachev 2001). For a uniform distribution of sources, the probability of chance clustering is reported to be as small as 4×10^{6}. Far from confirming what seemed a fascinating discovery, the recent analysis reported by the HiRes Collaboration showed that the data is consistent with no small-scale anisotropy among the highest energy events. The discovery of such clusters would be a tremendous breakthrough for the field, but the case for them is not yet proven. To calculate a meaningful statistical significance in such an analysis, it is important to define the search procedure a priori in order to ensure it is not inadvertently devised especially to suit the particular data set after having studied it. In the analyses carried out by AGASA Collaboration, for instance, the angular bin size was not defined ahead of time. Very recently, with the aim to avoid accidental bias on the number of trials performed in selecting the angular bin, the original claim of AGASA Collaboration was re-examined considering only the events observed after the claim. This study showed that the evidence for clustering in the AGASA data set is weaker than was previously claimed, and consistent with the null hypothesis of isotropically distributed arrival directions. Summing up, the clustering on small angular scale at the upper end of the spectrum remains an open question, and the increase in statistics and improved resolution attainable with Pierre Auger Observatory was expected to solve the issue (see paragraph 5.3.1)

5.2.2 UHECRs propagation: the GZK-cutoff

In this section the relevant interactions that CRs suffer on their journey to Earth are described. Ever since the discovery of the cosmic microwave background standard physics implies there would be a cutoff in the observed CR-spectrum. In the mid-60s Greisen, Zatsepin, and Kuzmin (GZK) (Greisen 1966) pointed out that this photonic molasses makes the universe opaque to protons of sufficiently high energy, i.e., protons with energies beyond the photopion production threshold:

$$E^{th}_{p\gamma CMB} = \frac{m_\pi(m_p + \frac{m_\pi}{2})}{\epsilon_{CMB}} \approx 6.8 \times 10^{19} \left(\frac{\epsilon_{CMB}}{10^{-3} eV}\right)^{-1} \text{eV} \tag{5.6}$$

where $m_p(m_\pi)$ denotes the proton (pion) mass and $\epsilon_{CMB} \sim 10^{-3}$ eV is a typical CMB photon energy. After pion production, the proton (or perhaps, instead, a neutron) emerges with at least 50% of the incoming energy. This implies that the nucleon energy changes by an e-folding after a propagation distance $\leq (\sigma_{p\gamma} n_\gamma y)^1 \sim 15$ Mpc. Here, $n_\gamma \approx 410 cm^{-3}$ is the number density of the CMB photons, $\sigma_{p\gamma} \geq 0.1$ mb is the photopion production cross section, and y is the average energy fraction (in the laboratory system) lost by a nucleon per interaction. Energy losses due to pair production become relevant below $\sim 10^{19}$ eV. For heavy nuclei, the giant dipole resonance can be excited at similar total energies and hence, for example, iron nuclei do not survive fragmentation over comparable distances. Additionally, the survival probability for extremely high energy ($\approx 10^{20}$ eV) γ-rays (propagating on magnetic fields $\gg 10^{11}$ G) to a distance d, $p(> d) \approx \exp \frac{d}{6.6 \mathrm{Mpc}}$, becomes less than 10^4 after traversing a distance of $50 \mathrm{Mpc}^3$. In recent years, several studies on the propagation of CRs (including both analytical analyses and numerical simulations) have been carried out. A summary of the UHECR attenuation lengths for the above mentioned processes (as derived in these analyses) is given in figure 5.2.

It is easily seen that our horizon shrinks dramatically for energies $\geq 10^{20}$ eV. Therefore, if UHECRs originate at cosmological distances, the net effect of their interactions would yield a pile-up of particles around 4 - 5×10^{19} eV with the spectrum droping sharply thereafter. As one can infer from picture 5.2, the subtleties of the spectral shape depend on the nature of the primary species, yielding some ambiguity in the precise definition of the GZK cutoff.

5.3 UHECRs anisotropy: history

While the majority of cosmic rays detected at Earth is quite isotropic as discussed in the previous paragraph, an observable anisotropy appears to occur (Watson et al. 1996) at the approach of energies of order 10^{18} eV. This effect is quite natural, because the gyroradius of protons of that energy in a 1 G magnetic field is about 1 kpc, i.e., of O(10^{10}) larger than the scale of the random component of the Galactic magnetic field (GMF). Although the typical values of the regular component of the GMF are expected to be from 4 to 6 times higher for most of our Galaxy (Beck et al. 1996), the cosmic-ray spectrum extends to energies higher by at least 2.5 orders of magnitude. UHECRs are observed by the air showers that they initiate when the primary UHECR particles interact in the atmosphere. Ground-based air shower arrays register the arrival of a large shower by the coincidental arrival of a large number of charged particles in distant particle detectors. Since the flux of UHECRs is extremely low ($0.5 \ \mathrm{km}^{-2} \mathrm{yr}^{-1} \mathrm{sr}^{-1}$ above 10^{19} eV), the shower arrays designed for their detection are by necessity very sparse, and the detection yields only the energy and the arrival direction of the primary UHECR. The nature

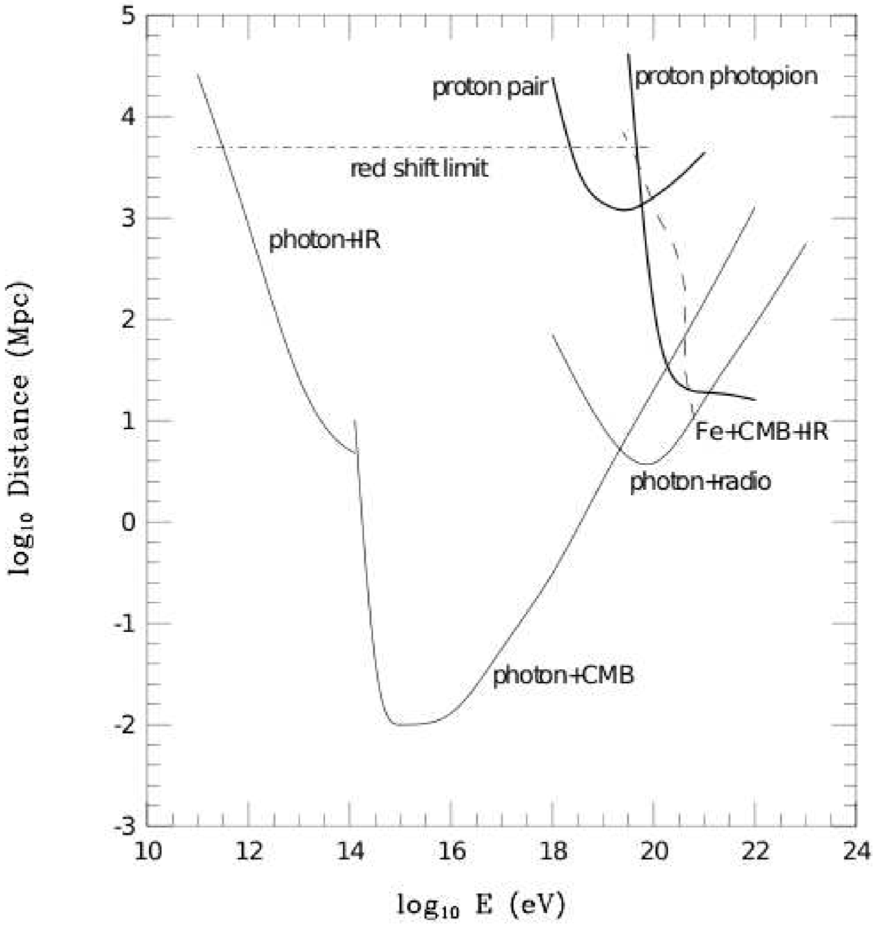


Figure 5.2: *Attenuation length of γs, ps, and ^{56}Fes in various background radiations as a function of energy. The 3 lowest and left-most thin solid curves refer to γ-rays, showing the attenuation by infra-red, microwave, and radio backgrounds. The upper, right-most thick solid curves refer to propagation of protons in the CMB, showing separately the effect of pair production and photo-pion production. The dashed-dotted line indicates the adiabatic fractional energy loss at the present cosmological epoch. The dashed curve illustrates the attenuation of iron nuclei.*

of that particle could be only derived from a large enough statistical sample. Although the origin, and even the nature, of cosmic rays with such high energy is yet unknown, their value for revealing the general structure of the GMF (in the natural assumption that they are mostly ionized hydrogen atoms (protons)) and whether their is a connection between their observed anisotropy and the large scale distribution of candidate extragalactic sources, is remarkable. Protons of energy above 10^{19} eV do not suffer significant energy loss on galactic scale lengths. They are only deflected in magnetic fields extending on scales larger than 1 kpc. If their arrival distribution on their entry into our Galaxy is not strictly isotropic, the GMF would have a focusing (or defocusing) effect on their arrival distribution at Earth, which will reflect the general GMF structure rather than the local magnetic field in the vicinity of the solar system. The propagation of UHECRs in the GMF has been studied previously, although mostly in terms of the general anisotropy at energies above 10^{17} eV related to the relative strength of the field strengths at large and small scales.

5.3.1 Recent developments

The Pierre Auger Observatory records cosmic ray showers through an array of 1,600 particle detectors placed 1.5 kilometres (about one mile) apart in a grid spread across 3,000 square kilometres located on the large plain known as the Pampa Amarilla in western Argentina. Twenty four specially designed telescopes detect the emission of fluorescence light from the air shower produced by very energetic cosmic rays impinging on the upper layers of the atmosphere. The combination of particle detectors and fluorescence telescopes provides a powerful instrument for the search of UHECRs and the study of the distribution of their arrival directions. The Pierre Auger Collaboration has very recently announced that high energy cosmic rays arrival directions can be mostly correlated with the positions of a particular type of galaxy, known as Active Galactic Nuclei (The Pierre Auger Collaboration 2007). The Auger observatory has been the first experiment to be able to determine the direction of these energetic particles with sufficient accuracy, and also the first to gather a sufficiently large number of such events to study them meaningfully. Prior to this achievements, researchers could only theorize where these high energy particles were being produced, while now there is an evidence for UHECRs delivered by active galactic nuclei and a new way of probing these extreme environments. Active Galactic Nuclei (AGN) are found at the hearts of some galaxies and are thought to be powered by supermassive black holes that are consuming large amounts of matter (for details on AGNs, see chapter 11). AGNs have long been considered sites where high-energy particle production might take place, even if the exact mechanism of how AGNs can accelerate particles to energies 100 million times higher than the most powerful particle accelerator is still a doubtful piece of information. This result announces a new window to the nearby universe and the beginning of cosmic-ray astronomy, since as more and more data will be collected, it will become possible to look at individual galaxies from a completely new viewpoint. Only the rare highest-energy cosmic rays can be linked to their sources with sufficient precision. Scientists working in the Auger collaboration so far have recorded 81 cosmic rays with energy above 4×10^{19} eV, representing the largest sample of UHECRs with energy above 40 EeV recorded by any observatory. At these ultra-high energies, the uncertainty in the direction from

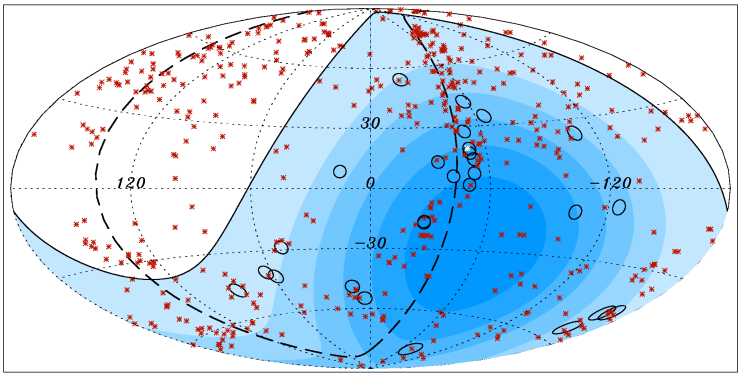


Figure 5.3: *The celestial sphere in galactic coordinates (Aitoff projection) showing the arrival directions of the 27 highest energy cosmic rays detected by Auger, shown as circles of radius* $3.1°$*, with energies greater than* 5.7×10^{18} *eV. The positions of 472 AGN within 75 Mpc are shown as red asterisks. The blue region defines the field of view of Auger; deeper blue indicates larger exposure. The solid curve marks the boundary of the field of view, where the zenith angle equals* $60°$*. The closest AGN, Centaurus A, is marked as a white star. Two of the 27 cosmic rays have arrival directions within* $3°$ *of this galaxy. The supergalactic plane, delineating a region where a large numbers of nearby galaxies, including AGNs, are concentrated, is indicated by the dashed curve.*

which the cosmic ray arrived is only a few degrees, allowing to determine the location of the particles cosmic source. It has been showed that the 27 highest-energy events (with energy $E < 5.7 \times 10^{19}$ eV) arrival directions are not homogeneously spread across the sky, but the clustering of these events correlated well with the known locations of 381 AGNs (see figure 5.3).

Cosmic rays with energy higher than about 6.0×10^{19} eV lose energy in collisions with the cosmic microwave background, but cosmic rays from nearby sources are less likely to lose energy because of scattering on their relatively short trip to Earth. A further results of this work was that most of the 27 events with energy $E > 5.7 \times 10^{19}$ eV came from locations in the sky including the nearest AGNs. The Auger collaboration is developing plans for a second, larger installation in located in the northern hemisphere (Colorado, USA) to extend the coverage to the entire sky while substantially increasing the number of high-energy events recorded.

5.3.2 Our work

The paper presented in the next chapter 6 has been the first to investigate the possibility that the measurements of the Pierre Auger Observatory could discriminate between an isotropic and a clustered distribution of UHECRs arrival directions, and determined the minimal number of events needed to do so. In particular, a large scale structure model for the UHECRs origin which evaluates the expected anisotropy in the UHECR arrival distribution starting from a given astronomical catalogue of the local universe has been studied. The method has been applied to the IRAS PSCz catalogue, deriving the minimum statistics needed to significantly reject the hypothesis that UHECRs trace the baryonic distribution in the universe, in particular providing a forecast for the Auger experiment. An obvious development of this work would be the application of the same methodology to other catalogues of extra-galactic sources, both using photometric redshifts measurements that may increase the number of galaxies used to reconstruct the clustering of matter in the nearby Universe of at least one order of magnitude, and selected samples of specific types of galaxies (like AGNs, Seyfert galaxies, etc.) in order to cast light on the physical mechanisms producing the energetic cosmic rays. This paper has stimulated the discussion on this subject and has triggered a deep interest toward a possible cosmic rays astronomy, whose feasibility in the last days has been confirmed by the results reported in paragraph 5.3.1.

Published as: Cuoco, A., D'Abrusco, R., Longo, G., Miele, G., Serpico, P. D. – *"The footprint of large scale cosmic structure on the ultra-high energy cosmic ray distribution."* 2006, Journal of Cosmology and Astroparticle Physics, 01(2006)009

Chapter 6

Application I: UHECRs and Large Scale Structure of the Universe

Abstract

Current experiments collecting high statistics in ultra-high energy cosmic rays (UHECRs) are opening a new window on the universe. In this work we discuss a large scale structure model for the UHECR origin which evaluates the expected anisotropy in the UHECR arrival distribution starting from a given astronomical catalogue of the local universe. The model takes into account the main selection effects in the catalogue and the UHECR propagation effects. By applying this method to the IRAS PSCz catalogue, we derive the minimum statistics needed to significantly reject the hypothesis that UHECRs trace the baryonic distribution in the universe, in particular providing a forecast for the Auger experiment.

6.1 Introduction

Almost a century after the discovery of cosmic rays, a satisfactory explanation of their origin is still lacking, the main difficulties being the poor understanding of the astrophysical engines and the loss of directional information due to the bending of their trajectories in the galactic (GMF) and extragalactic magnetic field (EGMF).

More in detail, given the few-μG intensity of regular and turbulent GMF, a diffusive confinement of cosmic rays of galactic origin is expected up to rigidity $\mathcal{R} \equiv p\,c/Z\,e \simeq$ few $\times 10^{17}$ V, p being the cosmic ray momentum, Z its charge in units of the positron one, and c the speed of light. Still at $\mathcal{R} \simeq$ few $\times 10^{18}$ V cosmic rays are strongly deflected, and no directional information can be extracted. Around $\mathcal{R} \sim 10^{19}$ V the regime of relatively small deflections in the GMF starts. The transition decades $\mathcal{R} \simeq 10^{17}$–$10^{19}$ V, though not yet useful for "directional" astronomy, may still show a rich phenomenology (drifts, scintillation, lensing) which is an interesting research topic of its own (Roulet 2004).

At energies above a few $\times 10^{19}$ eV, which we will refer to as the ultra-high energy (UHE) regime, protons propagating in the Galaxy retain most of their initial direction. Provided that EGMF is negligible, UHE protons will therefore allow to probe into the nature and properties of their cosmic sources. However, due to quite steep

CR power spectrum, UHECRs are extremely rare (a few particles km^{-2} century^{-1}) and their detection calls for the prolonged use of instruments with huge collecting areas. One further constraint arises from an effect first pointed out by Greisen, Zatsepin and Kuzmin (Greisen 1966, Zatsepin and Kuz'min 1966) and since then known as GZK effect: at energies $E \gtrsim 5\times10^{19}$ eV the opacity of the interstellar space to protons drastically increases due to the photo-meson interaction process $p + \gamma_{\rm CMB} \to \pi^{0(+)} + p(n)$ which takes place on cosmic microwave background (CMB) photons. In other words, unless the sources are located within a sphere with radius of $\mathcal{O}(100)$ Mpc, the proton flux at $E \gtrsim 5\times10^{19}$ eV should be greatly suppressed. However, due to the very limited statistics available in the UHE regime (cf. Volcano Ranch (Linsley 1963), SUGAR (Winn et al. 1986), Haverah Park (Lawrence et al. 1991, Ave et al. 2000), Fly's Eye (Bird et al. 1993, Bird et al. 1994, Bird et al. 1995), Yakutsk (Efimov and et al. 1991) AGASA (Takeda et al. 1998), HiRes (Abbasi 2004, Abu-Zayyad et al. 2005), and, very recently, also Auger (The Pierre Auger Collaboration 2005a)), the experimental detection of the GZK effect has not yet been firmly established. It has to be stressed that the theoretical tools available to probe this extremely interesting part of the CRs spectrum are still largely inadequate: both the modelling and the data interpretation impose either strong assumptions based on little experimental evidence or the extrapolation by orders of magnitudes of available knowledge. For instance, the structure and magnitude of the EGMF are poorly known. Only recently, magnetic fields were included in simulations of large scale structures (LSS) (Dolag et al. 2004, Sigl et al. 2004). Qualitatively the simulations agree in finding that EGMFs are mainly localized in galaxy clusters and filaments, while voids should contain only primordial fields. However, the conclusions of Refs. (Dolag et al. 2004) and (Sigl et al. 2004) are quantitatively rather different and it is at present unclear whether deflections in extragalactic magnetic fields will prevent astronomy even with UHE protons or not. Another large source of uncertainty is our ignorance on the chemical composition of UHECRs, mainly due to the need to extrapolate for decades in energy the models of hadronic interactions. They are an essential input for the Monte Carlo simulations used in the analysis and reconstruction of UHECRs showers, but the predictions of such simulations differ appreciably already in the *knee* region (around 10^{15} eV), even when high quality data and deconvolution techniques are used (Antoni et al. 2005). Future accelerator measurements of hadronic cross sections in higher energy ranges will ameliorate the situation, but this will take several years at least. From now on, therefore, we shall work under the assumptions that UHE astronomy is possible, namely: i) proton primaries, for which $e\mathcal{R} = E$; ii) EGMF negligibly small; iii) extragalactic astrophysical sources are responsible for UHECR acceleration. Now the question arises: might one support this scenario using the directional information in UHECRs? A possibility favoring these hypothesis is that relatively few, powerful nearby sources are responsible for the UHECRs, and the small scale clustering observed by

AGASA (Takeda et al. 1999) may be a hint in this direction. However, the above quoted clustering has not yet been confirmed by other experiments with comparable or larger statistics (Abbasi 2004), and probably a final answer will come when the Pierre Auger Observatory (Cronin 1992) will have collected enough data. Independently on the observation of small-scale clustering, one could still look for large scale anisotropies in the data, eventually correlating with some known configuration of astrophysical source candidates. In this context, the most natural scenario to be tested is that UHECRs correlate with the luminous matter in the "local" universe. This is particularly expected for candidates like gamma ray bursts (hosted more likely in star formation regions) or colliding galaxies, but it is also a sufficiently generic hypothesis to deserve an interest of its own.

Aims of this work are: i) to describe a method to evaluate the expected anisotropy in the UHECR sky starting from a given catalogue of the local universe, taking into account the selection function, the blind regions as well as the energy-loss effects; ii) to assess the minimum statistics needed to significantly reject the null hypothesis, in particular providing a forecast for the Auger experiment. Previous attempts to address a similar issue can be found in (Waxman et al. 1997, Evans et al. 2002, Smialkowski et al. 2002, Singh et al. 2004). Later in the paper we will come back to a comparison with their approaches and results.

The catalogue we use is IRAS PSCz (Saunders et al. 2000). This has several limitations, mainly due to its intrinsic incompleteness, but it is good enough to illustrate the main features of the issue, while still providing some meaningful information. This work has to be intended as mainly methodological. An extension to the much more detailed 2MASS (Jarrett et al. 2000, Jarrett 2004) and SDSS (York et al. 2000, Adelman-McCarthy et al. 2006) galaxy catalogues is presently investigated.

The paper is structured as follows: the catalogue and the related issues are discussed in Section 6.2. In Section 6.3 we describe the technique used for our analysis. The results are discussed in Section 6.4, where we compare our findings with those obtained in previous works. In Section 6.5 we give a brief overview on ongoing research and experimental activities, and draw our conclusions. Throughout the paper we work in natural units $\hbar = k_B = c = 1$, though the numerical values are quoted in the physically most suitable units.

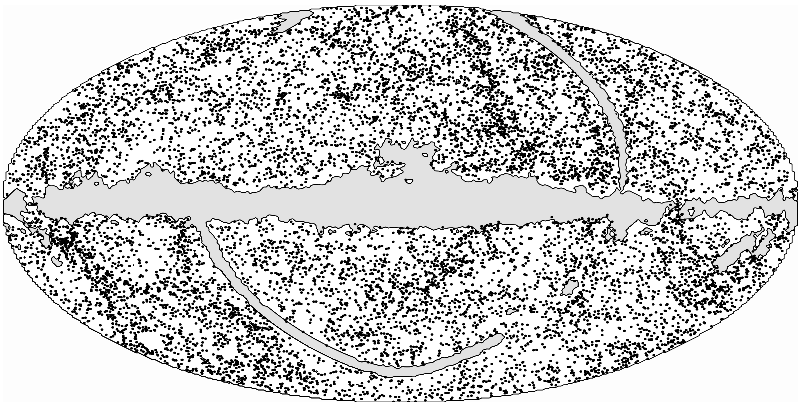


Figure 6.1: *PSCz catalogue source distribution and related mask in galactic coordinates.*

6.2 Astronomical Data

6.2.1 The Catalogue

Two properties are required to make a galaxy catalogue suitable for the type of analysis discussed here. First, a great sky coverage is critical for comparing the predictions with the fraction of sky observed by the UHECR experiments (the Auger experiment is observing all the Southern hemisphere and part of the Northern one). Second, the energy-loss effect in UHECR propagation requires a knowledge of the redshifts for at least a fair subsample of the galaxies in the catalogue. Selection effects both in fluxes and in redshifts play a crucial role in understanding the final outcome of the simulations.

Unfortunately, in practical terms this two requirements turn out to be almost complementary and no available catalogue matches both needs simultaneously. A fair compromise is offered by the IRAS PSCz catalogue (Saunders et al. 2000) which contains about 15 000 galaxies and related redshifts with a well understood completeness function down to $z \sim 0.1$ —i.e. down to a redshift which is comparable to the attenuation length introduced by the GZK effect— and a sky coverage of about 84%. The incomplete sky coverage is mainly due to the so called zone of avoidance centered on the Galactic Plane and caused by the galactic extinction and to a few, narrow stripes which were not observed with enough sensitivity by the IRAS satellite (see Fig. 6.1). These regions are excluded from our analysis with the use of the binary mask available with the PSCz catalogue itself.

6.2.2 The Selection Function

No available galaxy catalogue is complete in volume and therefore completeness estimates derived from the selection effects in flux are needed. More in detail, the relevant quantity to be derived is the fraction of galaxies actually observed at the various redshifts, a quantity also known as the *redshift selection function* $\phi(z)$ (Peebles 1980). A convenient way to express $\phi(z)$ is in terms of the galaxy luminosity function (i.e. the distribution of galaxy luminosities) $\Phi(L)$ as

$$\phi(z) = \frac{\int_{L_{\min}(z)}^{\infty} \mathrm{d}\, L\; \Phi(L)}{\int_0^{\infty} \mathrm{d}\, L\; \Phi(L)}. \tag{6.1}$$

Here $L_{\min}(z)$ is the minimum luminosity detected by the survey in function of redshift. By definition, for a flux-limited survey of limiting flux $f_{\lim}$, $L_{\min}(z)$ is given in terms of the luminosity distance $d_L(z)$ as

$$L_{\min}(z) = 4\pi d_L^2(z) f_{\lim}. \tag{6.2}$$

The luminosity distance depends on the cosmology assumed, though for small redshifts ($z \lesssim 0.1$) it can be approximated by $d_L(z) \simeq z/H_0$.

Generally $\phi(z)$ is inferred from the catalogue data itself in a self-consistent way, using the observational galaxy luminosity distribution to estimate $\Phi(L)$ (Saunders et al. 2000, Sandage et al. 1979, Efstathiou et al. 1988). The quantity $n(z)/\phi(z)$ represents the experimental distribution corrected for the selection effects, which must be used in the computations. A detailed discussion of this issue can be found in Ref. (Blanton et al. 2001). Furthermore, we wish to stress that up to $z \sim 0.1$ evolution effects are negligible and the local universe galaxy luminosity function can be safely used. In the case of deeper surveys like SDSS, cosmological effects cannot be neglected and our approach can still be employed even though a series of corrections, like evolutionary effects or scale-dependent luminosity, must be taken into account (Tegmark et al. 2004). These corrections are needed since luminous galaxies, which dominate the sample at large scales, cluster more than faint ones (Davis et al. 1988). In the case of the PSCz catalogue the selection function is given as (Saunders et al. 2000)

$$\phi(r) = \phi_* \left(\frac{r}{r_*}\right)^{1-\alpha} \left[1 + \left(\frac{r}{r_*}\right)^{\gamma}\right]^{-\left(\frac{\beta}{\gamma}\right)}, \tag{6.3}$$

with the parameters $\phi_* = 0.0077$, $\alpha = 1.82$, $r_* = 86.4$, $\gamma = 1.56$, $\beta = 4.43$ that respectively describe the normalization, the nearby slope, the break distance in Mpc, its sharpness and the additional slope beyond the break (see also Fig. 6.2).

It is clear, however, that even taking into account the selection function we cannot use the catalogue up to the highest redshifts ($z \simeq 0.3$), due to the rapid loss of statistics. At high z, in fact, the intrinsic statistical fluctuation due to the selection effect starts

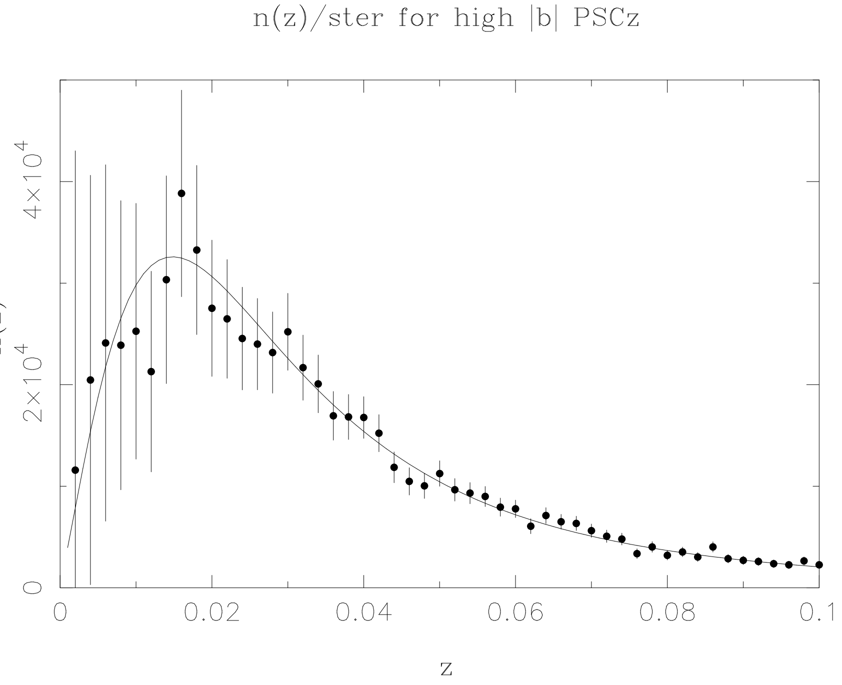


Figure 6.2: *Experimental redshift distribution of the PSCz catalogue galaxies and prediction for an homogeneous universe from the selection function $\phi(z)$ (from (Saunders et al. 2000)); both are normalized in order to represent the number of sources per unit of redshift per steradian.*

to dominate over the true matter fluctuations, producing artificial clusterings not corresponding to real structures ("shot noise" effect). This problem is generally treated constructing from the point sources catalogue a smoothed density field $\rho(\hat{\Omega}, z)$ with a variable smoothing length that effectively increases with redshift, remaining always of size comparable to the mean distance on the sphere of the sources of the catalogue. We minimize this effect by being conservative in setting the maximum redshift at $z = 0.06$ (corresponding to $180\,\mathrm{Mpc}$) where we have still good statistics while keeping the shot noise effect under control. With this threshold we are left with $\sim 11,500$ sources of the catalogue. Furthermore, for the purposes of present analysis, the weight of the sources rapidly decreases with redshift due to the energy losses induced by the GZK effect. In the energy range $E \geq 5{\times}10^{19}$ eV, the contribution from sources beyond $z \simeq 0.06$ is subdominant, thus allowing to assume for the objects beyond $z = 0.06$ an effective isotropic source contribution.

6.3 The Formalism

In the following we describe in some detail the steps involved in our formalism. In Sec. 6.3.1 we summarize our treatment for energy losses, in Sec. 6.3.2 the way the "effective" UHECR map is constructed, and in Sec. 6.3.3 the statistical analysis we perform.

6.3.1 UHECRs Propagation

The first goal of our analysis is to obtain the underlying probability distribution $f_{\rm LSS}$ defined as $f_{\rm LSS}(\hat{\Omega}, E)$ to have a UHECR with energy higher than E from the direction $\hat{\Omega}$. For simplicity here and throughout the paper we shall assume that each source of our catalogue has the same probability to emit a UHECR, according to some spectrum at the source $g(E_i)$. In principle, one would expect some correlation of this probability with one or more properties of the source, like its star formation rate, radio-emission, size, etc. The authors of Ref. (Singh et al. 2004) tested for a correlation $L_{\rm UHECR} \propto L_{\rm FIR}^{\kappa}$, $L_{\rm UHECR}$ being the luminosity in UHECRs and $L_{\rm FIR}$ the one in far-infrared region probed in IRAS catalogue. The results of their analysis do not change appreciably as long as $0 \lesssim \kappa \lesssim 1$. We can then expect that our limit of $\kappa = 0$ might well work for a broader range in parameter space, but this is not of much concern here, since we do not stick to specific models for UHECR sources. The method we discuss can be however easily generalized to such a case, and eventually also to a multi-parametric modelling of the correlation.

In an ideal world where a volume-complete catalogue were available and no energy losses for UHECRs were present, each source should then be simply weighted by the geometrical flux suppression $\propto d_L^{-2}$. The selection function already implies the change of the weight into $\phi^{-1} d_L^{-2}$. Moreover, while propagating to us, high-energy protons lose energy as a result of the cosmological redshift and of the production of $e^{\pm}$ pairs and pions (the dominant process) caused by interactions with CMB. For simplicity, we shall work in the continuous loss approximation (Berezinskii and Grigor'eva 1988). Then, a proton of energy E_i at the source at $z = z_i$ will be degraded at the Earth ($z = 0$) to an energy E_f given by the energy-loss equation[1]

$$\frac{1}{E}\frac{dE}{dz} = -\frac{dt}{dz}\times(\beta_{\rm rsh} + \beta_{\pi} + \beta_{e^{\pm}}). \tag{6.4}$$

Eq.(6.4) has to be integrated from z_i, where the initial Cauchy condition $E(z = z_i) = E_i$

[1] We are neglecting diffuse backgrounds other than CMB and assuming straight-line trajectories, consistently with the hypothesis of weak EGMF.

is imposed, to $z = 0$. The different terms in Eq. (6.4) are explicitly shown below

$$-\frac{dt}{dz} = [(1+z)H_0\sqrt{(1+z)^3\Omega_M + \Omega_\Lambda}]^{-1}, \tag{6.5}$$

$$\beta_{\rm rsh}(z) = H_0\sqrt{(1+z)^3\Omega_M + \Omega_\Lambda}, \tag{6.6}$$

$$\beta_\pi(z,E) \simeq C_\pi(1+z)^3, \;\; E \geq E_{\rm match} \tag{6.7}$$

$$A_\pi(1+z)^3 e^{-\frac{B_\pi}{E(1+z)}}, \;\; E \leq E_{\rm match} \tag{6.8}$$

$$\beta_{e^\pm}(z,E) \simeq \frac{\alpha^3 Z^2}{4\pi^2}\frac{m_e^2 m_p^2}{E^3}\int_2^\infty {\rm d}\xi \frac{\varphi(\xi)}{\exp[\frac{m_e m_p \xi}{2ET_0(1+z)}] - 1}, \tag{6.9}$$

where we assume for the Hubble constant $H_0 = 71^{+4}_{-3}$ km/s/Mpc, and $\Omega_M \simeq 0.27$ and $\Omega_\Lambda \simeq 0.73$ are the matter and cosmological constant densities in terms of the critical one (Spergel et al. 2003). In the previous formulae, m_e and m_p are respectively the electron and proton masses, T_0 is the CMB temperature, and α the fine-structure constant. Since we are probing the relatively near universe, the results will not depend much from the cosmological model adopted, but mainly on the value assumed for H_0. More quantitatively, the r.h.s of Eq. (6.4) changes linearly with H_0^{-1} (apart for the negligible term $\beta_{\rm rsh}$), while even an extreme change from the model ($\Omega_M = 0.27$; $\Omega_\Lambda = 0.73$) to ($\Omega_M = 1$; $\Omega_\Lambda = 0.0$) (the latter ruled out by present data) would only modify the energy loss term by 6% at $z \simeq 0.06$, the highest redshift we consider. The parameterization for β_π as well as the values:

$$\{A_\pi, B_\pi, C_\pi\} = \{3.66{\times}10^{-8}{\rm yr}^{-1}, 2.87{\times}10^{20}\,{\rm eV}, 2.42{\times}10^{-8}{\rm yr}^{-1}\} \tag{6.10}$$

are taken from (Anchordoqui et al. 1997), and $E_{\rm match}(z) = 6.86\,e^{-0.807\,z}{\times}10^{20}\,$eV is used to ensure continuity to $\beta_\pi(z,E)$. An useful parameterization of the auxiliary function $\varphi(\xi)$ can be found in (Chodorowski et al. 1992), which we follow for the treatment of the pair production energy loss. In practice, we have evolved cosmic rays over a logarithmic grid in E_i from 10^{19} to 10^{23} eV, and in z from 0.001 to 0.3. The values at a specific source site has been obtained by a smooth interpolation.

Note that in our calculation i) the propagation is performed to attribute an "energy-loss weight" to each z in order to derive a realistic probability distribution $f_{\rm LSS}(\hat{\Omega}, E)$; ii) we are going to "smooth" the results over regions of several degrees in the sky (see below), thus performing a sort of weighted average over redshifts as well. Since this smoothing effect is by far dominant over the single source stochastic fluctuation induced by pion production, the average effect accounted for by using a continuous energy-loss approach is a suitable approximation.

In summary, the propagation effects provide us a "final energy function" $E_f(E_i, z)$ giving the energy at Earth for a particle injected with energy E_i at a redshift z. Note that, being the energy-loss process obviously monotone, the inverse function $E_i(E_f, z)$ is also available.

6.3.2 Map Making

Given an arbitrary injection spectrum $g(E_i)$, the observed events at the Earth would distribute, apart for a normalization factor, according to the spectrum $g(E_i(E_f, z))dE_i/dE_f$. In particular we will consider in the following a typical power-law $g(E_i) \propto E_i^{-s}$, but this assumption may be easily generalized. Summing up on all the sources in the catalogue one obtains the expected differential flux map on Earth

$$F(\hat{\Omega}, E_f) \propto \sum_k \frac{1}{\phi(z_k)} \frac{\delta(\hat{\Omega} - \hat{\Omega}_k)}{4\pi d^2_{L(z_k)}} E_i^{-s}(E_f, z_k) \frac{dE_i}{dE_f}(E_f, z_k), \tag{6.11}$$

where the selection function and distance flux suppression factors have been taken into account. However, given the low statistics of events available at this high energies, a more useful quantity to employ is the integrated flux above some energy threshold $E_{\rm cut}$, that can be more easily compared with the integrated UHECR flux above the cut $E_{\rm cut}$. Integrating the previous expression we have

$$\begin{aligned} f_{\rm LSS}(\hat{\Omega}, E_{\rm cut}) &\propto \sum_k \frac{1}{\phi(z_k)} \frac{\delta(\hat{\Omega} - \hat{\Omega}_k)}{4\pi d^2_{L(z_k)}} \int_{E_i(E_{\rm cut}, z_k)}^{\infty} E^{-s} {\rm d}E \\ &= \sum_k f_{\rm LSS}(k)\, \delta(\hat{\Omega} - \hat{\Omega}_k), \end{aligned} \tag{6.12}$$

that can be effectively seen as if at every source k of the catalogue it is assigned a weight $f_{\rm LSS}(k)$ that takes into account geometrical effects (d_L^{-2}), selection effects (ϕ^{-1}), and physics of energy losses through the integral in dE. In this "GZK integral" the upper limit of integration is taken to be infinite, though the result is practically independent from the upper cut used provided it is much larger than 10^{20} eV.

It is interesting to compare the similar result expected for an uniform source distribution with constant density; in this case we have (in the limit $z \ll 1$)

$$f_{\rm LSS}(\hat{\Omega}, E_{\rm cut}) \propto \int {\rm d}z \frac{\left[E_{i(E_{\rm cut}, z)}\right]^{-s+1}}{s-1} \equiv \int {\rm d}z\, p(z, E_{\rm cut}, s) \tag{6.13}$$

where the integral in dE has been explicitly performed and the flux suppression weight is cancelled by the geometrical volume factor. The integrand $p(z, E_{\rm cut}, s)$ containing the details of the energy losses also provides an effective cut at high z. The integrand —when normalized to have unit area— can be interpreted as the distribution of the injection distances of CR observed at the Earth. It also suggests the definition of the so-called "GZK sphere" as the sphere from which originates most (say 99%) of the observed CR flux on Earth above an energy threshold $E_{\rm cut}$. In Fig. 6.3 we plot the distribution p for different values of $E_{\rm cut}$ and s. We see that around a particular threshold $z_{\rm GZK}$ the distribution falls to zero: the dependence of $z_{\rm GZK}$ on $E_{\rm cut}$ is quite critical as expected, while there is also a softer dependence on s. This suggests naturally the choice $E_{\rm cut} = 5{\times}10^{19}$ eV for the chosen value $z_{\rm GZK} \simeq 0.06$; at the same time, the energy cut chosen is not too restrictive, ensuring indeed that a significant statistics might be achieved in a few years. For this $E_{\rm cut}$ the isotropic contribution to the flux is sub-dominant; however we can take it exactly into account and the weight of the isotropic part is given by[2]

$$w_{\rm iso} \propto \int_{z_{\rm GZK}}^{\infty} {\rm d}z\, p(z, E_{\rm cut}). \tag{6.14}$$

Finally, to represent graphically the result, the spike-like map (6.12) is effectively smoothed through a gaussian filter as

$$f_{\rm LSS}(\hat{\Omega}, E_{\rm cut}) \propto \sum_k f_{\rm LSS}(k) \exp\left(-\frac{d_s^2[\hat{\Omega}, \hat{\Omega}_k]}{2\sigma^2}\right) + \frac{w_{\rm iso}}{4\pi} 2\pi\sigma^2 \mu(\hat{\Omega}) \tag{6.15}$$

In the previous equation, σ is the width of the gaussian filter, d_s is the spherical distance between the coordinates $\hat{\Omega}$ and $\hat{\Omega}_k$, and $\mu(\hat{\Omega})$ is the catalogue mask (see Section 6.2.1) such that $\mu(\hat{\Omega}) = 0$ if $\hat{\Omega}$ belongs to the mask region and $\mu(\hat{\Omega}) = 1$ otherwise.

6.3.3 Statistical Analysis

Given the extremely poor UHECR statistics, we limit ourselves to address the basic issue of determining the minimum number of events needed to significantly reject "the null hypothesis". To this purpose, it is well known that a χ^2-test is an extremely good estimator. Notice that a χ^2-test needs a binning of the events, but differently from the K-S test performed in (Singh et al. 2004) or the Smirnov-Cramer-von Mises test of (Smialkowski et al. 2002), it has no ambiguity due to the 2-dimensional nature of the problem, and indeed a similar approach was used in (Waxman et al. 1997). A criterion guiding in the choice of the bin size is the following: with N UHECRs events available and M bins, one would expect $\mathcal{O}(N/M)$ events per bin; to allow a reliable application of the χ^2-test, one has to impose $N/M \geq 10$. Each cell should then cover at least a solid angle of $\Delta_M \sim 10{\times}\Delta_{\rm tot}/N$, $\Delta_{\rm tot}$ being the solid angle accessible to the experiment. For $\Delta_{\rm tot} \sim 2\pi$ (50% of full sky coverage), one estimates a square window of side $454°/\sqrt{N}$, i.e. $45°$ for 100 events, $14°$ for 1000 events. Since the former number is of the order of present

[2] The normalization factor is fixed consistently with Eqs. (6.12)-(6.13).

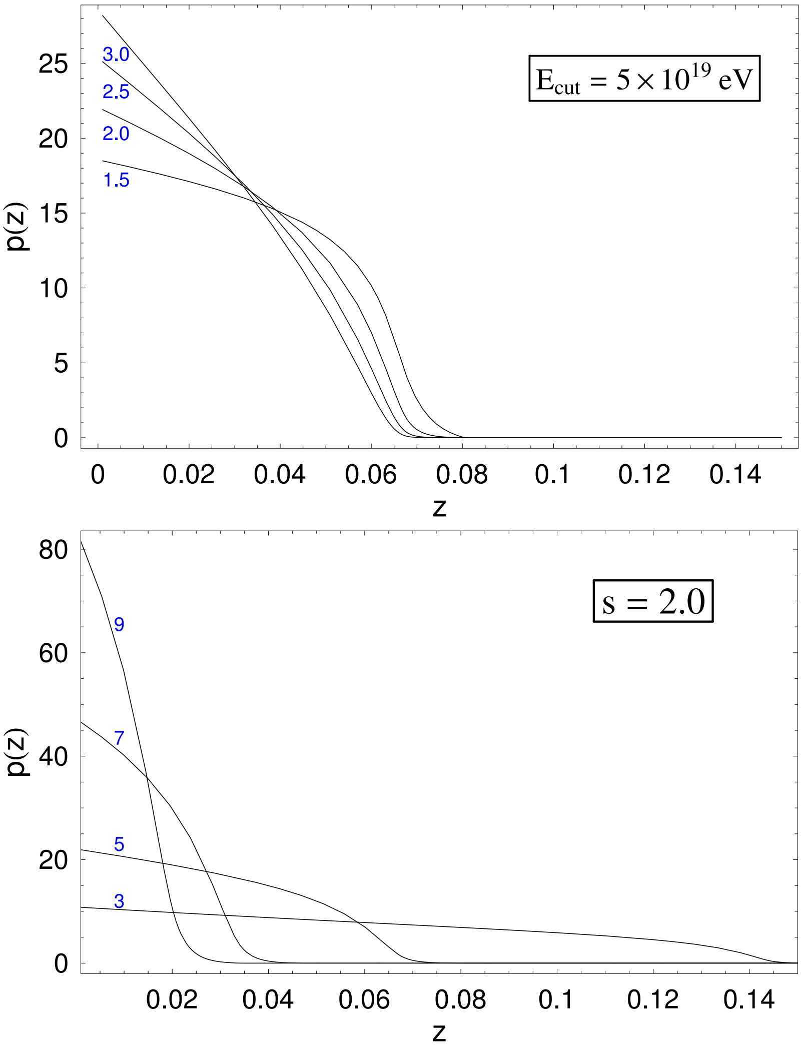


Figure 6.3: *Distribution of the injection distances of CR observed at the Earth for fixed* $E_{\rm cut} = 5{\times}10^{19}$ *eV (top) and* $s = 1.5, 2.0, 2.5, 3.0$ *and for fixed spectral index* $s = 2.0$ *(bottom) and varying* $E_{\rm cut} = 3, 5, 7, 9{\times}10^{19}$ *eV. The area subtended by* $p(z)$ *has been normalized to unity.*

world statistics, and the latter is the achievement expected by Auger in several years of operations, a binning in windows of size 15° represents quite a reasonable choice for our forecast. This choice is also suggested by the typical size of the observable structures, a point we will comment further at the end of this Section. Notice that the GMF, that induces at these energies typical deflections of about 4° (Kachelrieß et al. 2007), can be safely neglected for this kind of analysis. The same remark holds for the angular resolution of the experiment.

Obviously, for a specific experimental set-up one must include the proper exposure $\omega_{\rm exp}$, to convolve with the previously found $f_{\rm LSS}$. The function $\omega_{\rm exp}$ depends on the

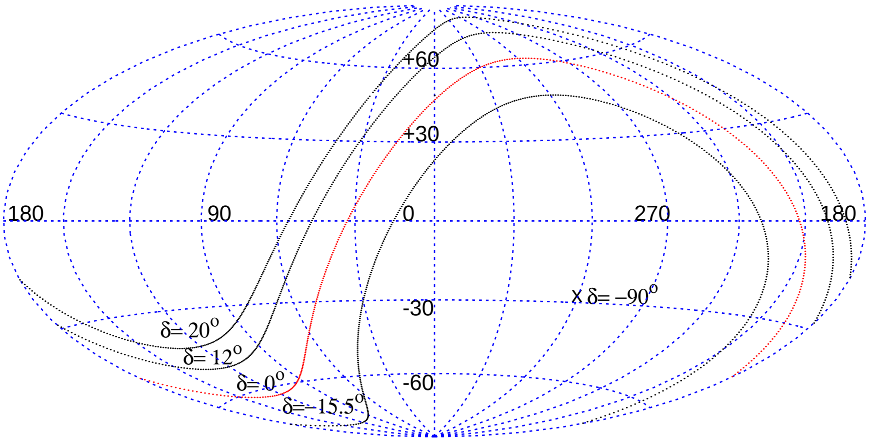


Figure 6.4: *Galactic coordinate reference frame and contours enclosing 68%, 95% and 99% of the Auger exposure function, with the corresponding declinations. The celestial equator ($\delta = 0^\circ$) and south pole ($\delta = -90^\circ$) are also shown.*

declination δ, right ascension RA, and, in general, also on the energy. For observations having uniform coverage in RA, like AGASA or Auger ground based arrays, one can easily parameterize the relative exposure as (Sommers 2001)

$$\omega_{\rm exp}(\delta) \propto \cos\theta_0 \sin\alpha_m \cos\delta + \alpha_m \sin\theta_0 \sin\delta, \tag{6.16}$$

where θ_0 is the latitude of the experiment ($\theta_0 \approx -35^\circ$ for Auger South), α_m is given by

$$\alpha_m = \begin{cases} 0\,, & \text{if } \ \xi > 1 \\ \pi\,, & \text{if } \ \xi < -1 \\ \cos^{-1}\xi\,, & \text{otherwise} \end{cases} \tag{6.17}$$

and

$$\xi \equiv \frac{\cos\theta_{\rm max} - \sin\theta_0 \ \sin\delta}{\cos\theta_0 \ \ \cos\delta}\,, \tag{6.18}$$

$\theta_{\rm max}$ being the maximal zenith angle cut applied (we assume $\theta_{\rm max} = 60^\circ$ for Auger). Contour plots for the Auger exposure function in galactic coordinates are shown in Fig. 6.4.

For a given experiment and catalogue, the null hypothesis we want to test is that the events observed are sampled —apart from a trivial geometrical factor— according to the distribution $f_{\rm LSS}\,\omega_{\rm exp}\,\mu$. Since we are performing a forecast analysis, we will consider test realizations of N events sampled according to a random distribution on the (accessible) sphere, i.e. according to $\omega_{\rm exp}\,\mu$, and determine the confidence level (C.L.) with which the hypothesis is rejected as a function of N. For each realization of N events we calculate the two functions

$$\mathcal{X}^2_{\rm iso}(N) = \frac{1}{M-1}\sum_{i=1}^{M}\frac{(o_i-\epsilon_i[f_{\rm iso}])^2}{\epsilon_i[f_{\rm iso}]}, \tag{6.19}$$

$$\mathcal{X}^2_{\rm LSS}(N) = \frac{1}{M-1}\sum_{i=1}^{M}\frac{(o_i-\epsilon_i[f_{\rm LSS}])^2}{\epsilon_i[f_{\rm LSS}]}, \tag{6.20}$$

where o_i is the number of "random" counts in the i-th bin Ω_i, and $\epsilon_i[f_{\rm LSS}]$ and $\epsilon_i[f_{\rm iso}]$ are the theoretically expected number of events in Ω_i respectively for the LSS and isotropic distribution. In formulae (see Eq. (6.12)),

$$\begin{aligned}\epsilon_i[f_{\rm LSS}] &= N\alpha\frac{\sum_{j\in\Omega_i} f_{\rm LSS}(j)\omega_{\rm exp}(\delta_j)\mu(j)+w_{\rm iso}/4\pi\, S[\Omega_i]}{\sum_j f_{\rm LSS}(j)\omega_{\rm exp}(\delta_j)\mu(j)+w_{\rm iso}/4\pi\, S_\omega}, &(6.21)\\ \epsilon_i[f_{\rm iso}] &= N\alpha\frac{S[\Omega_i]}{S_\omega}, &(6.22)\end{aligned}$$

where $S[\Omega_i]=\int_{\Omega_i}d\Omega\,\omega_{\rm exp}\mu$ is the spherical surface (exposure- and mask-corrected) subtended by the angular bin Ω_i, and similarly $S_\omega=\int_{4\pi}d\Omega\,\omega_{exp}\mu$. The mock data set is then sampled $\mathcal{N}$ times in order to establish empirically the distributions of $\mathcal{X}^2_{\rm LSS}$ and $\mathcal{X}^2_{\rm iso}$, and the resulting distribution is studied as function of N (plus eventually $s, E_{\rm cut}$, etc.). The parameter

$$\alpha \equiv \frac{\int d\Omega\;\omega_{exp}(\delta)\mu(\Omega)}{\int d\Omega\;\omega_{exp}(\delta)} \tag{6.23}$$

is a mask-correction factor that takes into account the number of points belonging to the mask region and excluded from the counts o_i. Note that the random distribution is generated with N events in all the sky view of the experiment, but, effectively, only the region outside the mask is included in the statistical analysis leaving us with effective $N\alpha$ events to study. This is a limiting factor due to quality of the catalogue: With a better sky coverage the statistics is improved and the number of events required to asses the model can be reduced.

As our last point, we return to the problem of choice of the bin size. To assess its importance we studied the dependence of the results on this parameter. For a cell side larger than about $\sim 25°$ the analysis loses much of its power, and a very high N is required to distinguish the models and obtain meaningful conclusions. This is somewhat expected looking at the map results that we obtain, where typical structures have dimensions of the order $15°-20°$. A greater cell size results effectively in a too large smoothing and a consequent lost of information. On the other hand, a cell size below $4^\circ - 6^\circ$ makes the use of a χ^2 analysis not very reliable, because of the low number of events in each bin expected for realistic exposure times. In the quite large interval $\sim 6^\circ - 20^\circ$ for the choice of the cell size, however, the result is almost independent of the bin size, that makes us confident on the reliability of our conclusions.

6.4 Results

In Fig. 6.5 we plot the smoothed maps in galactic coordinates of the expected integrated flux of UHECRs above the energy threshold $E_{\rm cut} = 3, 5, 7, 9{\times}10^{19}$ eV and for slope parameter $s = 2.0$, with smoothing angle $\sigma = 3^\circ$ and contours enclosing 95%, 68%, 38%, 20% of the corresponding distribution; the isotropic part has been taken into account and the ratio of the isotropic to anisotropic part $w_{\rm iso}/\sum_k f_{\rm LSS}(k)$ is respectively $83\%, 3.6\%, \ll 1\%, \ll 1\%$.

Only for $E_{\rm cut} = 3{\times}10^{19}$ eV the isotropic background constitutes then a relevant fraction, since the GZK suppression of far sources is not yet present. For the case of interest $E_{\rm cut} = 5{\times}10^{19}$ eV the contribution of $w_{\rm iso}$ is almost negligible, while it practically disappears for $E_{\rm cut} \geq 7{\times}10^{19}$ eV. Varying the slope for $s = 1.5, 2.0, 2.5, 3.0$ while keeping $E_{\rm cut} = 5{\times}10^{19}$ eV fixed produces respectively the relative weights $8.0\%, 3.6\%, 1.8\%, 0.9\%$, so that only for very hard spectra $w_{\rm iso}$ would play a non-negligible role (see also Fig. 6.3).

Due to the GZK-effect, as it was expected, the nearest structures are also the most prominent features in the maps. The most relevant structure present in every slide is the Local Supercluster. It extends along $l \simeq 140^\circ$ and $l \simeq 300^\circ$ and includes the Virgo cluster at $l = 284^\circ, b = +75^\circ$ and the Ursa Major cloud at $l = 145^\circ, b = +65^\circ$, both located at $z \simeq 0.01$. The lack of structures at latitudes from $l \simeq 0^\circ$ to $l \simeq 120^\circ$ corresponds to the Local Void. At higher redshifts the main contributions come from the Perseus-Pisces supercluster ($l = 160^\circ, b = -20^\circ$) and the Pavo-Indus supercluster ($l = 340^\circ, b = -40^\circ$), both at $z \sim 0.02$, and the very massive Shapley Concentration ($l = 250^\circ, b = +20^\circ$) at $z \sim 0.05$. For a more detailed list of features in the map, see the key in Fig. 6.6.

The $E_{\rm cut}$-dependence is clearly evident in the maps: as expected, increasing $E_{\rm cut}$ results in a map that closely reflects the very local universe (up to $z \sim 0.03 - 0.04$) and its large anisotropy; conversely, for $E_{\rm cut} \simeq 3, 4{\times}10^{19}$ eV, the resulting flux is quite isotropic and the structures emerge as fluctuations from a background, since the GZK suppression is not yet effective. This can be seen also comparing the near structures with the most distant ones in the catalogue: while the Local Supercluster is well visible in all slides, the signal from the Perseus-Pisces super-cluster and the Shapley concentration is of comparable intensity only in the two top panels, while becoming highly attenuated for $E_{\rm cut} = 7{\times}10^{19}$ eV, and almost vanishing for $E_{\rm cut} = 9{\times}10^{19}$ eV. A similar trend is

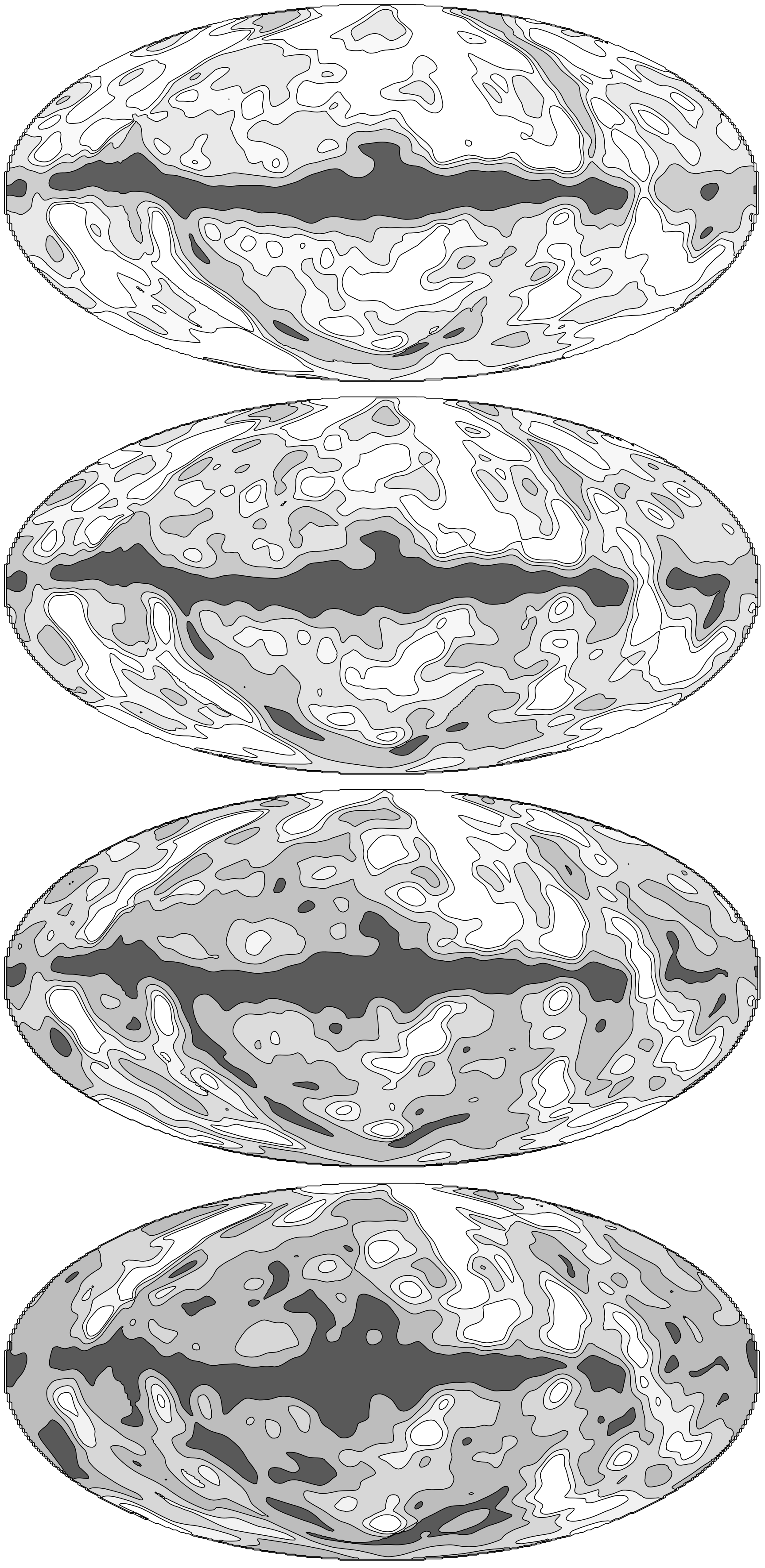


Figure 6.5: *Equal area Hammer-Aitoff projections of the smoothed UHECRs arrival directions distribution (Eq. (6.15)) in galactic coordinates obtained for fixed* $s = 2.0$ *and, from the upper to the lower panel, for* $E_{\rm cut} = 3, 5, 7, 9 \times 10^{19}$ *eV.*

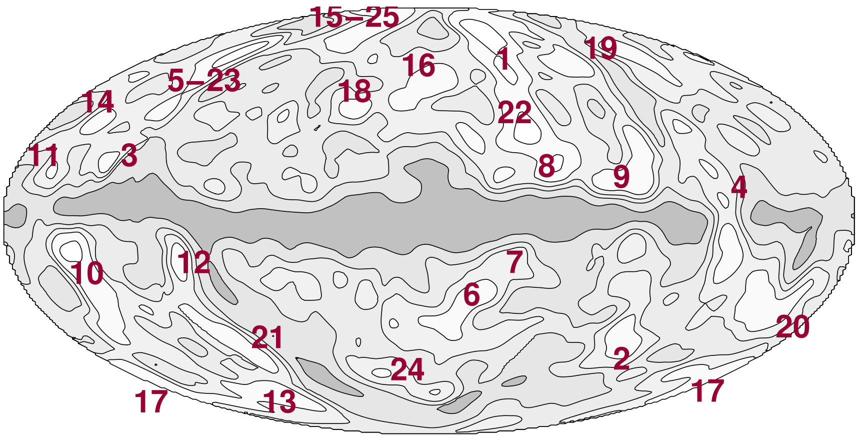


Figure 6.6: *Detailed key of the structures visible in the UHECR maps; arbitrary contour levels. Labels correspond to: (1) Southern extension of Virgo and Local Supercluster; (2)Fornax-Eridani Cluster; (3) Cassiopea Cluster; (4) Puppis Cluster; (5) Ursa Major Cloud; (6-7) Pavo-Indus and "Great Attractor" region; (8) Centaurus Super-Cluster; (9) Hydra Super-Cluster; (10) Perseus Super-Cluster; (11) Abell 569; (12) Pegasus Cluster; (13-17) Pisces Cluster; (14) Abell 634; (15) Coma Cluster; (16-18) Hercules Supercluster; (19) Leo Supercluster; (20) Columba Cluster; (21) Cetus Cluster; (22) Shapley Concentration; (23) Ursa Major Supercluster; (24) Sculptor Supercluster; (25) Bootes Supercluster.*

observed for increasing s at fixed $E_{\rm cut}$, though the dependence is almost one order of magnitude weaker. Looking at the contour levels in the maps we can have a precise idea of the absolute intensity of the "fluctuations" induced by the LSS; in particular, for the case of interest of $E_{\rm cut} = 5{\times}10^{19}$ eV the structures emerge only at the level of 20%-30% of the total flux, the 68% of the flux actually enclosing almost all the sky. For $E_{\rm cut} = 7,9{\times}10^{19}$ eV, on the contrary, the local structures are significantly more pronounced, but in this case we have to face with the low statistics available at this energies. Then in a low-statistics regime it's not an easy task to disentangle the LSS and the isotropic distributions.

The structures which are more likely to be detected by Auger (see also Fig. 6.4) are the Shapley concentration, the Southern extension of the Virgo cluster, the Local Supercluster and the Pavo-Indus super-cluster. Other structures, such as the Perseus-Pisces supercluster and the full Virgo cluster are visible only from the Northern hemisphere and are therefore within the reach of experiments like Telescope Array, or the planned North extension of the Pierre Auger Observatory. Moreover, the sky region obscured by the heavy extinction in the direction of the Galactic Plane reflects a lack of information about features possibly "hidden" there. Unfortunately, this region falls just in the middle of the Auger field of view, thus reducing —for a given statistics N— the significance of the check of the null hypothesis. Numerically, this translates into a smaller

$N \setminus s$	1.5	2.0	2.5	3.0
50	(42:6)	(47:8)	(52:10)	(52:10)
100	(55:9)	(60:12)	(66:14)	(69:16)
200	(72:27)	(78:33)	(84:40)	(86:43)
400	(92:61)	(95:72)	(97:80)	(98:83)
600	(98:85)	(99:91)	(100:96)	(100:97)
800	(100:95)	(100:98)	(100:99)	(100:100)
1000	(100:98)	(100:100)	(100:100)	(100:100)

Table 6.1: *The probability (in %) to reject the isotropic hypothesis at (90%:99%) C.L. when UHECRs follow the LSS distribution, as a function of the injection spectral index and of the observed number of events, fixing* $E_{\rm cut} = 5{\times}10^{19}$ *eV.*

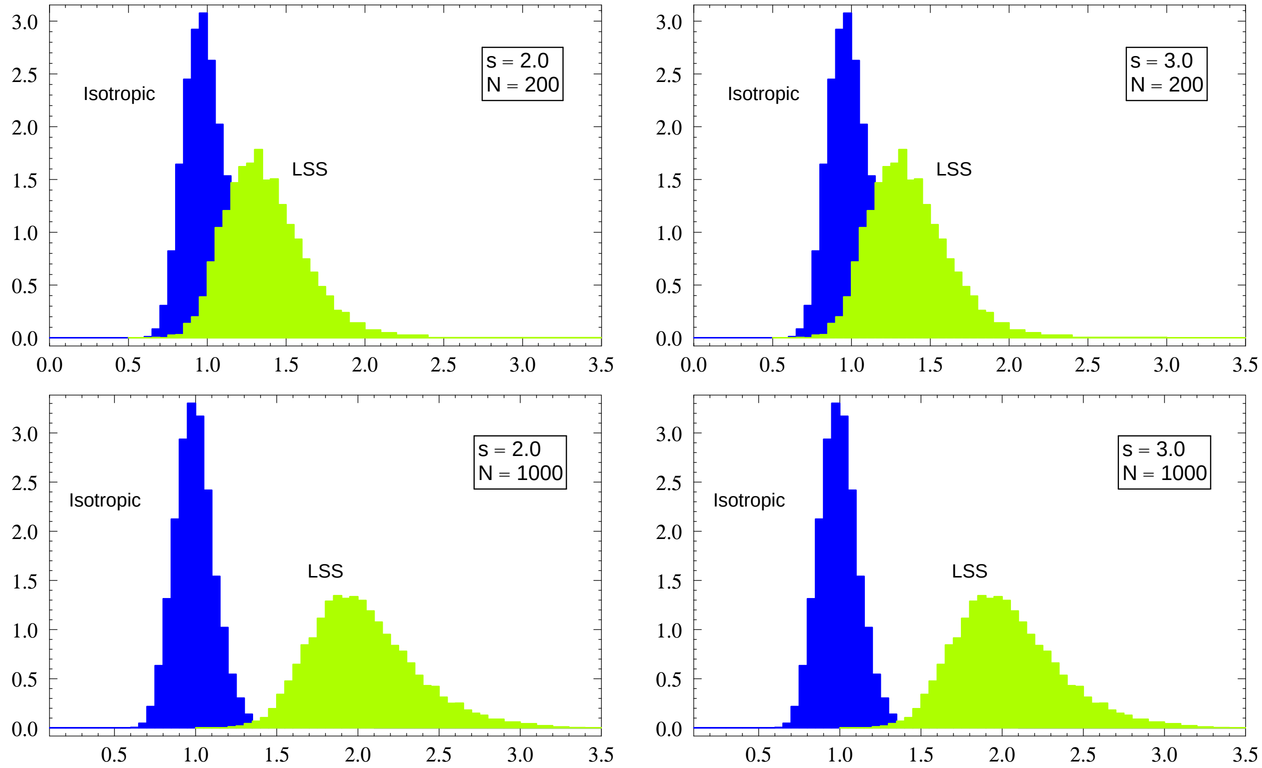


Figure 6.7: *The probability distributions of the estimators* $\mathcal{X}^2_{\rm iso}$ *and* $\mathcal{X}^2_{\rm LSS}$ *for the cases* $s = 2.0, 3.0$ *and for* $N = 200, 1000$ *events, fixing* $E_{\rm cut} = 5{\times}10^{19}$ *eV. The distribution are the results of 10000 monte-carlo simulation like described in the text.*

value of the factor α of Eq. (6.23) with respect to an hypothetical "twin" Northern Auger experiment.

A quantitative statistical analysis confirms previous qualitative considerations. In

Table 6.1 we report the probability to reject the isotropic hypothesis at 90% and 99% C.L. when UHECRs follow the LSS distribution, as a function of the injection spectral index and of the observed number of events, fixing $E_{\rm cut} = 5{\times}10^{19}$ eV. In Figure 6.7 we show the distributions of the functions $\mathcal{X}^2_{\rm iso}$ and $\mathcal{X}^2_{\rm LSS}$ introduced in the previous section for $s = 2.0, 3.0$ and $N = 200, 1000$, for the same cut $E_{\rm cut} = 5{\times}10^{19}$ eV. It is clear that a few hundreds events are hardly enough to reliably distinguish the two models, while $N =$800–1000 should be more than enough to reject the hypothesis at 2-3 σ, independently of the injection spectrum. Steeper spectra however slightly reduce the number of events needed for a given C.L. discrimination. It is also interesting to note that, using different techniques and unconstrained LSS simulations, it was found that a comparable statistics is needed to probe a magnetized local universe (Sigl et al. 2004). It is worthwhile stressing that our conclusions should be looked as conservative, since only proton primaries have been assumed, and constant source properties. Variations in individual source power and a mixed composition could increase the "cosmic variance" and make more difficult to distinguish among models for the source distribution (Sigl et al. 2004).

With respect to previous literature on the subject, our analysis is the closest to the one of Ref. (Waxman et al. 1997). Apart for technical details, the greatest differences with respect to this work arise because of the improved determination of crucial parameters undergone in the last decade. Just to mention a few, the Hubble constant used in (Waxman et al. 1997) was 100 km s^{-1} Mpc^{-1}, against the presently determined value of 71^{+4}_{-3} km s^{-1} Mpc^{-1}: this changes by a 30% the value of the quantity $z_{\rm GZK}$ (see Sec. 6.3.2). Moreover, the catalogue (Fisher et al. 1995) that was used in (Waxman et al. 1997) contains about 1/3 of the objects we are considering, has looser selection criteria and larger contaminations (Saunders et al. 2000). Finally, the specific location of the Southern Auger observatory was not taken into account. All together, when considering these factors, we find quite good agreement with their results.

Some discrepancy arises instead with the results of (Singh et al. 2004), whose maps appear to be dominated by statistical fluctuations, which mostly wash away physical structures. This has probably to be ascribed to two effects, the energy cut $E_{\rm cut} = 4{\times}10^{19}$ eV and the inclusion of high redshift object (up to $z \sim 0.3$) of the catalogue (Saunders et al. 2000) in their analysis. Their choice of $E_{\rm cut} = 4{\times}10^{19}$ eV implies indeed $z_{\rm GZK} \simeq 0.1$, i.e. a cutoff in a redshift range where shot noise distortions are no longer negligible. The same remarks hold for Ref. (Smialkowski et al. 2002), which also suffers of other missing corrections (Singh et al. 2004). Also, in both cases, the emphasis is mainly in the analysis of the already existing AGASA data than in a forecast study. Our results however clearly show that AGASA statistics —only 32 data at $E \geq 5{\times}10^{19}$ eV in the published data set (Hayashida et al. 2000), some of which falling inside the mask— is too limited to draw any firm conclusion on the hypothesis considered.

6.5 Summary and conclusion

In this work we have summarized the technical steps needed to properly evaluate the expected anisotropy in the UHECR sky starting from a given catalogue of the local universe, taking into account the selection function, the blind regions, and the energy-loss effects. By applying this method to the catalogue (Saunders et al. 2000), we have established the minimum statistics needed to significantly reject the null hypothesis, in particular providing a forecast for the Auger experiment. We showed with a χ^2 approach that several hundreds data are required to start testing the model at Auger South. The most prominent structures eventually "visible" for this experiment were also identified.

Differently from other statistical tools based e.g. on auto-correlation analysis, the approach sketched above requires an Ansatz on the source candidates. The distribution of the luminous baryonic matter considered here can be thought as a quite generic expectation deserving interest of its own, but it is also expected to correlate with many sources proposed in the literature. In any case, if many astrophysical sources are involved in UHECR production, it is likely that they should better correlate with the local baryonic matter distribution than with an isotropic background.

As already stated, this work has to be intended as mainly methodological. Until now, the lack of UHECR statistics and the inadequacy of the astronomical catalogues has seriously limited the usefulness of such a kind of analysis. However, progresses are expected in both directions in forthcoming years. From the point of view of UHECR observatories, the Southern site of Auger is almost completed, and already taking data. Working from January 2004 to June 2005, Auger has reached a cumulative exposure of 1750 km^2 sr yr, observing 10 events over $10^{19.7}$ eV=5×10^{19} eV (see: `www.auger.org/icrc2005/spectrum.html`). Notice that statistical and systematic errors are still quite large, and a down-shift in the $\log_{10} E$ scale of 0.1 would for example change the previous figure to 17 events. Once completed, the total area covered will be of 3000 km^2, thus improving by one order of magnitude present statistics in a couple of years (The Pierre Auger Collaboration 2005b). The idea to build a Northern Auger site strongly depends on the possibility to perform UHECR astronomy, for which full sky coverage is of primary importance. In any case, the Japanese-American Telescope Array in the desert of Utah is expected to become operational by 2007 (Kasahara 2005). It should offer almost an order of magnitude larger aperture per year than AGASA in the Northern sky, with a better control over the systematics thanks to a hybrid technique similar to the one employed in Auger.

The other big step is expected in astronomical catalogues. The 2MASS survey (Jarrett et al. 2000) has resolved more than 1.5 million galaxies in the near-infrared, and has been explicitly designed to provide an accurate photometric and astrometric knowledge of the nearby Universe. The observation in the near IR is particularly sensitive to the stellar component, and as a consequence to the luminous baryons. Though the redshifts of the sources have to be obtained via photometric methods, the larger error on the distance estimates (about 20% from the 3-band 2MASS photometry (Jarrett 2004)) is more than compensated by the larger statistics. An analysis of this catalogue for UHECR purposes is in progress. Independently of large sky coverage, deep surveys like SDSS (Adelman-McCarthy et al. 2006) undoubtedly have an important role in mapping the local universe as well. For example, the information encoded in such catalogues can be used to validate methods —like the neural networks (Tagliaferri et al. 2002, Collister and Lahav 2004, Vanzella et al. 2004)— used to obtain photometric redshifts. An even better situation is expected from future projects like SDSS II (see: `www.sdss.org`). Finally, a by-product of these surveys is the discovery and characterization of active galactic nuclei (Best, Kauffmann, Heckman, Brinchmann, Charlot, Ivezić and White 2005, Best, Kauffmann, Heckman and Ivezić 2005), which in turn could have interesting applications in the search for the sources of UHECRs.

Acknowledgments

We thank M. Kachelrieß and M. Paolillo for reading the manuscript and for useful comments. A.C. thanks the Astroparticle and High Energy Physics Group of Valencia for the nice hospitality during the initial part of the work and acknowledges the Italian-Spanish *Azione Integrata* for financial support.

Chapter 7

A possible origin of cosmic Voids

Science must begin with myths, and with the criticism of myths

K. Popper

7.1 Voids and Large Scale Structure: the current scenario

The large-scale galaxy distribution is highly inhomogeneous. Groups, clusters and superclusters of galaxies and large voids can be observed. During last decades, much attention was paid on the analysis of bound structures as groups and clusters. Recently, new superclusters catalogues were constructed from redshift surveys, like the 2dFGRS (Sadler et al. 2002), and compared with large cosmological simulations (Einasto 2007b, Einasto 2007a). In few words, there are large regions in the universe without bright galaxies so called cosmic voids. Early on very large voids over $50h^{-1}$ Mpc diameter were found by Gregory and Thompson (Gregory et al. 1978) and Kirshner (Kirshner et al. 1981), while much more common are voids with typical diameters of about 10 h^{-1} Mpc that fill most of cosmic space. The explanation of the origin and and evolution of such structures is not obvious. According to the standard paradigm of cosmological structure formation, negative potential wells from primordial unhomogeneities attract all matter in bound structures. In the same way, positive potential perturbations expel matter, but observed voids are too large for complete emptying. Therefore, in addition to the dilution of matter, the galaxy formation probability should be suppressed in underdense regions (Lee and Shandarin 1998, Madsen et al. 1998). Recently two authors (Furlanetto and Piran 2006), applied these ideas within the excursion set formalism of gravitational instability. These analytical theories derived void size distributions that are peaked typically at diameters below 10 h^{-1} Mpc which seem to be smaller than observed void sizes. Voids were routinely identified in all wide-field redshift surveys as the CfA (de Lapparent et al. 1986, Vogeley et al. 1994), the SSRS2 (El-Ad and Piran 1997), the LCRS (Müller et al. 2000), the IRAS-survey (El-Ad and Piran 2000), the 2dFGRS (Hoyle and Vogeley 2004, Croton et al. 2004, Patiri et al. 2006), the SDSS (Rojas et al. 2004, Rojas et al. 2005, Patiri et al. 2006), and the DEEP2 survey with an analysis

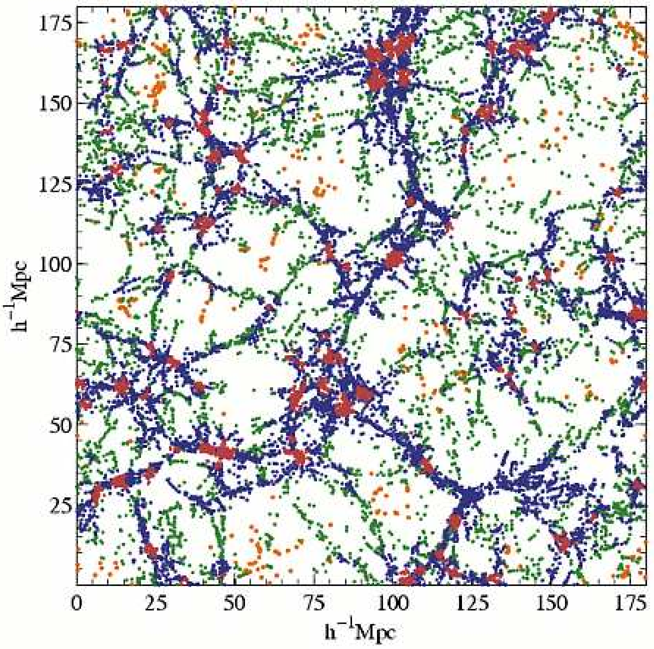


Figure 7.1: *Classification of halo environments in a slice of 10 Mpc/h thickness from a cosmological simulations of the large scale structure distribution of galaxies (Hahn et al. 2007), where each point corresponds to a halo of at least 10 particles. Haloes are colour coded for four different environments: clusters (red), filaments (blue), sheets (green) and voids (orange).*

of voids up to redshift $z \approx 1$ (Conroy et al. 2005). However, many void searches are only devoted to the identification of large voids, while other void finding algorithms depend crucially on special procedures as a previous identification of wall galaxies by an overdensity criterion and then a specific search for voids bounded by wall galaxies (El-Ad and Piran 1997, Hoyle and Vogeley 2004). Furthermore, the void search depends on the galaxy sample used for defining voids, in particular on the limiting magnitude of the galaxy sample. In an influential paper, Peebles (Peebles 2001) derived from nearest neighbour statistics that galaxies of different brightness seem to give shape to the same voids. He claimed that this contradicts the standard CDM scenario of galaxy and structure formation that seem to predict a hierarchy of galactic structures with smaller structure for fainter objects sitting in less massive dark matter halos, i.e also smaller voids for fainter objects (see figure 7.1). In a follow up theoretical study (White et al. 2002), it was showed from high-resolution simulation that voids defined by bright galaxies are also underdense in faint galaxies, i.e. that bright and faint galaxies respect similar voids.

More recently, another author (Colberg 2007) compared different void search algorithms and found that most proposed algorithm find comparable locations and sizes of large voids. This is very likely not the case for the large number of small voids that fill a significant part of space. In any case, the statistical characterization of the largest Voids is robust since their observational properties are nearly constant with different search algorithm and respect to different galaxy populations.

7.2 Observational features of Voids

The observational studies concerning cosmological Voids, like those discussed in the previous section, have been mainly devoted to the investigation of the visible matter content of these large structures, i.e. the characteristics of galaxies contained in the Voids. Since at the present stage of observations, only few galaxies have been observed in the inner regions of Voids, in order to get some statistically robust insight into their features, it has been necessary to investigate the content of the outer regions of the cosmic Voids. In general, three kinds of indirect observations are used to explore the content of cosmic Voids, namely the study of the morphology of galaxies nearby the Void, the measure of the dispersion of the peculiar velocities of galaxies nearby the Void and the observation of the effects and distortion introduced by the Void when acting as a gravitational lens on the background galaxy distribution and on the cosmic microwave background observed radiation anisotropy. The most detailed observational study of the distribution and properties of cosmic Voids to day have been produced by a research group at the Department of Physics of the Drexel University in Philadelphia (Hoyle and Vogeley 2004). These studies are based on an automated algorithm developed by (El-Ad and Piran 1997) and named Voidfinder, which allows to construct a void survey starting from a galaxy spectroscopic survey. Applying Voidfinder to the 2dFGRS galaxy survey, (Hoyle and Vogeley 2004) have found 289 cosmological voids, whose observational characteristics are summarized in the list below:

- The average radius of the Voids selected is (12 ± 2) Mpc;
- The density contrast with respect to the cosmological background is ~ 0.95;
- The fraction of the volume of the entire Universe occupied by the Voids is $\sim 40\%$;
- No sign of morphological biasing of the population of galaxies hosted by the Voids is observed;
- Neighbouring galaxy population shows a velocity dispersion significantly smaller than average.

It is interesting to notice that the current measurements of the velocity dispersion in large samples of galaxies located near the boundaries of cosmic Voids suggest that there could be some amount of matter inside the Voids which has not yet been observed.

7.3 Models of Voids formation

The observational evidences mentioned in the previous section has generated great interest about the formulation of models capable of explaining the actual distribution of matter inside cosmological voids and their mechanism of formation. It is generally accepted that there is strong correlation between the morphology of galaxies and the density of the environment where they are placed. More in particular, early-type galaxies are more likely to form in higher density environments (for example in clusters of galaxies), while late-type are commoner where the density is lower. This morphology-density correlation, originally foreseen in the so called biased galaxy formation picture, is supported by several observational studies (for example, (Goto et al. 2003) and is particularly relevant to investigate the possible presence of matter inside the cosmic Voids. In fact, if the Voids are completely devoided of matter as suggested by observations, then the galaxies we observe nearby their boundaries should present a strongly characteristic morphological distribution because they formed in an environment in which the density of visible baryonic matter is rapidly falling from values near to the average of cosmological background distribution (outside the Void) to extremely low values (inside the underdense Void). As already stated, so far there is no conclusive observational evidence confirming the existence of this characteristic distribution. In order to get through this apparent contradiction, the possibility that the internal part of the Voids contains some form of dark matter (as originally stated by (Peebles 2001)) has been considered. Some authors have stated that Voids could contain large overdensities of very low brightness galaxies that have escaped observations so far, accounting for the mass which is allegedly reported to be placed in their centres (Kirshner et al. 1981). The problem of the current composition of Voids in terms of luminous and dark matter is strictly correlated to the other important issue concerning Voids formation. Among various models proposed in the literature, particularly interesting is the one originally introduced by (Friedmann and Piran 2001), according to which Voids have formed from the comoving expansion of negative primordial perturbations in the density field, just like the observed overdensities (clusters and groups of galaxies) have developed from the comoving expansion of primordial positive fluctuations in the density field . Several N-body simulations of this formation mechanism based on the cold dark matter scenario produced results which are consistent with observational data. More recently another formation mechanism was proposed by (Stornaiolo 2002), which has been object of our observational investigation. In this scenario, the collapse of extremely large wavelengths positive perturbations led to the formation of low density/high mass black holes (Cosmological Black Holes or CBH). Voids have then formed by the comoving expansion of the matter surrounding the collapsed perturbation. According to this model, at the centre of each cosmological Void (assumed to be spherical), is located a CBH having mass expressed by the equation:

$$M_{Void} = \frac{4}{3}\pi\Omega_{cbh}\rho_{c0}R^3 \tag{7.1}$$

where the parameter Ω_{cbh} can be expressed as:

$$\Omega_{cbh} = \frac{\rho_{cbh}}{\rho_{c0}} \tag{7.2}$$

and represents the ratio of the density ρ_{cbh} of all the CBHs contained in the Universe to the critical density as observed at the present time ρ_{c0}, while R is the radius of the Void. The mass of the CBH partially or completely compensate the lack of visible matter in the volume occupied by the Void according to the values of Ω_{cbh}. The actual value of this parameter represents a degree of freedom of the model, and will to be modified according to future observational estimates of CBH mass. Nevertheless, at this stage, it is reasonable to assume that ρ_{cbh} has the same value of the observed matter density ρ_{matter} scaled by the fraction of the volume of the universe occupied by the voids. This assumption implies that $\Omega_{cbh} \sim 0.2$. As an example, a cosmic Void with a diameter of ~ 25 Mpc, according to the previously explained model, should host a CBH with mass of the order of $10^{14}M_{\odot}$, which is comparable to the mass of a cluster of galaxies.

7.4 Voids as gravitational lenses

The gravitational lensing properties of Voids have so far remained a topic treated only in theoretical papers. This is due to the fact that if the Voids are filled with low brightness (and consequently not visible in the currently available observations) ordinary galaxies, then they would have no observable peculiar gravitational lensing effect and, as pointed out by (Amendola et al. 1999), this would be the case even if Voids are completely devoided of matter. One case of interest is represented by the CBH model briefly discussed in the previous section and reported in the chapter 8. The authors examined the results of simulations of a Void/CBH/background galaxies system, focusing in particular on the lensing effects introduced in the background galaxies distribution by the presence of a massive CBH in the centre of the Void. The result is that a CBH contained in the Void acts as a Schwarzschild lens, with observable consequences on the luminosity and spatial distributions of galaxies in a region around the projected centre of the Void.

Published as: Serpico, M., D'Abrusco, R., Longo, G., Stornaiolo, C. – *"Cosmological black holes as voids progenitors: Simulations."* 2007, General Relativity and Gravitation, vol. 39, 10, pag. 1551-1562

Chapter 8

Application II: Observational scrutiny of a model of the origin of the Voids.

Abstract

Cosmological black holes (CBH), i.e. black holes with masses of the order of $10^{13} \div 10^{14} M_{\odot}$*, have been proposed as possible progenitors of galaxy voids. The presence of a CBH in the central regions of a void should induce significant gravitational lensing effects and in this paper we discuss such gravitational signatures using simulated data. These signatures may be summarized as follows: i) a blind spot in the projected position of the CBH where no objects can be detected; ii) an excess of faint secondary images; iii) an excess of double images having a characteristic angular separation. All these signatures are shown to be detectable in future deep surveys.*

8.1 Introduction

Voids are among the largest structures known in the Universe with typical diameters ranging between 10 and 25 Mpc and with just a few of about 30 Mpc (Hoyle and Vogeley 2004) (Goldberg et al. 2005). Their origin is controversial. In the first models (see for example (Friedmann and Piran 2001)), it was assumed that voids formed from the evolution of negative primordial perturbations in the density field. More in detail, according to these models, void formation is the result of two correlated processes. The first one is the comoving expansion of the negative fluctuations. The second arises from the biased galaxy formation picture: galaxies are less likely to form in the underdense regions created by this expansion. Several N-body simulations of this formation mechanism based on the cold dark matter scenario (Benson et al. 2003) produced results which are consistent with observational data.

But, as noted by Peebles in (Peebles 2001), the small relative velocity dispersion in the CfA sample shows that, when $\Omega_m = 1$, most of the mass has to be in the voids (see references in (Peebles 2001)). This could occur in a natural way in the CDM model: if ordinary galaxies formed preferentially in high-density regions, they would be strongly clustered, leaving most of the mass in the voids. In the same paper Peebles suggested that this might be possible also when $\Omega_m < 1$.

Following this remark, it was proposed by Stornaiolo in (Stornaiolo 2002) (see also (Capozziello et al. 2004)) a different scenario for the void formation. According to this scenario the voids are the concomitant result of the collapse of extremely large wavelength positive perturbations, which led to the formation of low density/high mass black holes (Cosmological Black Holes or CBH) and the comoving expansion of the matter surrounding the collapsed perturbation. It is assumed that the wavelengths of the perturbations are of the order of the void comoving diameter and that the obey the observed perturbation spectrum.

According to this scenario the so-called *Swiss Cheese* universe (Einstein and Straus 1946) (Kantowski 1969) is a suitable model to describe voids. Voids are assumed to be spherical cavities, each with a spherical black hole with mass

$$M = \frac{4}{3}\pi\Omega_{cbh}\rho_c R^3_{void}, \tag{8.1}$$

at their center. The parameter

$$\Omega_{cbh} = \frac{\rho_{cbh}}{\rho_c} \tag{8.2}$$

represents the fraction of density due the totality of these black holes with respect to the total density of the universe; $\rho_c = 1.88 \times 10^{-29}\,\mathrm{g\,cm^{-3}}h^2\Omega$ is the present-day critical density of the universe. Applying formula (8.1) to a void with a typical radius $R_{void} = 12Mpc$ and a presumable value of $\Omega_{cbh} = 0.2$ it results that the corresponding CBH has a mass $M \sim 4 \times 10^{14}\,M_\odot$. In the following we shall use this values of the parameters.

If we identify the comoving dimensions of the voids with the wavelengths of the cosmological perturbations that generated the CBH, then, according to the observed power spectrum of density fluctuations (see (Tegmark and Zaldarriaga 2002), (Tegmark et al. 2004) and (Spergel et al. 2007)), we find that these objects must have masses comprised between 10^{13} and $10^{14}\,M_\odot$. This shows that our assumption (i.e. $\Omega_{cbh} \sim \Omega_{matter}$) is consistent with observations.

The Swiss Cheese model can be extended even in presence of the cosmological constant as shown in (Balbinot et al. 1988). This result assures that this model holds in different cosmological scenarios, so allowing us to implement our simulations using a Einstein - De Sitter model and aiming at generalizing straightforwardly the results to more complex models.

According to the Swiss Cheese model, if the CBHs compensate the voids, then there would be no direct interaction between them and the external structures.

But as observed in (Capozziello et al. 2004), the model studied here can justify the main properties of the voids. Voids are characterized by not being completely empty structures, because it has been observed an underdensity in the distribution of the galaxies of the order of 10% with respect to the external density. The galaxies are distributed close to the border of the void (see (Hoyle and Vogeley 2004) and (Goldberg et al. 2005)).

Observations of cosmic velocity fields (Faber et al. 1994) show that large scale structure around the voids does not present velocity fields converging toward the voids, but toward the visible clusters and superclusters around the voids. In other words voids appear as repulsive structures. This apparent shortcoming, in the framework of the model considered here, can be overcome by the Birkoff theorem which states that the stationary solutions are also static if the spherical symmetry is restored. So, a fraction of galaxies is attracted by clusters and superclusters ″outside″ the void while another fraction shows no dynamics since it has been already attracted ″inside″ the void. This fact could be interpreted as an early selection due to a competitive mechanism between CBHs and external matter contained in clusters and superclusters.

However, if the Swiss-Cheese model were always valid such a selection would never have been achieved; instead, in a more realistic situation, the model holds only approximately so then we have to expect galaxies inside and outside the void due to the deviations from the spherical symmetry and to the perturbations of the CBH mass.

The aim of this paper is to discuss how to test the CBH model. It needs to be stressed that the current observational samples of data (void surveys, morphological classification of void galaxies) (Rojas et al. 2005) are too small to allow any conclusive test of any model for voids. But it was observed in (Stornaiolo 2002), the CBH model can be tested principally through the lensing effect of a CBH on the images of the background galaxies. In this paper we analyze in detail the observational implications of the gravitational lensing effects and we discuss whether they might be actually observable or not.We argue that our qualitative results, i.e. the expected trends of the considered observables, can be trusted against future observations.

This paper is structured as follows. In Section 8.2 we present the simulations which were performed in order to derive the possible gravitational signatures of the CBH, summarized in Section 8.3. In Section 8.4 we discuss whether such effects may or may not be observed in existing or ongoing surveys. In Section 8.5 we draw some conclusions. In a subsequent paper we shall discuss the weak lensing effects possibly induced by a CBH on a background galaxy distribution.

8.2 The simulations

In order to evaluate the gravitational lensing effects induced by a CBH located in the center of a void we produced two sets of simulations: one set for a reference unperturbed universe, *i.e.* for a void in an otherwise uniform universe (without the CBH) and a second set obtained from the previous by adding a CBH in the center of the void. In both cases we assumed, for simplicity, an Einstein-de Sitter cosmological background model, with $\Omega_M = 1$, $H_0 = 100h\ \mathrm{Km\,s^{-1}\,Mpc^{-1}}$ and no cosmological constant. Of course, we should consider that in the last few years evidences for acceleration of the cosmological expansion have accumulated to a level that they can't be ignored, especially when attempting to predict observational effects. Nevertheless, it is widely accepted that in the nearby universe the substantial degeneracy in the predictions derived from different cosmological models cannot be removed. In our case, the choice of the cosmological model only affects the relation $D = D(z, H_0, \Omega_M, \Omega_\Lambda)$ between the distance D and the redshift z of the background galaxies which is involved in the calculation of the apparent magnitudes and of the angular positions. Using approximate expressions (Kantowski and Thomas 2001) it can be seen that in a redshift range $[\,0, 0.4]$ the maximum deviation from the standard Einstein-de Sitter model would be of the order of 10% (see for example figure 1 in (Choudhury and Padmanabhan 2005) for a graphical estimate of this calculation for the $\Omega_M = 0.3$, $\Omega_\Lambda = 0.7$ cosmology). Such deviation would affect the numerical results of our simulations but not the typical qualitative signatures that can be derived from them. On the other side, introducing a more realistic cosmological background model would strongly complicate the computational aspect of our simulations. In the light of these considerations we decided that the Einstein-DeSitter model is accurate enough for our purposes. As mentioned, the reference universe is described by an Einstein-de Sitter background containing a spherical void of radius R_{void} located at the comoving distance D_{void} from the observer which does not produce any detectable deflection or magnification effects on background galaxies (as pointed out in (Amendola et al. 1999)). The perturbed universe is instead described by a Swiss-Cheese model, with a CBH in the center of the void, which, as above, has comoving radius R_{void} and is located at the comoving distance D_{void} from the observer. The CBH has a mass given in equation (8.1). It needs to be stressed that while the background universe is described by a FLRW metric, inside the void region holds a Schwarzschild metric. The boundary conditions at the transition between the two regimes are discussed in (Kantowski 1969). In particular, it is well known that light crossing an expanding void (or in general a non-linear perturbation) is also affected by a net red-shift. Following the calculations in (Kantowski 1969) we can estimate this effect in the case of photon crossing a hole of the Swiss-Cheese model which contains a CBH. The variation Δz in the photon redshift is then given by

$$1 + \Delta z = 1 + O\left(\frac{R_{void}}{c/H_0}\right)^3 + O\left(\frac{R_{void}}{c/H_0}\frac{R_S}{a}\right) \tag{8.3}$$

where a is the impact parameter of the photon and R_s is the Schwartzschild radius of the CBH. If we set $R_{void} \sim 10h^{-1}\mathrm{Mpc}$, $c/H_0 \simeq 3000h^{-1}\mathrm{Mpc}$, $R_S \sim 0.2h^{-1}\mathrm{Kpc}$ in (8.3) we obtain that the variation in redshift remains under one part over 10^3 if:

$$a \geq \frac{10}{3} R_S \tag{8.4}$$

In the following we assume that the condition (8.4) is always verified and therefore that changes in the sources redshift, due to the propagation of light across the edge of the void, is negligible.This assumption is reasonable also because this change in sources redshift is about one order of magnitude lower than the typical error of photometric redshift estimates (D'Abrusco et al. 2007). In conclusion we describe the void-CBH system as a Schwarzschild lens enclosed in a Einstein-de Sitter cosmological model. More in particular we simulated a slice of universe covering a solid angle defined by the void diameter and a redshift range comprised between the far edge of the void and $z = 0.4$. This conical volume was populated with a randomly distributed (in the comoving frame) galaxy population drawn from the r-band SDSS (Sloan Digital Sky Survey) luminosity function for the field (Blanton et al. 2003). Evolutionary effects induced by the redshift on the luminosity function were neglected. Furthermore we treated galaxies as material points since their physical extension is not relevant to the following discussion. The lensing effects induced by the CBH were then derived in the weak lensing approximation by assuming that the light from the sources passed at a distance from the CBH much larger than its Schwarzschild radius (about $30\,\mathrm{arcsec}$ for $M = 10^{14}\ M_{\odot}$ and $D_{void} = 50\,\mathrm{Mpc}$). The deflection angle produced by the CBH for radiation approaching with impact parameter ξ is given by

$$\hat{\alpha} = \frac{4GM}{c^2\xi} \tag{8.5}$$

For a source at distance D_s from the observer, the Einstein angle can be written in the form:

$$\hat{\alpha}_0 = \sqrt{\frac{4GM}{c^2}\frac{D_{ds}}{D_{void}D_s}} \tag{8.6}$$

where D_{ds} is the distance of the source from the lens.
If $\hat{\beta}$ is the unlensed angular position of the source with respect to the observer, we know that the effect of the lens will be the creation of two images of the source with angular positions:

$$\hat{\theta}_{1,2} = \frac{1}{2}\left(\hat{\beta} \pm \sqrt{4\hat{\alpha}_0^2 + \hat{\beta}^2}\right) \tag{8.7}$$

and with magnifications given by:

$$\mu_{1,2} = \frac{1}{4}\left(\frac{\tilde{\beta}}{\sqrt{\tilde{\beta}^2+4}} + \frac{\sqrt{\tilde{\beta}^2+4}}{\tilde{\beta}} \pm 2\right) \tag{8.8}$$

where $\tilde{\beta} = \hat{\beta}/\hat{\alpha}_0$.

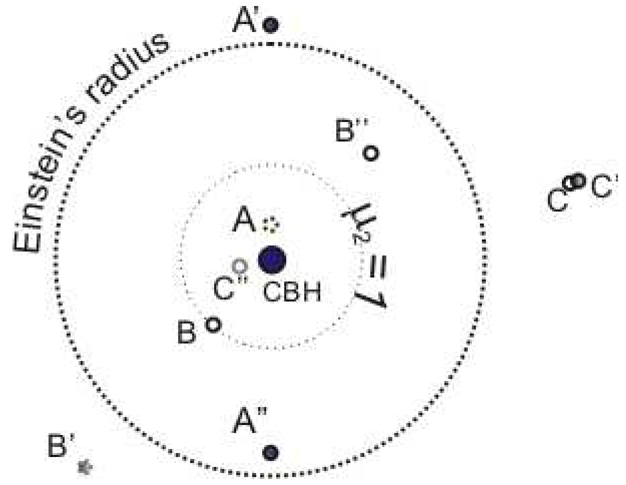


Figure 8.1: *Schematic layout of how the CBH gravitational effects affect the formation of double images as a function of the angular distance of the source from the CBH (see text for a more detailed explanation).*

Simulations were then performed for different values of Ω_{cbh} and R_{void} (0.05 to 0.4 with step 0.05 and from 10 to 20 Mpc, step 2, respectively).

8.3 Qualitative description of observable quantities

The simulations showed that the CBH leaves three different types of signatures on the background galaxy distribution. In order to better quantify what happens we refer to Fig. 1 which shows the images produced in three different relative positions of source and lens. When the "real" object (A) is very close to the CBH, we have the formation of two images, namely A' and A'' which are respectively outside and inside the Einstein Radius and very close to it. Both images are brighter than the unlensed image and A' is brighter than A''. Then when the source moves away from the CBH, the amplification factor with respect to the secondary image tends to 1. The dashed inner circle marks the position of such locus. If the source B lays on this circle, then it will produce two images B' and B''. B' is outside of the Einstein Radius at a larger distance than A' and is brighter than B, while B'' lays inside the Einstein Radius closer to the CBH than A''. Finally let us consider the case of a source C which, if unperturbed would fall outside of the Einstein Radius. Also in this case we shall see two images C' and C''. The first one will almost coincide with the position of C and will have an almost identical brightness, while C'' will be almost invisible and very close to the CBH. When a random distribution of background galaxies is considered, the overall effect of the lensing will be the formation of 4 different areas on the sky, namely the regions A, B, C and D, shown in Fig. 2.

The inner circle A, which we call 'blind spot', is the region where no image can be detected due to the increasing demagnification of secondary images when they move towards the center of the void. The size of the blind spot depends on the mass of the CBH and on the density of the background galaxy distribution.
The annulus B is characterized by the presence of a large number of secondary images, and it is the second observable feature associated with the CBH.
The third zone C, is an annular zone that we call the 'deficit zone'. It encompasses the average Einstein Radius of the galaxy sample and is characterized by a relatively low number of background galaxy images. It can be understood reminding that, at this angular distance from the center of the void, one can find only those primary and secondary images which originates from sources having angular position falling well within the Einstein Radius. Therefore, in this annulus it is expected to find a lack of both primary and secondary images: as the source moves away from the Einstein Radius, the primary image tends to coincide with the original position of the source while the secondary image becomes fainter and moves more and more towards the CBH.
Finally, the D zone does not present any particular feature produced by lensing, and coincides with the homogeneous background galaxy distribution.

In total we run 48 sets of simulations assuming the void distance at $50\ Mpc$ (*i.e.* matching the distance of the nearest void) and covering a grid defined by: $\Omega_{CBH} = 0.05 \rightarrow 0.4$ with step 0.05, and $R_{void} = 10 \rightarrow 20\ Mpc$ with step 2 Mpc. For each grid point the procedure was iterated 1000 times randomly changing at each iteration the positions of the background galaxies. Each simulation produced a catalogue of galaxy positions and magnitudes and each group of simulations was then used to derive average quantities.

8.4 Results

The main observational signatures left by the CBH can be summarized as follows.

8.4.1 Blind spot

For every simulation, we determined an estimate of the angular radius θ_{blind} of the blind spot, defined as the minimum angular distance from the CBH of the secondary images (brighter than $m_{lim} = 23.5$) and then, at each simulation grid point we took the average value over the 1000 simulations. In Table 1 we list some representative values.

As expected, the size of the blind spot, increases with Ω_{CBH} and with R_{void}. Assuming, for instance, a typical value of $30\,\mathrm{arcsec}$ (cf. Table 1), it is apparent that the blind spot should be rather difficult to observe. In fact, while it should be easily detectable

Ω_{CBH}/R_{void}	**10**	**12**	**20**
0.05	16±5	25±7	36±8
0.2	33±8	36±9	53±10
0.4	39±8	44±9	69±12

Table 8.1: *Average value of θ_{blind} (in arcsec) as a function of Ω_{CBH} and R_{void} (in Mpc). The quoted errors are the r.m.s. of the 1000 individual simulations.*

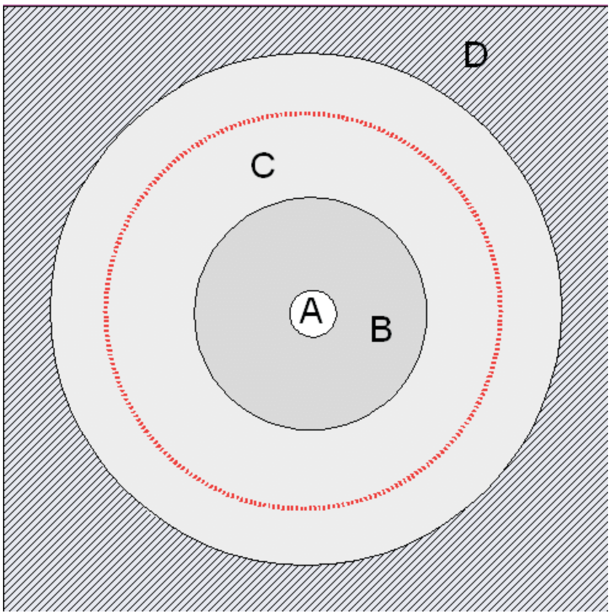


Figure 8.2: *The three regions described in the text. Region A (greatly enlarged to make it visible) is the blind spot; Region B: region where we expect the excess of secondary images; Region C: deficit region (the circle inside the C region shows the average Einstein Radius).*

in very deep number counts, its size is of the same order of the average angular separation between bright galaxies at intermediate redshift. Possible effects connected with the distortion induced on extended background objects which happen to fall near the line of sight of the CBH will be addressed in a forthcoming paper.

8.4.2 Galaxy number counts and radial profile

The presence of the CBH affects the number counts. In order to estimate the size of such effect and in absence of a priori information on the size of the void we adopted the following procedure. First we introduced an annular zone (defining an inner and outer region) centered on the blind spot. The radius of the zone was then found by maximizing the difference between the average galaxy counts in the inner and outer regions.

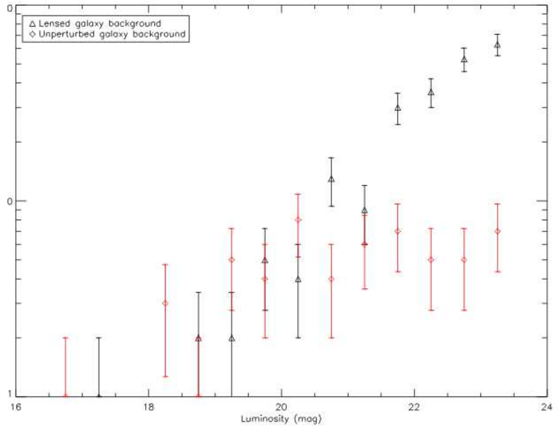


Figure 8.3: *The figure shows in red the average number counts as a function of the apparent magnitude in the r band, obtained, respectively, in the outer region (see text) and in black those obtained in the inner region.*

As it can be seen in Fig. 3, the average number counts associated with the inner area show a systematic difference with respect to those extracted from the background. This effect however becomes significant only at magnitudes fainter than ~ 21.0.

In order to quantify such difference we performed a Kolmogorov-Smirnov test on both distributions of points. Each data set was split into two parts including galaxies brighter or fainter than 21 mag, respectively. The brighter parts of the distribution do not present any statistically significant difference, while for the fainter parts we derived a probability higher than 98%, that the two samples are drawn from different populations.

As it was discussed in the previous paragraphs, the presence of a CBH induces a typical pattern in the number counts radial profiles. Such pattern is characterized by a peak in the range of distances intermediate between the blind spot radius and the average Einstein Radius (caused by the secondary images concentration), followed by a dip, which corresponds to a slight underdensity of objects and than at distances comparable with the Einstein Radius it raises up again to smoothly reach the value expected for the background galaxy distribution.

In Fig. 4a we show the number counts profile extracted from the simulation grid–point at $\Omega_{CBH} = 0.2$ and $R_{void} = 12$ Mpc. The first point of the profile is located in the blind spot and is followed by an isolated peak which rapidly falls in the dip.

The only two other examples of possible radial profiles obtained when not placed on the center of the void are illustrated in Figg. 4b and 4c. The first image, which can be roughly described by an initial peak, larger than the previous case, located at the first steps of the profiles followed by the dip, is the typical pattern generated when the

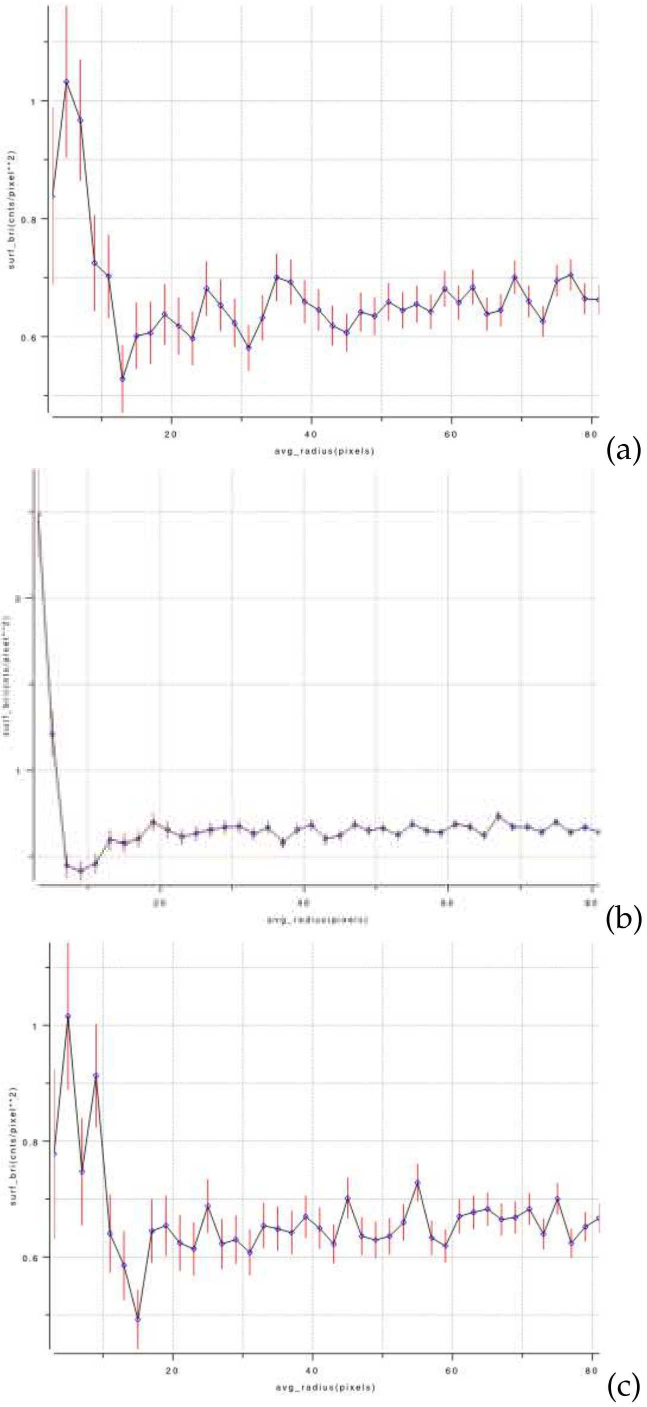


Figure 8.4: *Galaxy number counts radial profiles obtained by integrating over annular regions centered in different regions. Panel a - centered on the blind spot (zone A); Panel b - center falls in the secondary images overdensity region (zone B); Panel c - center falls in the deficit region (zone C).*

central point of the radial profile is positioned inside the secondary images overdensity; the second profile, characterized by the presence of an initial low spot followed by two distinct peaks and the dip, is reproduced when the center of the radial density profile lays out of the circular overdensity created by the secondary images, and inside the deficit annulus.

8.4.3 Multiple images

The third signature comes just from the pairs of double images produced by the CBH. We expect angular separations for these pairs of the order of two times the average Einstein's angle of the sample. This means angular separations of the order of some arcminutes. This forecast is particularly striking because if these doubles really exist, the only way to observe them practically is to have an estimate of their wide angular separation. We measured the distribution of the angular separation between the double images produced by the CBH obtaining the two points angular correlation function $w(\theta)$ for the simulated galaxy distribution. A standard Landy–Szalay (Landy and Szalay 1993) estimator was used:

$$w(\theta) = \frac{\langle DD \rangle + \langle RR \rangle - 2\langle DR \rangle}{\langle RR \rangle} \tag{8.9}$$

where $\langle DD \rangle$, $\langle DR \rangle$, $\langle RR \rangle$, are pair counts in bins of $\theta \pm \delta\theta$ of: data–data, data–random and random–random points, respectively. The statistic has been demonstrated to be close to a minimum of variance estimator and to be robust with respect to the number of random points (Kerscher et al. 2000). In Fig. 5 we show the above defined correlation function obtained for the simulation grid-point at $\Omega_{CBH} = 0.2$ and $R_{void} = 12$ Mpc; a well defined peak corresponding to an angular distance of 6 arcmin, comparable with the average diameter of the secondary images overdensity region (Zone B), is visible. Moreover, a slight anticorrelation is found at distance greater than 10 arcmin, while for a random distribution of points no correlation of any sort should be detected. It needs to be stressed that the peak observed at very short angular separations is an artifact produced by the fact that in our simulations galaxies are assumed to be point-like and would disappear if galaxies were approximated with extended objects.

8.5 Discussion and conclusion

The results of our simulations which were based on a number of assumptions concerning the cosmological model can be regarded as quite general and robust in their final predictions. In fact they clearly show that the presence of a CBH close to the center of

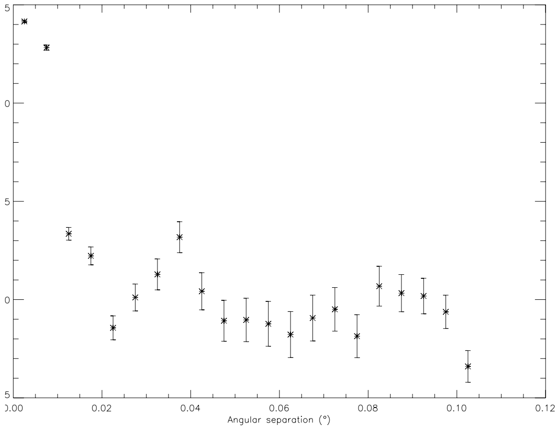


Figure 8.5: *Angular 2–point correlation function derived from the simulated galaxy distribution (see text).*

a void would leave unambiguous signatures on the background galaxy distribution as a result of the gravitational lensing properties of the CBH. Replacing the Einstein - De Sitter model used for our simulations with more realistic cosmological models could produce variations in the distribution of galaxy redshifts up to 10%. Nonetheless the observables considered are robust to such fluctuations at least at a qualitative level, so that detection of CBH would prove reliable. In other words, we trust that the CBH lensing signatures give clear indications of expected effects. Unfortunately, such signatures can be detected only at faint light levels, i.e. at magnitudes fainter than the completeness limit of most existing photometric surveys which include voids in their field. Furthermore, the only survey which, at least in theory, should be deep enough to allow at least the radial profile test above described (namely the Sloan Digital Sky Survey, cf. (Stoughton et al. 2002)) does not satisfactorily cover any previously known void. It needs to be stressed however that our results clearly show that such tests will be possible on any of the planned deep digital surveys which will become available in the near future (cf. for instance the VST extragalactic survey (Capaccioli et al. 2003)).

Chapter 9

Groups of galaxies

What's in a name?
That which we call a rose
By any other name
would smell as sweet.

Romeo and Juliet, W. Shakespeare

9.1 Introduction

Groups of galaxies can be found in a variety of configurations going from the most compact ones to the more dispersed ones. In any case they are characterized by a low velocity dispersion around the centre of mass. Theoretical models suggest that a combination of high density regions and low velocity dispersion favours both strong and weak gravitational interactions, together with the occurrence of phenomena of cannibalism and merging between galaxies. This is the reason why the groups of galaxies are ideal places for the study of these interesting events. In addition, diffused gas haloes are observed in groups of galaxies, indicating the presence of a potential well where galaxies are placed, while providing strong clues of the presence of a fraction of dark matter hidden in the haloes of groups through the measurement of their spatial extension and temperature. Groups of galaxies are of vital importance also from the cosmological point of view, because their observed properties are strictly related to cosmological parameters, such as the baryon fraction of the Universe or the cosmological density parameter Ω. From a theoretical point of view, there are many questions still without answer, regarding their formation, their connection with the large scale structure of the galaxy distribution and the influence that groups have on the processes of formation and evolution of the galaxies. Shakhbazian's groups of galaxies are a class of structures observed and catalogued by Armenian astronomer Shakhbazian. These peculiar groups of galaxies have been only marginally studied, and still today very little about their physical nature and their evolutionary state is known. Shakhbazian' groups can be found in a wide range of redshift: $0.00 < z < 0.3$, and are characterized by heterogeneous observational features. An interesting and still doubtful piece of information about this class of observed groups is whether they represent true physical structures: if this is the case, the determination of their global properties regarding the content of galaxies, their evolutionary stage compared to other class of objects (cf. Hickson groups) and the relation of the observed properties with the environment can provide useful insights into the physics of processes driving the formation of gravitationally bound systems of galaxies.

9.2 Groups and clusters of galaxies

The formation of groups and clusters of galaxies, at very large scale, can be efficiently modelled as a positive perturbation of a homogeneous density distribution propagating through the Universe, i.e. a region of the space-time with locally a matter density greater then the average value of the surrounding Universe: $\rho_{per} > \bar{\rho}$. In the hypothesis of a constant and finite spherical overdensity, the evolution of such perturbation is described by the simple equation

$$\ddot{R} = -\frac{GM}{/}R^2 \tag{9.1}$$

The perturbation will expand following the Hubble expansion of the Universe until it reaches its maximum size. Then, decoupling from the Hubble stream occurs and, after a characteristic time τ_{ta} called *turnaround time*, the gravitational collapse of the matter contained in the region of the perturbation and the process of virialization start. The *turnaround time* can be expressed in terms of the matter density at the the *turnaround* ρ_{ta}:

$$\tau_{ta} = \left(\frac{32G\rho_{ta}}{3\pi}\right)^{-1/2} . \tag{9.2}$$

Once the central region of the perturbation has collapsed, the surrounding shells undergo the same process in a time τ_{ta} depending on the average density of matter inside the shell. This process is defined *secondary infall*. In the case of cosmology $\Omega = 1$, simple scale laws for the matter contained in the overdensity can be determined:

$$M_{ta} \sim t^{2/3} \qquad R_{ta} \sim t^{8/9} \qquad \bar{\rho}_{ta} \sim t^{-2} \, . \tag{9.3}$$

The turnaround radius R_{ta} in the approximation of a *top-hat* perturbation can be calculated using equation 9.2:

$$R_{ta} = \left(\frac{8GM_{ta}t_0^2}{\pi^2}\right)^{1/3} , \tag{9.4}$$

which is notably independent of Ω_0, and with t_0 the age of the Universe. The richest and most compact structure of galaxies are called clusters and superclusters, and have been intensively studied since they represent the largest overdensities observable in the large scale structure of the Universe. The study of clusters and superclusters of galaxies provides information regarding a large number of cosmological and astronomical questions, ranging from the dynamics of large gravitational structures formations to the quantity and composition of visible and dark matter in the densest regions of the Universe, the evolution and characteristics of galaxies as a function of the environment. Superclusters are the largest observed systems of galaxies, and even if not completely virialized, they extend up to $100h^{-1}$ Mpc and there is a precise correlation between their position and the position of the knots of the web-like large scale structure of the Universe, since these very high contrast overdensity appear to be placed in the regions of space where filaments (long almost one dimensional pattern formed by galaxies) meet and intersect. Cluster of galaxies are mostly virialized systems gravitationally bound defined according to Hubble's definition (Hubble 1958) as systems where can be found at least 30 galaxies brighter than $m_3 + 2$ magnitudes, where m_3 is the magnitude of the third most brilliant galaxy member of the cluster, within a radius $R \simeq 1.5\mathrm{h}^{-1}$ Mpc from their centre. The number of galaxies placed within this radius represents the Abell's richness of the cluster. The typical velocity of a galaxy inside a cluster is $750\ \mathrm{km\ s}^{-1}$, measured as the average velocity dispersion along the line of sight, corresponding to a virial mass encircled in the Hubble's radius of $\sim 5 \times 10^{14}\ \mathrm{h}^{-1}\ \mathrm{M}_{\odot}$. Rich clusters also contain intra-cluster medium composed by hot plasma whose distribution follows almost perfectly the shape of the gravitational potential well containing the galaxies members of the system; the density of this diffused matter has been estimated as $\sim 10^{-3}\ \mathrm{electrons\ cm}^{-3}$, with average temperature $\sim 5\mathrm{keV}$. The presence of the plasma is confirmed by the observation of diffused X ray emission caused by the thermal *bremsstralhung*, with typical values of the flux of $\sim 2 \div 14\ \mathrm{keV}$. Rich clusters are very rare objects, since their estimated spatial density is $\sim 2 \times 10^{-7}\ \mathrm{clusters\ Mpc}^{-3}$, 5 order of magnitudes lower than the density of bright field galaxies ($\sim 10^{-2}\ \mathrm{galaxies\ Mpc}^{-3}$).

9.3 Compact groups of galaxies

Groups and poor clusters of galaxies provide a natural and smooth extension of the rich clusters briefly described in the previous paragraph towards smaller richness, mass, luminosity and dimension. An advantage of low richness systems is that, even if more difficult to detect due to the lower luminosity and smaller overdensity of the projected galaxy distribution, they are quite abundant ($\sim 55\%$ of all the galaxies are contained in the Universe are members of this class of systems). The mass of a group (and consequently the richness and the total luminosity) are related to the initial density contrast of the primordial perturbation from which the stage of evolution can be inferred. For example, loose groups of galaxies are identified with density contrast between 20 and 80, so that a they are yet experiencing the phase of collapse of their evolution. In particular, compact groups of galaxies are easily detected thanks to their high spatial density $(10^{-3} \sim 10^{-5})\ \mathrm{h}^3\,\mathrm{gal}\,\mathrm{Mpc}^{-3}$, since they contain from 2 to 30 galaxies inside a conventional radius $R = (0.1 - 1)\ \mathrm{h}^{-1}\,\mathrm{Mpc}$. For this reason, this class of groups is of great importance for the comprehension of the effects of the environment on the evolution of galaxies: the high spatial density favour interactions between galaxies so that remnants and current evidences of ongoing or past mergers are visible in a large fraction of compact groups. In other words, the evolution and the behaviour of galaxies in compact groups are influenced non only by the global properties of the structure (similarly at what happens for clusters and superclusters), but also by the peculiar features of the single members.

9.3.1 Projected overdensity and real structures

Since only three components of the vector of the 6-dimensional phase-space associated to each galaxy belonging to a group can be measured (2-dimensional position of the galaxy on the celestial sphere and the component of its velocity along the line of sight), groups of galaxies are subject to projection effects whose worst effect is the misidentification of apparent overdensity as real physical structures. Measures of the redshift of galaxies that supposedly belong to a group has often showed that a certain fraction of galaxies have significantly different redshift from that of the large majority of galaxies, which can be identified as the redshift of the whole group. The existence of these objects has raised concerns regarding the possibility that these members are objects that accidentally are projected along the same line of sight of the considered group, or that there are other phenomena not still noticed that can generate such situations (for example, amplification of the background galaxies due to gravitational lensing). The key point is to understand whether the total number of objects with discordant redshift is compatible or not with the effects of accidental projection. It has been shown that the probability that a group of galaxies is misidentified due to chance projection effect is a decreasing function of the richness of the group. The combination of observational data of compact groups of Hickson and compact groups visible in the South hemisphere with Montecarlo simulations (Iovino and Hickson 1997) allows to conclude that for almost all the groups, the number of objects with discordant redshift is consistent with the effects of an accidental projection. This and other results, though not entirely addressing the issue, give strong clues in favour of a correct interpretation of discordant redshifts and chance projection effects. Nonetheless, a projected surface density higher than the average is not sufficient to assert the physical nature of an agglomerate of galaxies, since it constitutes only one of the necessary conditions for this to occur. Other observables, as the presence of infrared emission, tidal tails or distorted velocity profiles (typical signs of interaction between galaxies) and presence of gas with diffused X ray emission strongly support the reality of a gravitationally bound system of galaxies.

9.3.2 Dynamical properties

Several studies on compact groups of galaxies have shown that their crossing time is $t_{cr} = 0.02H_0^{-1}$, where $H_0^{-1} = 2 \times 10^{10}$ years is the Hubble's time, suggesting that these groups are collisional systems. The mass/luminosity ratios observed in compact groups (M/L $= 50h$ in solar unity) are intermediate between the value observed in single galaxies and rich clusters. Since M/L ratio for isolated galaxies is approximately $7h$, galaxies members of such groups contain only 15% of the total mass of the group. This can be explained by the fact that the greater fraction of the mass is associated with global features (like the intra-cluster plasma and gas and the dark matter halo) of the whole system and not to the single constituent galaxies. Because of the reduced number of galaxies belonging to compact groups, the measurement of velocity of the members and physical separations between galaxies suffer of several sources of error, so that large and homogeneous samples of groups are necessary to perform a robust statistical evaluation of such quantities. A typical value of velocity dispersion derived for compact groups is $\sigma \sim 250\,\mathrm{km}\,rms^{-1}$ (Hickson 1997), similar to the value found for loose groups of galaxies and significantly smaller than rich cluster dispersion velocity.

9.3.3 Shape and orientation of groups of galaxies

Spatial distribution of galaxies within compact groups provides useful information on the nature of these systems, since if groups are mainly chance projections configurations of galaxies not physically bound the observed distributions would be consistent with a random distribution. On the other hand, in the case members of group are interacting inside a gravitational potential well, their spatial distribution would show typical signatures reflecting the origin and the physical evolution of the system. In particular, the three dimensional shape and axial orientation of the galaxy distribution of compact groups are sensible diagnostics of the physical features and dynamical state of the structure. Even if still discussed, several studies have shown that compact groups of galaxy have spatial distributions not consistent with the hypothesis of chance projection effects as cause of the apparent overdensity of galaxies. Using static simulations, Hickson (Hickson et al. 1992) concluded that the observed spatial distributions of galaxies of a sample of compact groups are explainable with a prolate three-dimensional shape;

the same conclusion was reached by Oleak (Oleak et al. 1995) studying 95 Shakhbazian groups. The shapes derived by Hickson are consistent with those found from projected dynamical simulations of compact groups seen as substructures of loose rich groups. If the intrinsic shapes of the compact groups are connected to their formation process, a correlation between the orientation and the environment of such systems should be observed. The analysis of the environments and shapes of Hickson groups carried out by Palumbo (Palumbo et al. 1995) showed that no correlation is found between the major axis orientation and the spatial distribution of galaxies surrounding the groups, in conflict with the previous statement. If compact groups of galaxies are effectively real physical structures, a density profile concentrated in the core of the structure is expected, similarly at what is observed in clusters of galaxies. Also in this case, groups are formed by a few galaxies that do not provide a sufficient statistics to derive meaningfully the projected density profile of matter. Some works indicate that the stacked density profile of a sample of Hickson groups is consistent with the expectation of a centrally peaked profile. Montoya (Montoya et al. 1996) derived density profiles of single Hickson groups which clearly show centrally concentrated shape, suggesting this class of compact groups of galaxies are characterized by a unique spatial scale which could be explained invoking hierarchical clustering scenario for structure formation. Many authors in last decades have stressed, using simulations and spectroscopical observations, the existence of correlations, more or less pronounced, between the apparent axial ratio q of the three-dimensional shape of single groups of galaxies and other observational quantities, namely the richness, velocity dispersion and evolutionary stage. Nonetheless, these claims have not yet been confirmed on a statistically sound basis.

9.3.4 Galaxies in compact groups

Several studies of the morphology of the galaxies in compact groups have provided a picture according which the fraction of late-type galaxies f_s in these systems seems to be significantly smaller than for the field galaxy population, with values of $f_s \approx 0.49 \div 0.59$ respect to a value of ~ 0.82 for the field. The fact that the population of galaxies within compact groups is largely less assorted in terms of morphology than the field population with a clear preference for early type galaxies, rules out the hypothesis that such systems are merely effects of chance projection on the sky of not interacting galaxies, since if this was the case the morphological mix should have been the same in both environments. Correlations between morphology and other properties of the groups could explain this observational circumstance; at the moment, the strongest observed correlation involving the morphology of galaxies found in compact groups appears to be the correlation between morphology and velocity dispersion σ, indicating growing fraction of early-type galaxies for growing velocity dispersion. A crucial point for the comprehension of the formation and evolution of compact groups is the lack of the well known density/morphology relation observed in rich clusters, suggesting that the velocity dispersion σ of galaxies (more than the spatial density, i.e. the richness of the system) is the real physical parameter driving the dynamics of the evolution of member galaxies. This simple static scenario does not take into account other important mechanisms which can as well influence the evolution of the system, namely the possible contamination due to interlopers on the late-type fraction f_s measure in spatially ill-defined groups (systems whose borders are not well defined because of poor statistics), or differences of evolutionary stage between compact groups that could affect the observed mixture of morphological types.

Submitted as: Capozzi, D., D'Abrusco, R., Longo, Paolillo, M., De Filippis, E. 2007 – *"Shakhbazian groups din the SDSS sample"*, Monthly Notices of the Royal Astronomical Society

Chapter 10

Application III: Shakhbazian groups in the SDSS

Abstract

We discuss the properties of the subsample of Shakhbazian groups (SHKGs) covered by the Sloan Digital Sky Survey Data Release–5 (SDSS-5). SHKGs were defined as compact groups of compact galaxies and subsequent studies have shown that their members probe an environment with characteristics which are intermediate between those of loose and very compact groups. Using the SDSS-5 spectroscopic data and the photometric redshifts derived in (D'Abrusco et al. 2007) we searched for overdensities in the 3-D space and then used a variety of diagnostic tools to investigate its nature (physical or chance alignment) and to identify candidate members of the groups. These diagnostics include, radial galaxy counts profile, density map, individual luminosity function, etc. For each confirmed group we derived a complete set of global parameters, such as richness, size, mean photometric redshift and the fraction of early type galaxies. Our study confirms that almost all groups are physical entities with richness in the range 3–13 and properties ranging between those of loose and compact groups. The global properties of SHKG support the existence of two subclasses of groups, the first one being formed by compact and isolated groups and the second formed by compact structures embedded into more extended ones. SHK groups appear to be abnormally rich in early type galaxies, and there is a dependence of the morphological composition on the density of the environment which respects the morphology-density relation. Our results, and in particular the high content of early galaxies, cast some doubts on the evolutionary scenario which predicts groups to evolve from loose, to [core+halo], to compact configurations.

10.1 Introduction

In spite of the fact that small groups of galaxies are the most common extragalactic environment (Tully 1987), their physical properties, origin and evolution are still poorly understood. This is particularly true in the low mass density regime, id est in that environmental range which bridges the field with the richest (clusters) or most compact (e.g. Hickson Compact groups) structures. This is mainly due to observational selection effects which render the poor and less compact structures much more difficult to detect and study, especially at intermediate and high redshift. The lack of statistically complete and well measured samples of groups at different redshifts is very unfortunate since they are needed to constrain all models for the formation and evolution of cosmic structures. For instance, in the hierarchical models it would be crucial to measure the epoch of group formation and to discriminate whether it does or doesn't exist a down-sizing effect for which high-mass groups at high z form before lower-mass ones at lower z. Other crucial issues are related to the dynamical status of groups, to their relaxation time and on how the evolution of the structure affects that of their galaxy members. Groups of galaxies are very numerous and show spatial densities of about $10^{-4} \div 10^{-5}\ \mathrm{h}^3\,\mathrm{gal\,Mpc}^{-3}$. Their gravitational potential wells are about as deep as that of individual galaxies, whose random velocities in such systems are a few hundreds of $\mathrm{km\,s}^{-1}$. Under these conditions, galaxies strongly interact both among themselves and with the global potential of the group. Therefore, groups of galaxies are collisional systems evolving toward virial equilibrium through collisions and collisionless interactions among their member galaxies. Because of this, they are also the ideal laboratory for studying gravitational interaction phenomena among galaxies, like merging and ram pressure stripping and constitutes environments that evolve themselves during the phases of collapse and virialization, while affecting the properties of their member galaxies.

According to hierarchical formation theories,as structures of galaxies increase their dimensions, they decouple from the Hubble flow and then collapse and virialize. Therefore, groups of galaxies, as structures of galaxies in general, form as a consequence of the gravitational collapse of galaxies dwelling small and localised over-densities. It's possible to consider these over-densities as a constant finite spherical perturbation in a homogeneous Universe, evolving in the following way: it expands following the Hubble expansion then decouples from it, turns around and then collapses. Subsequently a virialization process settles the structure, in relatively brief times, on an equilibrium configuration.

Numerical simulations have extensively shown that structures of high spatial densities undergo fast dynamical evolution and merging (Barnes 1985, Barnes 1989, Mamon 1986, Bode et al. 1993, Diaferio et al. 1993, Diaferio et al. 1994, Governato et al. 1996). The expected number of groups inferred through theoretical estimates of crossing times is in contrast with the observed one. The Secondary Infall Scenario suggested by Gunn (Gunn and Gott 1972) and then reconsidered by Mamon (Mamon 1996, Mamon 1999, Mamon 2006), allows a secondary collapse of galaxies surrounding the formed structures. This secondary aggregation provide the structures a way to increase their lifetimes, according with the observed number of groups of galaxies. Groups of galaxies seem to evolve assuming different configurations: loose, [core+halo] and compact.

Compact configurations form in the last phases of the evolving process and their final by-product is a giant elliptical galaxy surrounded by a hot X-ray emitting gas halo. This relic is known as Fossil Group.
The paper is structured as it follows. In Sect. 10.2 we shortly summarize the main properties of Shakhbazian groups as they are known in the literature and in Sect. 12.2 we describe the data used in the analysis. In Sect. 10.4 we describe in some detail the method used to derive the observable quantities, while in Sect. 10.5 we summarize the main properties of each group in our sample. In Sect. 10.6 we discuss the global properties for all groups in our sample and in Sect. 12.7 we finally draw our conclusions.

10.2 Shakhbazian groups

Shakhbazian groups of galaxies (hereafter SHKG) were originally defined as compact groups of mainly red compact galaxies and selected by visually inspecting the printed version of the First Palomar Sky Survey (POSS) using rather empirical and ill defined selection criteria:

- They must contain 5-15 member galaxies.
- Each galaxy's apparent magnitude in the POSS red band must be comprised between $14^m \div 19^m$.
- They are compact, id est the relative distances of the member galaxies must be $3 \div 5$ times the characteristic diameter of a member galaxy.
- Almost all galaxies must be extremely red; there must not be more than $1 - 2$ blue galaxies.
- Galaxies are compact (high surface brightness and border not diffuse).
- The group must be isolated.

The search lead to a long series of papers (Shakhbazian 1973; Shakhbazian & Petrosian 1974; Baier et al. 1974; Petrosian 1974; Baier & Tiersch 1975, 1976a, 1976b, 1978, 1979; Petrosian 1978) identifying a total of 377 groups which, due to the poor resolution of the POSS and to the compactness requirement, appeared initially to be strongly contaminated by stars mistaken for galaxies and, furthermore, resulted affected by many systematics. For these reasons, until a few years ago, the SHKG sample has not received as much attention as other more homogeneous and better defined samples such as, for instance, the Hickson's compact groups (hereafter HCG) one. Detailed photometry by

(Kodaira et al. 1990) showed that in most cases, those which had been believed compact galaxies were rather normal ellipticals and S0 galaxies with slightly redder colours ($\Delta(V - R) \sim 0.2$) than field ones. Furthermore, even though some contamination by stars is indeed present, many of the objects initially suspected to be stars were found to be galaxies. The extensive and detailed studies by Tovmassian and collaborators (Tovmassian et al. 1999; Tovmassian & Tiersch 200; Tovmassian et al. 2003; Tovmassian et al. 2003a; Tovmassian et al. 2004; Tovmassian et al. 2005; Tovmassian et al. 2005a; Tovmassian et al. 2005b; Tovmassian et al. 2006; Tovmassian et al. 2007; Tiersch 1976; Tiersch et al. 1994; Tiersch et al. 1999a; Tiersch et al. 1999b; Tiersch et al. 2002) of 44 groups, have shown that SHKGs form a rather intriguing class of physically bound and moderately compact structures composed by $5 - 15$ galaxies separated by distances of the order of $3 - 5$ galactic radii and including at most $1 - 2$ blue objects. The spatial densities of SHKGs found in these works, span a wide range: from slightly higher than those of loose groups ($10 - 10^2$ $\mathrm{gal/Mpc^3}$ up to values comparable to the cores of rich clusters or HCGs ($10^4 - 10^5$ $\mathrm{gal/Mpc^3}$). Therefore the observed spatial densities of SHKGs imply that they may be at different stages of dynamical and morphological evolution and can be used both to probe the effects of environment on galaxy evolution and to constrain the formation mechanisms of low richness structures.

From the morphological point of view, SHKGs appear to be dominated by early type (E/S0) galaxies (77% against 51% in HCGs and 40% in the field, which are on average very red ($(B-V) \geq 1.0$ and $(R-K) = 2.9 \pm 0^m.6$). Detailed investigations showed also that the frequency of interacting galaxies is rather high. The fact that enhanced FIR emission was detected in only a small fraction ($\sim$ 7%) of the SHCGs (Tovmassian et al. 1999) seems to conflict against the more pronounced excess found in HCGs ($> 60\%$, (Kleinmann et al. 1987) but can be understood in part as a result of the large fraction of early type gas poor galaxies and in part as a consequence of the higher mean redshift of the SHCG sample with respect to the HCG one (0.06 against ~ 0.1). The virial radii are of the order of ~ 160 kpc and the mean mass-weighted radial velocity dispersion (=RVD) is 330 ± 170 $\mathrm{km\ s^{-1}}$, (with a wide spread, from 88.5 up to 667.1 $\mathrm{km\ s^{-1}}$), thus implying dynamical crossing times in the range from 2×10^6 yr to 1.9×10^8 yr, on average smaller than what has been found for the HCGs.

The mean value of the mass-to-luminosity ratio is $M/L \sim 32 \pm 29$ while HCGs and clusters show respectively values of about $M/L \sim 35$ and $M/L \sim 105 - 140$. Finally, SHKGs as most HGCs (Vennik et al. 1993), appear to be embedded into large, loose structures (Tovmassian et al. 2001; Tovmassian 2001, 2002; Tovmassian & Chavushyan 2000; Tovmassian & Tiersch 2001)

The fact that some of the densest SHK groups were not included in the Hickson's list can be easily explained either because they violate Hickson's isolation criterion or because, due to the compactness of the images of the member galaxies, they were just overlooked.

10.3 The data

We made use of the Sloan Digital Sky Survey - Data release 5 (hereafter SDSS-DR5) public archive. When finished, SDSS will cover 10,000 sq. deg. of the celestial sphere in 5 bands $ugriz$ (with effective wavelengths, u=3540 Å, g=4760 Å, r=6280 Å, i=7690 Å, z=9250 Å) and is complemented by an extensive spectroscopic survey providing spectroscopic redshifts for ~ 1 million galaxies. A detailed description of the survey can be found in (Stoughton et al. 2002) and (Eisenstein et al. 2001). The spectroscopic survey is almost complete for galaxies with $r < 17.7$, while at fainter light levels it includes mainly Luminous Red Galaxies (hereafter LRG) (Eisenstein et al. 2001). The cross correlation of the region covered by the SDSS-DR5 with the SHKGs positions as given in the above quoted papers produced the list of groups in Table 10.1. In this table we also provide the most relevant references for the groups which have been studied in detail.

For each of these groups we extracted from the DR5 all objects comprised within a projected distance of 3 Mpc from the assumed centroid of the group. The extracted data from SDSS are summarized in Table 10.3 and can be regarded as features (id est parameters used for analysis) or labels (parameters used for evaluating the results).

The coordinates of the centroids were taken from the above quoted Stoll et al. (1993-1997), Tovmassian et al. and Tiersch et al. papers quoted above, in order to obtain a first order estimate of the distance of the groups, whenever possible we used the spectroscopic redshift estimates available in the literature, otherwise we took the average of the SDSS-DR5 spectroscopic redshifts obtained by cross-correlating the lists of confirmed members positions with the SDSS-DR5 objects. Photometric redshifts for all the objects in these field were then extracted from the catalogue by (D'Abrusco et al. 2007) we refer to this work for all details on how these redshifts were evaluated and for a thorough discussion of their accuracy. It needs to be stressed that for all those groups which have not been the target of specific studies, the spectroscopic redshifts can be considered only as an indication since, very often, only one redshift for group is available and in many cases the corresponding object turns out to be a spurious member accidentally projected against the group. This fact explains also many of the discrepancies between the spectroscopic redshifts and the photometric listed in Table 10.1 and discussed in greater detail in the paragraph 10.5. In this paper we assumed $H_0 = 70$ km s^{-1} Mpc^{-1}.

SHK	RA	DEC	z_{spec}	z_{phot}	σz_{phot}	References
1	10:55:05.70	+40:27:30.0	0.117	0.11	0.01	Tovmassian, H. M., et al., 1999, ApJ,523,87
5	11:17:06.75	+54:55:10.3	0.139	0.15	0.01	Stoll, D., et al., 1993-1999, AN
6	11:18:49.93	+51:44:29.6	0.079	0.10	0.02	SDSS DR5
8	16:03:41.02	+52:21:13.7	0.110	0.13	0.02	Tovmassian, M., et al., 2005, A&A, 439, 973
10	14:10:51.45	+46:16:34.6	0.134	0.142	0.007	SDSS DR5
11	14:11:09.17	+44:40:11.8	0.094	0.18	0.02	SDSS DR5
14	14:25:19.75	+47:15:09.5	0.073	0.09	0.01	Tovmassian, M., et al., 2005, A&A, 439, 973
19	13:28:30.20	+15:50:25.6	0.069	0.03	0.02	Tovmassian, M., et al., 2005, A&A, 439, 973
22	15:45:43.74	+55:06:58.8	0.082	0.07	0.02	Tovmassian, M., et al., 2005, A&A, 439, 973
29	16:08:42.16	+52:26:19.0	0.035	0.18	0.02	Tovmassian, H. M., et al., 1999, ApJ, 523, 87
31	00:58:17.99	+13:54:38.7	0.187	0.19	0.02	Tovmassian, H. M., et al., 2003, Revista Mexicana de Astronomía y Astrofísica, 39, 275
54	10:40:34.82	+40:13:58.6	0.090	0.10	0.02	SDSS DR5
55	10:43:35.98	+48:22:42.4	0.143	0.14	0.01	SDSS DR5
57	10:45:26.92	+49:31:43.1	0.174	0.18	0.02	SDSS DR5
60	11:24:35.54	+40:25:14.4	0.108	0.21	0.02	SDSS DR5
63	11:29:33.90	+42:26:32.8	0.181	0.22	0.01	SDSS DR5
65	11:30:53.05	+35:01:43.9	0.152	0.17	0.02	SDSS DR5
70	12:01:20.76	+41:15:14.4	0.118	0.11	0.02	SDSS DR5
74	14:21:06.05	+43:03:46.7	0.104	0.13	0.02	Tovmassian, H. M., et al., 2005, Revista Mexicana de Astronomía y Astrofísica, 39, 275
95	08:28:37.52	+50:18:01.1	0.081	0.08	0.01	SDSS DR5
96	08:38:02.64	+52:37:46.3	0.097	0.14	0.02	SDSS DR5
104	09:27:13.60	+52:58:40.5	0.167	0.1185	0.0001	Tovmassian, H. M., et al., 2007, Revista Mexicana de Astronomía y Astrofísica, 43, 45
120	11:04:28.47	+35:52:50.6	0.070	0.14	0.02	Tovmassian, H. M., et al., 2007, Revista Mexicana de Astronomía y Astrofísica, 43, 45
123	11:44:43.31	+57:31:23.7	0.116	0.11	0.02	SDSS DR5
128	13:19:54.65	+55:45:26.8	0.145	0.16	0.02	SDSS DR5
152	09:38:50.93	+01:58:19.7	0.093	0.472	0.008	SDSS DR5
154	11:22:53.28	+01:06:46.3	0.073	0.09	0.02	Tiersch, H. et al., 2002, A&A, 392, 33
181	08:28:01.06	+28:15:56.3	0.093	0.11	0.02	Tovmassian, H. M., et al., 2004, A&A, 415, 803
184	09:08:11.29	+30:35:56.0	0.133	0.13	0.03	SDSS DR5
186	09:22:52.20	+28:55:29.6	0.077	0.09	0.02	SDSS DR5
188	09:56:59.23	+26:10:27.3	0.080	0.09	0.02	Tovmassian, H. M., et al., 2005, Revista Mexicana de Astronomía y Astrofísica, 39, 275
191	10:48:09.20	+31:28:51.7	0.118	0.14	0.01	Stoll, D., et al., 1993-1999, AN
202	12:19:47.53	+28:24:13.5	0.028	0.03	0.01	Stoll, D., et al., 1993-1999, AN
205	12:35:23.55	+27:34:45.7	0.096	0.13	0.02	Stoll, D., et al., 1993-1999, AN
213	13:45:12.24	+26:53:44.0	0.058	0.11	0.02	Stoll, D., et al., 1993-1999, AN
218	14:33:39.13	+26:41:02.6	0.095	0.10	0.02	Tovmassian, H. M., et al., 1999, ApJ, 523, 87
223	15:49:42.86	+29:09:37.5	0.083	0.10	0.02	Tovmassian, H. M., et al., 2007, Revista Mexicana de Astronomía y Astrofísica, 43, 45
229	09:00:44.51	+33:45:19.9	0.124	0.038	0.002	SDSS DR5
231	10:01:40.12	+38:18:43.8	0.146	0.15	0.01	SDSS DR5
237	11:05:29.36	+38:00:48.6	0.030	0.09	0.02	Stoll, D., et al., 1993-1999, AN
245	12:24:45.80	+31:57:17.3	0.063	0.06	0.02	Kodaira, K., et al., 1991, PASJ, 43, 169
248	13:12:16.40	+36:11:17.4	0.271	0.19	0.01	Tovmassian, H. M., et al., 1999, ApJ, 523, 87
251	13:36:54.80	+36:49:37.7	0.061	0.06	0.02	Tovmassian, H. M., et al., 2005, Revista Mexicana de Astronomía y Astrofísica, 39, 275
253	13:52:23.70	+37:30:59.7	0.073	0.09	0.02	Stoll, D., et al., 1993-1999, AN
254	13:56:19.02	+35:11:20.6	0.171	0.17	0.01	SDSS DR5
258	15:23:35.34	+32:24:39.2	0.113	0.17	0.03	SDSS DR5
344	08:47:32.54	+03:42:01.0	0.077	0.08	0.01	Tovmassian, H. M., et al., 2004, A&A, 415, 803
346	09:15:10.16	+05:14:21.4	0.135	0.14	0.02	Tovmassian, H. M., et al., 1999, ApJ, 523, 87
348	09:26:35.17	+03:26:39.7	0.088	0.10	0.02	Tovmassian, H. M., et al., 2005, Revista Mexicana de Astronomía y Astrofísica, 39, 275
351	11:10:19.20	+04:47:31.8	0.030	0.05	0.02	Stoll, D., et al., 1993-1999, AN
352	11:21:37.95	+02:53:20.2	0.049	0.06	0.02	Tovmassian, H. M., et al., 1999, ApJ, 523, 87
355	13:12:11.33	+07:18:28.8	0.093	0.11	0.01	Stoll, D., et al., 1993-1999, AN
357	13:42:10.29	+02:13:42.5	0.076	0.09	0.02	Stoll, D., et al., 1993-1999, AN
358	14:23:45.67	+06:35:03.7	0.050	0.07	0.02	SDSS DR5
359	14:29:56.51	+18:50:20.0	0.033	0.11	0.02	Tovmassian, H. M., et al., 1999, ApJ, 523, 87
360	15:41:26.72	+04:44:09.7	0.108	0.13	0.01	Tiersch, H. et al., 2002, A&A, 392, 33
371	11:43:33.32	+21:53:57.0	0.130	0.14	0.02	Tovmassian, H. M., et al., 1999, ApJ, 523, 87
376	13:56:34.42	+23:21:48.5	0.067	0.03	0.02	Tovmassian, H. M., et al., 2003, astro-ph, 0302105 v1

Table 10.1: *Column 1: identification; column 2 and 3: right ascension and declination (equinox 2000.0, REF); column 4: number of spectroscopic redshifts from literature; column 5: photometric redshift; column 6: error on photometric redshift; column 7: Reference.*

10.4 The method

In order to investigate whether a physical overdensity is present in correspondence of the SHK centroids listed in Table 10.1, for each group in our sample we derived: density

N	Parameter	Description	P/L
	objID	Objects observed photometrically	-
	SpecobjID	Objects observed spectroscopically	-
	satur_centre	Saturation at object's centre	L
	saturated	Presence of saturated pixels	L
	blended	Deblending parent	L
	deblended_at_the_edge	Deblending process at frame edges	L
	child	Product of an attempt to deblend a blended object	L
	nodeblend	Absence of a deblending process	L
	nChild	Number of children of a deblended object	L
	parentID	SDSS identification code of the Parent object	-
1	RA	Right Ascension (J2000)	P
2	DEC	Declination (J2000)	P
3	modelMag$_i$	Asymptotic magnitude in the bands (u,g,r,i,z)	P
4	dered$_i$	Asymptotic magnitude corrected for reddening	P
5	psfMag$_i$	PSF magnitude in the bands (u,g,r,i,z)	P
6	isoA$_i$	Projected major semiaxis in the various bands	P
7	isoB$_i$	Projected minor semiaxis in the various bands	P
8	isoPhi$_i$	Position angle in the various bands	P
9	type	Classification	P
10	type$_i$	Classification in the various bands	P
11	Zspec	Spectroscopic redshift	P
12	ZspecErr	Spectroscopic redshift error	P

Table 10.2: *List of the parameters extracted from the SDSS database and used in the analysis. Column 1: running number, for parameters only. Column 2: SDSS code. Column 3: short explanation. Column 4: type of parameter: F = feature, L = label.*

maps and photometric redshift distributions, projected radial profiles, luminosity functions, colour-magnitude diagrams, surface brightness radial profiles, richnesses, early type fractions and sizes. As an exemplification, in Fig. 10.2 we give the relevant plots for one of two most extreme cases, namely the group SHK 154, one of the densest physically bound structure in our sample. These two cases will also be used to exemplify our procedure. The results for all groups are summarized in Tables 10.3 and 10.4.7 and the complete set of plots for all groups in the sample can be found at the VONeural project pages (http://VONeural.na.infn.it/).

10.4.1 Star-Galaxy Classification

Due to the selection criteria, the original member lists of the Shakhbazian groups are heavily contaminated by stars mistaken for compact galaxies. The better resolution of the SDSS with respect to the older POSS material should b itself ensure that many misclassified objects are removed from our lists but, as additional check we compared the reliability of the Star/Galaxy classification provided by the SDSS classification algorithm with another indicator introduced by Yasuda (Yasuda et al. 2001). The comparison lead to a rate of misclassification $< 1\%$ and the objects with conflicting classifications were excluded from our final catalogues.

10.4.2 Redshift distribution

The only non-ambiguous proof of the physical nature of a cosmic structure is a well defined excess in the redshift distribution, therefore, in our analysis we made use of all available redshift information both spectroscopic and photometric. Unfortunately SDSS-DR5 spectroscopic redshifts are too sparsely distributed to cover a significant number of objects per group. Only in a few cases, a significant excess of spectroscopic redshifts was observed in a $1\ \mathrm{Mpc}$ region around the group centroid. We then derived the photometric redshift distributions in three circular regions having the radii defined above (id est: $150,\ 500\ \mathrm{and}\ 1000\ \mathrm{kpc}$). In each histogram we also plotted the histogram of the background after renormalizing for the area (see the central figure in 10.2 in red). We must also recall that if the assumed spectroscopic redshift is not the right one, the regions considered would be ill-determined, with the effect of obtaining misleading redshift distributions. On the basis of these distributions, and in particular that relative to the region of radius $150\ \mathrm{kpc}$, we estimate the mean photometric redshift of each group and compare it with its spectroscopic estimate.

10.4.3 Projected and spatial galaxy distributions

Using the centroids and the redshifts listed in Table 10.1, around each group we defined three regions: i) an inner one, circular in shape and with a projected radius of $150\ \mathrm{Kpc}$; ii) an intermediate region, annular in shape and with an external radius of $1\ \mathrm{Mpc}$ and, finally; iii) an annular outer region comprised between 2 and $3\ \mathrm{Mpc}$ which defines what we shall call the 'local' background. For each group we also derived a 2-dimensional histogram using the coordinates of the galaxies within a region of radius $R = 1\ \mathrm{Mpc}$, using for both coordinates a bin of $\Delta = 75\ \mathrm{kpc}$, angularly rescaled according to the group's mean spectroscopic redshift. In the upper left figure in 10.2, we show as an example the density map for the group SHK 154, darker colours corresponding to higher galaxy density. In order to reduce the noise we also plot a smoothed version of the maps (with a smoothing factor of 0.4). For each group we also obtained a map of the inner $300\ \mathrm{kpc}$. Objects were coded according to the following criteria:

- Photometric: with photometric redshift (cyan cross in first figure in 10.2).

- Consistent: with photometric redshift such that $|Z_{phot} - Z_{spec}| \leq 3\sigma_{Z_{phot}}$ (blue square in first figure in 10.2).

- Foreground galaxies: with photometric redshift such that $Z_{phot} < (Z_{spec} - 3\sigma_{Z_{phot}})$ (yellow asterisk in first figure in 10.2).

- Background galaxies: with photometric redshift such that $Z_{phot} > (Z_{spec} - 3\sigma_{Z_{phot}})$ (green asterisk in first figure in 10.2).

- Extreme background galaxies: with photometric redshift estimate greater than 0.3 (black asterisk in first figure in 10.2).

- Galaxies with SDSS spectroscopic redshift, are marked by a red triangle. The groups are represented by all the "consistent" galaxies when the assumed spectroscopic redshift measure is right, otherwise we can see no structures or a structure at a different redshift or the superposition of different structures at different distances.

10.4.4 Radial profiles

For each group of the analyzed sample, the numeric surface density radial profile was obtained by considering all galaxies contained in the complete region of radius $3\ \mathrm{Mpc}$, with a bin in radial distance of $100\ \mathrm{kpc}$. The background level was first estimated using the mean numeric surface density in the last $500\ \mathrm{kpc}$ of the profile obtained and then subtracted from the radial profile in order to better highlight the possible presence of a projected overdensity. One example of background subtracted surface density radial profile extending over a region of radius $R = 1\ \mathrm{Mpc}$ can be seen in Fig. 10.1.

10.4.5 Luminosity function

For each group we obtained the background subtracted luminosity functions for galaxies comprised within a region of radius $1\ \mathrm{Mpc}$. Absolute magnitudes were derived using the assumed redshift of the group. The background was determined in the outer region ($2 < R < 3\ \mathrm{Mpc}$) and the bin size is of $1\ rmmag$.

10.4.6 Colour–magnitude diagrams

It is well known that in clusters and groups, early type galaxies define well defined sequences in colour–magnitude diagrams (cf. (Bernardi et al. 2003) and references therein), which over the years has been recognized as a proof of the existence of a structure. We therefore produced for each group color-magnitude diagrams for galaxies contained in a circular region of radius $500\ \mathrm{kpc}$ and with dereddened apparent magnitudes falling in the range $m_a + 3$, where m_a is the magnitude of the brightest galaxies in the inner region, with $|Z_{phot} - Z_{spec}| \leq 3\sigma_{Z_{phot}}$, and whose photometric redshift is such that $|Z_{phot} - Z_{spec}| \leq 3\sigma_{Z_{phot}}$. Another such diagram was produced for the same galaxies except for the different photometric redshift selection criterion used, that is: $|Z_{phot} - \overline{Z}_{phot}| \leq 3\sigma_{Z_{phot}}$.

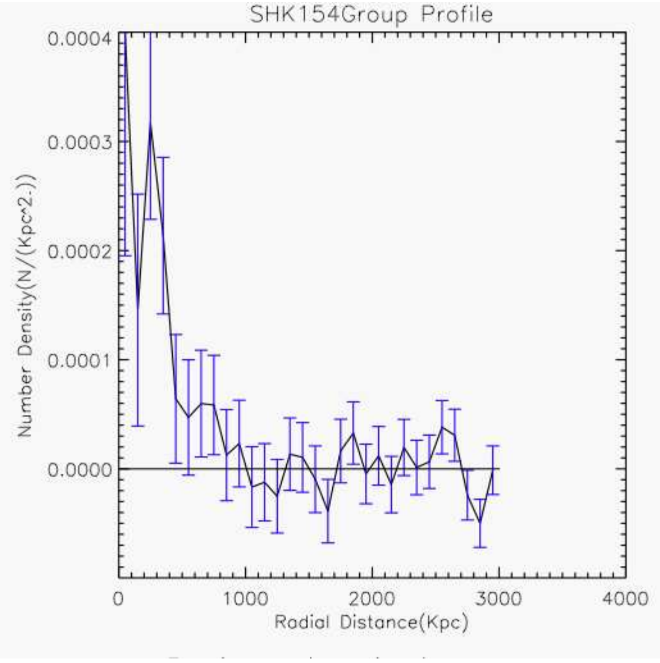


Figure 10.1: *Background subtracted surface density radial profile within a radius* $R = 1$ Mpc *for the group SHK 154.*

10.4.7 Richness

Richnesses were computed using four different criteria and for two different circular regions of radii $R = 150$ kpc and $R = 500$ kpc. All values were subtracted for the counts renormalized for the area and computed in the local background $2 < R < 3$ Mpc. The criteria used were:

i) Radial distance selection (**N1**): all galaxies with $R < 150$ kpc or $R < 500$ kpc

ii) Magnitude selection criterion (**N2**): galaxies obeying the radial distance criterion and such that $dered_r$ differs less than 3 mag from the brightest galaxy among them with radial distance lower than 150 kpc.

iii) Photometric redshift criterion (**N3**): galaxies obeying the radial distance criterion and such that $|Z_{phot} - Z_{spec}| \leq 3\sigma_{Z_{phot}}$.

iv) Global criterion **N4**: galaxies matching simultaneously all the three above listed selection criteria (in radial distance, magnitude and photometric redshift).

Numeric values for all richnesses are reported in table 10.3. In what follows we shall make use of the most conservative estimate, $N4$ which allows to compare groups more homogeneously since the completeness limit is the same for all groups.

Id.	$N1_{500}$	$N1_{150}$	$N2_{500}$	$N2_{150}$	$N3_{500}$	$N3_{150}$	$N4_{500}$	$N4_{150}$
1	2	3	4	5	6	7	8	9
1	35 ±15	18 ± 6	27 ± 6	15 ±4	39 ± 8	16 ± 4	24 ± 6	13 ± 4
5	2 ±12	6 ± 4	7 ± 4	6 ±3	11 ± 6	8 ± 3	6 ± 3	5 ± 2
6	144 ±24	30 ± 8	27 ± 6	8 ±3	49 ± 9	15 ± 4	23 ± 5	7 ± 3
8	-59 ±11	1 ± 4	-4 ± 3	3 ±2	0 ± 5	3 ± 2	1 ± 3	2 ± 2
10	69 ±14	22 ± 6	24 ± 5	8 ±3	53 ± 8	12 ± 4	22 ± 5	8 ± 3
11	31 ±17	0 ± 5	15 ± 6	1 ±2	10 ± 5	1 ± 2	8 ± 4	1 ± 1
14	60 ±22	21 ± 7	10 ± 4	7 ±3	18 ± 7	11 ± 4	10 ± 4	8 ± 3
19	-26 ±21	24 ± 8	20 ± 6	6 ±3	23 ± 7	8 ± 3	17 ± 5	5 ± 2
22	33 ±19	1 ± 5	4 ± 4	4 ±2	5 ± 6	5 ± 3	4 ± 3	5 ± 2
29	-81 ±38	2 ± 11	-11 ± 17	3 ±5	-20 ± 9	-2 ± 2	-22 ± 5	5 ± 2
31	26 ±10	9 ± 4	9 ± 4	6 ±2	18 ± 5	9 ± 3	11 ± 4	6 ± 2
54	8 ±18	7 ± 6	29 ± 8	6 ±3	39 ± 9	11 ± 4	31 ± 7	6 ± 3
55	-17 ±11	9 ± 4	6 ± 4	6 ±3	2 ± 4	6 ± 3	7 ± 3	7 ± 3
57	74 ±13	25 ± 6	19 ± 5	9 ±3	40 ± 7	15 ± 4	18 ± 5	10 ± 3
60	47 ±16	13 ± 5	11 ± 4	5 ±2	11 ± 6	5 ± 3	6 ± 3	5 ± 2
63	-4 ±9	8 ± 4	2 ± 2	5 ±2	-1 ± 4	3 ± 2	1 ± 2	3 ± 2
65	60 ±14	7 ± 4	46 ± 10	8 ±4	28 ± 7	6 ± 3	28 ± 7	6 ± 3
70	-7 ±12	2 ± 4	2 ± 4	2 ±2	10 ± 5	2 ± 2	6 ± 3	2 ± 1
74	4 ±15	1 ± 4	13 ± 7	3 ±3	24 ± 7	5 ± 3	20 ± 66	5 ± 2
95	-22 ±15	7 ± 5	1 ± 3	3 ±2	3 ± 6	4 ± 3	2 ± 2	3 ± 2
96	-19 ±14	-4 ± 4	-4 ± 6	2 ±2	-4 ± 5	-1 ± 1	1 ± 4	0 ± 1
104	-4 ±9	7 ± 4	7 ± 5	6 ±3	2 ± 4	4 ± 2	5 ± 4	4 ± 2
120	88 ±23	33 ± 8	13 ± 6	13 ±4	14 ± 7	9 ± 3	10 ± 4	4 ± 2
123	35 ±14	4 ± 4	15 ± 6	5 ±3	30 ± 8	5 ± 3	20 ± 5	5 ± 2
128	10 ±12	5 ± 4	2 ± 3	4 ±2	2 ± 4	5 ± 2	3 ± 3	5 ± 2
152	-38 ±15	1 ± 5	2 ± 5	-1 ±1	-1 ± 5	2 ± 2	5 ± 3	0 ± 1
154	140 ±23	16 ± 7	32 ± 6	9 ±3	47 ± 9	13 ± 4	30 ± 6	8 ± 3
181	39 ±16	20 ± 6	24 ± 5	10 ±3	42 ± 8	14 ± 4	24 ± 5	10 ± 3
184	6 ±12	2 ± 4	0 ± 3	2 ±2	3 ± 5	3 ± 2	4 ± 3	3 ± 2
186	1 ±21	7 ± 7	16 ± 5	5 ±2	21 ± 7	10 ± 4	13 ± 4	6 ± 2
188	-23 ±19	16 ± 7	5 ± 3	5 ±2	14 ± 7	13 ± 4	7 ± 3	6 ± 2
191	7 ±14	10 ± 5	19 ± 5	10 ±3	34 ± 8	11 ± 4	18 ± 5	11 ± 3
202	-84 ±51	6 ± 15	3 ± 2	3 ±2	66 ± 14	14 ± 5	2 ± 2	1 ± 1
205	54 ±18	9 ± 6	14 ± 5	5 ±2	22 ± 7	4 ± 3	16 ± 4	4 ± 2
213	98 ±26	10 ± 8	8 ± 4	6 ±3	18 ± 8	6 ± 3	7 ± 3	6 ± 2
218	10 ±16	13 ± 6	14 ± 5	11 ±3	7 ± 6	7 ± 3	2 ± 3	6 ± 3
223	72 ±20	12 ± 6	21 ± 5	9 ±3	46 ± 9	11 ± 4	21 ± 5	9 ± 3
229	10 ±13	-1 ± 4	2 ± 4	1 ±1	5 ± 5	0 ± 1	6 ± 3	1 ± 1
231	-1 ±11	3 ± 4	10 ± 4	5 ±2	11 ± 5	5 ± 3	11 ± 4	6 ± 2
237	-54 ±45	39 ± 15	12 ± 5	4 ±2	11 ± 10	6 ± 4	9 ± 4	3 ± 2
245	190 ±26	33 ± 8	18 ± 5	9 ±3	40 ± 9	12 ± 4	18 ± 5	9 ± 3
248	1 ±7	4 ± 3	1 ± 3	4 ±2	0 ± 1	1 ± 1	0 ± 1	1 ± 1
251	-66 ±24	15 ± 8	-1 ± 4	6 ±3	14 ± 7	13 ± 4	6 ± 4	6 ± 3
253	113 ±23	32 ± 8	15 ± 5	12 ±4	35 ± 8	20 ± 5	15 ± 4	13 ± 4
254	23 ±12	0 ± 3	10 ± 4	3 ±2	11 ± 5	3 ± 2	7 ± 3	3 ± 2
258	-6 ±13	10 ± 5	-5 ± 5	5 ±3	-7 ± 4	3 ± 2	-4 ± 3	2 ± 2
344	56 ±20	14 ± 7	7 ± 4	8 ±3	18 ± 7	12 ± 4	10 ± 4	8 ± 3
346	53 ±14	16 ± 5	20 ± 6	13 ±4	34 ± 8	8 ± 3	16 ± 5	8 ± 3
348	73 ±18	15 ± 6	18 ± 5	8 ±3	27 ± 7	10 ± 3	16 ± 4	6 ± 2
351	54 ±47	11 ± 14	15 ± 5	8 ±3	45 ± 13	18 ± 5	16 ± 5	9 ± 3
352	132 ±31	26 ± 10	28 ± 6	8 ±3	95 ± 13	24 ± 5	28 ± 6	9 ± 3
355	89 ±18	15 ± 6	14 ± 5	8 ±3	12 ± 6	5 ± 3	4 ± 3	3 ± 2
357	159 ±24	20 ± 7	38 ± 7	11 ±3	85 ± 11	27 ± 5	35 ± 6	12 ± 3
358	-134 ±28	-22 ± 7	5 ± 3	6 ±2	26 ± 9	12 ± 4	5 ± 3	6 ± 2
359	-337 ±41	17 ± 14	4 ± 8	10 ±4	-30 ± 8	-1 ± 3	-2 ± 3	3 ± 2
360	131 ±18	34 ± 7	25 ± 5	12 ±3	59 ± 9	21 ± 5	23 ± 5	11 ± 3
371	12 ±13	1 ± 4	2 ± 4	5 ±2	13 ± 6	3 ± 2	2 ± 3	5 ± 2
376	61 ±23	6 ± 7	12 ± 4	6 ±3	25 ± 8	9 ± 3	10 ± 4	7 ± 3

Table 10.3: *Column 1: SHK identification number; column 2: $N1_{500}$ richness derived within the 500 Kpc region; column 3: $N1_{150}$ richness derived within the 150 Kpc region; column 4: $N2_{500}$ richness derived within the 500 Kpc region; column 5: $N2_{150}$ richness derived within the 150 Kpc region; column 6: $N3_{500}$ richness derived within the 500 Kpc region; column 7: $N3_{150}$ richness derived within the 150 Kpc region; column 8: $N4_{500}$ richness derived within the 500 Kpc region; column 9: $N4_{150}$ richness derived within the 150 Kpc region.*

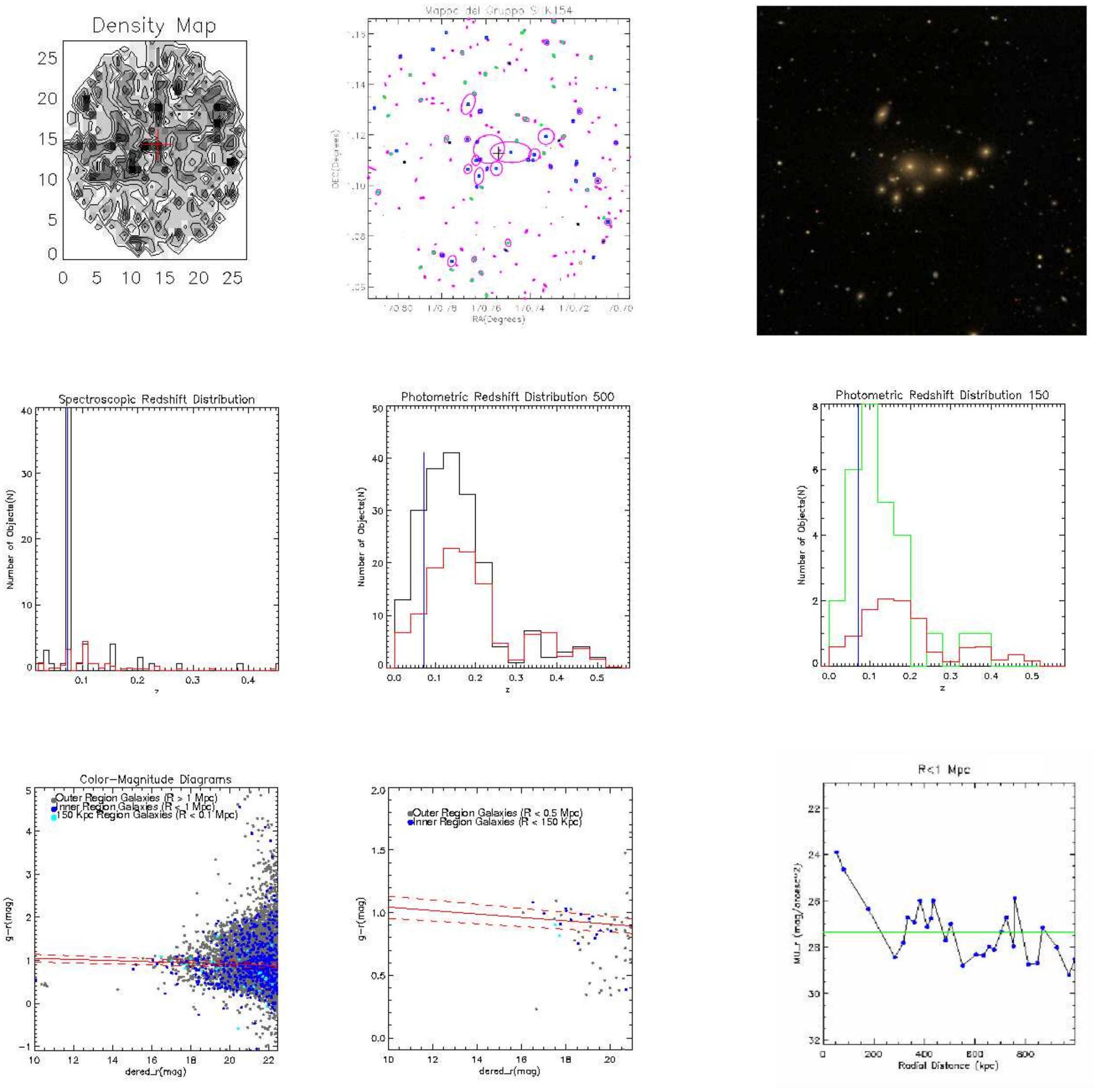


Figure 10.2: *Diagnostic plots for the rich group SHK 154. Upper left panel: Spectroscopic redshifts distributions for galaxies with* $R < 1\ \mathrm{Mpc}$ *(black line) and for local background's galaxies (red line); Upper right panel: Photometric redshifts distributions for galaxies with* $R < 500\ \mathrm{kpc}$ *(balck line) and for local background's galaxies (red line); Lower left panel: Photometric redshifts distributions for galaxies with* $R < 150\ \mathrm{kpc}$ *(green line) and for local background's galaxies (red line); Lower right panel: Color–magnitude diagrams for all the galaxies contained in all the analysed region around the SHK 154 group's centroid. Cyan, blue and grey dots represent respectively inner region galaxies, galaxies within a radial distance of* $R = 1\ \mathrm{Mpc}$ *and local background galaxies. Full red line represents the theoretical color–magnitude relation obtained by (Bernardi et al. 2003), while dashed red lines indicate the scatter in its slope.*

SHK	$f(E)_{gr}$	$f(E)_{env}$	$f(E)_{back}$	N_{env}	R_{spe}	$N_{R_{spe}}$	R_{fot}	$N_{R_{phot}}$
1	0.93	0.52	0.60	35 ± 8	63.502	12 ± 3	63.502	12 ± 3
5	1.00	0.59	0.46	8 ± 6	145.145	4 ± 2	145.145	4 ± 2
6	1.00	0.68	0.49	33 ± 7	102.454	6 ± 2	102.454	6 ± 2
8	0.67	0.47	0.53	-7 ± 4	40.65	3 ± 2	40.65	3 ± 2
10	1.00	0.78	0.37	28 ± 7	306.554	17 ± 4	306.554	17 ± 4
11	0.50	0.47	0.35	11 ± 6	0	0 ± 0	0	0 ± 0
14	0.62	0.45	0.46	4 ± 5	41.312	5 ± 2	41.312	5 ± 2
19	0.80	0.55	0.26	46 ± 8	10.486	2 ± 1	0	0 ± 0
22	0.80	0.46	0.46	8 ± 5	200.735	5 ± 2	200.735	5 ± 2
29	0.80	0.46	0.34	8 ± 5	0	0 ± 0	19.179	3 ± 2
31	1.00	1.00	0.65	5 ± 4	312.331	10 ± 3	312.331	10 ± 3
54	0.86	0.54	0.47	34 ± 10	248.296	14 ± 4	248.296	14 ± 4
55	1.00	0.57	0.45	3 ± 5	112.122	5 ± 2	112.122	5 ± 2
57	0.90	0.80	0.69	23 ± 6	219.366	10 ± 3	219.366	10 ± 3
60	1.00	0.83	0.45	1 ± 4	165.317	6 ± 2	174.598	2 ± 1
63	0.67	1.00	0.71	-3 ± 2	80.169	3 ± 2	174.598	2 ± 1
65	1.00	0.42	0.33	43 ± 11	171.318	9 ± 3	65.036	2 ± 1
70	1.00	0.47	0.45	3 ± 5	0	0 ± 0	0	0 ± 0
74	0.33	0.45	0.39	47 ± 10	240.456	18 ± 5	252.218	23 ± 5
95	1.00	0.73	0.62	-2 ± 4	56.043	2 ± 1	56.043	2 ± 1
96	1.00	0.30	0.36	7 ± 9	0	0 ± 0	109.289	1 ± 1
104	0.80	0.33	0.35	-14 ± 5	22.474	3 ± 2	69.717	2 ± 1
120	0.80	0.39	0.38	6 ± 6	15.225	3 ± 2	96.115	8 ± 3
123	0.67	0.46	0.46	7 ± 7	319.786	14 ± 4	319.786	14 ± 4
128	1.00	0.65	0.42	1 ± 4	36.207	2 ± 1	36.207	2 ± 1
152	0.00	0.30	0.48	3 ± 5	389.19	6 ± 3	0	0 ± 0
154	1.00	0.79	0.65	46 ± 9	119.184	9 ± 3	231.695	8 ± 3
181	0.90	0.58	0.37	27 ± 6	167.867	12 ± 3	150.548	11 ± 3
184	1.00	0.77	0.46	-0 ± 4	77.175	3 ± 2	77.175	3 ± 2
186	0.67	0.35	0.32	4 ± 4	63.141	6 ± 2	63.141	6 ± 2
188	1.00	0.73	0.56	-1 ± 4	125.852	6 ± 2	125.852	6 ± 2
191	1.00	0.78	0.62	24 ± 6	189.981	15 ± 4	266.3	17 ± 4
202	0.00	0.61	0.57	9 ± 4	641.46	3 ± 2	641.46	3 ± 2
205	1.00	0.52	0.40	16 ± 6	177.978	4 ± 2	177.978	4 ± 2
213	0.83	0.71	0.67	-2 ± 4	133.349	5 ± 2	212.026	4 ± 2
218	0.71	0.43	0.50	-9 ± 5	259.381	6 ± 3	135.688	8 ± 3
223	1.00	0.62	0.60	28 ± 7	146.144	8 ± 3	210.429	11 ± 3
229	1.00	0.52	0.38	11 ± 5	0	0 ± 0	194.373	11 ± 3
231	0.83	0.43	0.53	3 ± 5	144.082	5 ± 2	144.082	5 ± 2
237	1.00	0.47	0.49	21 ± 7	193.831	4 ± 2	34.6989	2 ± 1
245	1.00	0.71	0.46	35 ± 7	133.384	8 ± 3	133.384	8 ± 3
248	1.00	0.55	0.57	2 ± 3	0	0 ± 0	70.005	2 ± 1
251	0.57	0.52	0.34	-6 ± 5	88.026	5 ± 2	88.026	5 ± 2
253	0.92	0.55	0.43	13 ± 5	89.837	12 ± 3	89.837	12 ± 3
254	1.00	0.53	0.51	10 ± 6	251.579	8 ± 3	40.431	2 ± 1
258	0.67	0.29	0.43	3 ± 8	127.851	2 ± 2	0	0 ± 0
344	0.67	0.61	0.55	-4 ± 5	112.66	8 ± 3	112.66	8 ± 3
346	0.89	0.80	0.61	-4 ± 5	134.44	7 ± 3	150.763	7 ± 3
348	0.83	0.70	0.57	12 ± 5	205.022	8 ± 3	146.32	6 ± 2
351	0.55	0.51	0.63	19 ± 6	296.333	10 ± 3	296.333	10 ± 3
352	1.00	0.76	0.54	33 ± 7	92.092	8 ± 3	92.092	8 ± 3
355	1.00	0.60	0.47	-4 ± 5	29.388	3 ± 2	29.388	3 ± 2
357	1.00	0.87	0.68	43 ± 8	208.152	13 ± 4	208.152	12 ± 4
358	1.00	0.57	0.65	1 ± 3	85.339	5 ± 2	85.339	5 ± 2
359	0.50	0.40	0.43	-28 ± 6	28.234	2 ± 1	39.45	839 ± 3
360	1.00	0.87	0.50	34 ± 6	178.835	15 ± 4	178.835	15 ± 4
371	0.67	0.59	0.56	-12 ± 5	93.518	5 ± 2	93.518	5 ± 2
376	0.57	0.77	0.57	16 ± 6	78.003	6 ± 2	92.806	5 ± 2

Table 10.4: *Column 1: SHK groups identification number; column 2: fraction of ellipticals within SHK groups $f(E)_{gr}$; column 3: fraction of ellipticals in the outskirts of SHK groups $f(E)_{env}$; column 4: fraction of ellipticals in the background of SHK groups $f(E)_{back}$; column 5: density number of galaxies in the background N_{env}; column 6: radius of the SHK groups using spectroscopic redshift R_{spe}; column 7: density of groups galaxies after background subtraction with spectroscopic redshift $N_{R_{spe}}$; column 8: radius of the SHK groups using photometric redshift R_{fot}; column 9: density of groups galaxies after background subtraction with spectroscopic redshift $N_{R_{phot}}$.*

10.4.8 Surface Brightness Profile

As further information we extracted also the surface brightness profile of each group. Using the listed centroids, we binned the galaxy matching the iv criterion in Par. 10.4.7 in bins formed by three objects and, after integrating the fluxes we normalized for the area encompassing the centres of the objects. We then repeated the process for the other bins. The background level was evaluated by averaging the values corresponding to the objects comprised between $400\ \mathrm{kpc}$ and $1\ \mathrm{Mpc}$. The size of a group (see Tab. 10.4.7) was arbitrarily defined as the radial distance of the most distant surface brightness profile point brighter than $\mu_{back} - 2.5 \log 1.5$. We also obtained a second estimate of the radius using galaxies matching the further condition: $|Z_{phot} - \overline{Z}_{phot}| \leq 3\sigma_{Z_{phot}}$.

10.4.9 Morphology

In order to derive a morphological estimate we followed the approach by Strateva (Strateva et al. 2001) and Shimasaku (Shimasaku et al. 2001) based on the segregation of objects in the colour-colour plane and, for each group we inferred the early type fractions $f(E)_{150}$ for the inner region and obeying to the selection criterion iv in Par. 10.4.7, and the early type fraction in the intermediate annular region satisfying the selection criterion, $f(E)_{env}$. We also derived a further estimate for the outer (background) region, $f(E)_{bkg}$.

10.5 Individual properties

In this section we shortly discuss the individual properties of the groups in our sample.

SHK 1 An important and compact overdensity is detected, confirmed by the presence of a well defined ETS. The physical reality of this group is supported by a marked excess in photometric redshift distribution in correspondence of the spectroscopic redshift value found in literature. The size of this group is about $63\ \mathrm{kpc}$.

SHK 5 This group coincides with HCG 50. The overdensity corresponding to this group is confirmed by both the photometric redshift distribution and the presence of an evident ETS. This structure appears elongated, with a size of $\sim 145\ \mathrm{kpc}$.

SHK 6 Clear overdensity in all diagnostics with a well defined ETS. Presence of other and richer structures in the background.

SHK 8 Studied in detail by (Tovmassian 2005a). One of the objects in the original list by Shakhbazian is a star and another object is very likely in the foreground.

The central objects appear to be interacting and have accordant redshift. Our data confirm the existence of a poor and isolated triplet characterized by a minor overdensity. There is a slight mismatch between the spectroscopic redshift in (Tovmassian 2005a) and the photometric redshift distribution. The triplet diameter is ~ 40 kpc.

SHK 10 Clearcut rich group with elongated structure revealed in all diagnostics.

SHK 11 Spectroscopy shows a small group (at most 3-4 members). All other diagnostics fail to reveal a structure. We reject this group from further discussion.

SHK 14 Clearly defined small group possibly isolated (there is no evidence for surrounding looser structure). (Tovmassian 2005a) have shown that the central objects have all accordant redshifts. Mismatch between spectroscopic and photometric redshifts. The size of this group is ~ 41 kpc.

SHK 19 According to (Tovmassian 2005a) the group is composed by only 4 accordant redshift objects, with a size of about 10 kpc. We find in our data a clear overdensity but the redshift distribution indicates that there is an overlap with other structures in foreground and in background.

SHK 22 Small and compact group. Due to its low richness, the ETS is poorly populated even though all objects fall near the red sequence theoretically predicted in correspondence of the group's redshift. The excess in the photometric redshift distribution coincide with the assumed spectroscopic redshift of the group. Its size is about 200 kpc.

SHK 29 In the photometric redshift distribution a well defined structure is evident at $z_{phot} \sim 0.18$. The literature spectroscopic redshift likely refers to a foreground object. The group identified by Shakhbazian appears as the densest region of an high redshift cluster. No signs of overdensity or of an ETS are detected.

SHK 31 High redshift non compact group projected against nearer and poorer structure. Both structures are clearly evident in all diagnostics.

SHK 54 Complex structure: two clumps at the same redshift and of comparable richness. There is a well defined ETS.

SHK 55 Well defined compact and isolated group.

SHK 57 Well defined high redshift group with problems in ETS.

SHK 60 Poor compact group with larger and looser structure in the background.

SHK 63 Poor group projected in the background of a small group of nearer galaxies.

SHK 65 No evidence for structure in spectroscopic redshifts. In the other diagnostics we observe a large and loose structure with a wide spread in photometric redshifts and a very poor ETS. We reject this group from further discussion.

SHK 70 As previous group. We exclude it from the sample.

SHK 74 This group appears to be loose, while its mean photometric redshift and the spectroscopic one slightly disagree. We detect a well defined ETS.

SHK 95 Small excess of spectroscopic redshifts. Only moderate evidence of a structure in the other diagnostics. We consider it as a poor group at low redshift.

SHK 96 Optical group. We reject it from the sample.

SHK 120 Rich elongated structure projected against looser and more distant one. Poorly defined ETS.

SHK 123 This low multiplicity group shows a pronounced peak in the photometric redshift distribution and a marginally defined ETS. Two different groups of objects are present, maybe at similar redshifts and possibly belonging to the same structure. This groups is a loose and poor structure at the edges of SDSS field. We reject it from our sample.

SHK 128 Compact and isolated group of low multiplicity.

SHK 152 Optical group. Rejected from final sample.

SHK 154 The group has been studied in detail by (Tiersch et al. 2002) who found that most (5 out of 6) galaxies have accordant redshifts. They also detected signs of interactions among the members as well as an extended halo surrounding the group. They also find that some galaxies which appear to be projected on the main group have discordant redshifts. We find a strong overdensity with well defined ETS and a strong excess of photometric redshifts. The size of this group is ~ 180 kpc. The presence of a secondary peak at ~ 270 kpc in spectroscopic redshift distribution could indicate the presence of a second structure at higher redshift which is also confirmed by a double peak in the photometric redshifts distribution.

SHK 181 Rich group, with very well defined overdensity and elongated appearance. Detailed photometric and spectroscopic studies of this group have been performed by (Fasano and Bettoni 1994, Tovmassian et al. 2004) and proved beyond all doubts its physical nature. (Tovmassian et al. 2004) points out that its spatial density is rather low when compared to other SHKGs. Our data show the presence of a well defined ETS and of a pronounced excess in photometric redshift distribution.

SHK 184 Optical group. We reject it from further analysis.

SHK 186 Compact and elongated (chain-like) structure. Visible in all diagnostics.

SHK 188 Rich group slightly off-centred with respect to position listed in Stoll's catalogue. Well defined ETS.

SHK 191 The group extends out to $\sim 270\,\mathrm{kpc}$ and shows a well defined ETS. It is located into the Abell cluster A1097 (Tovmassian 2005b). The excess in the photometric redshift distribution is highly evident.

SHK 202 Very close and highly extended group which appears as a moderate overdensity slightly off-centred with respect to the position provided by Stoll. Other structures are clearly detected in background. Its closeness affects its detectability as a statistic excess. The ETS is quite evident.

SHK 205 Loose group of low richness settling on the expected ETS at its redshift. The average photometric and spectroscopic redshifts slightly disagree.

SHK 213 Loose group containing several more compact regions. Moderate overdensity and ETS well defined. Possible existence of a background group. The photometric redshift distribution presents a well defined excess at redshift higher than the assumed one, while spectroscopic redshifts seem to confirm the existence of a large structure at high redshifts ($z \sim 0.4$).

SHK 218 Well defined group confirmed by ETS and excess of photometric redshifts. There is strong evidence for a second group/cluster in the background.

SHK 223 Loose and not very rich group, having a well defined ETS. It shows slightly off-centred substructures.

SHK 229 Loose, not very rich structure with ill defined ETS. Rejected from further analysis.

SHK 231 Compact, elongated structure of low multiplicity. Visible in all diagnostics.

SHK 237 Possibly a nearby poor group. Low statistics makes it hard to detect in our diagnostic tools. Rejected.

SHK 245 Well defined overdensity of elongated appearance. The ETS is well defined and both the spectroscopic and photometric redshift distributions hint to the existence of a second background structure.

SHK 248 Negligible overdensity. Slight excess of photometric redshifts at a redshift different from the nominal one. This seems to indicate the existence of a small foreground group. We reject it from further discussion.

SHK 251 Well defined but not very strong overdensity which extends well beyond the central condensation. ETS detected but not conclusive. The background seems to indicate the presence of a rich structure at the edges of the field. The size of this group appears to be $\sim 88\ \mathrm{kpc}$, even if probably, because of an error in considering the background level, the real size is $\sim 220\ \mathrm{kpc}$.

SHK 253 Compact, isolated structure elongated in shape. Strong excess of photometric redshifts. Well defined red sequence. It extends out to $90\ \mathrm{kpc}$.

SHK 254 Probably an optical group. The spectroscopic assumed redshift and the mean photometric one strongly disagree. This is probably due to the spectroscopic assumed redshift that probably refers to a foreground galaxy. The photometric redshift distribution indicates the presence of a rich structure at a much higher redshift. ETS marginally defined. On the whole, it is possible the existence of a poor and disperse group.

SHK 258 Optical group. Rejected.

SHK 302 Similar to SHK 254. A few objects rather loosely distributed form an excess in the photometric redshift distribution at redshift much higher than the spectroscopic one listed in (Tiersch et al. 1999).

SHK 344 Detailed study by (Tovmassian et al. 2004) shows moderate signs of interaction among some members. In our data, we find a well defined overdensity with a clearly visible ETS and a second structure in the background.

SHK 346 Strong overdensity clearly visible in all diagnostics. There's evidence for a background compact substructure.

SHK 348 Loose group at redshift slightly higher than what is listed in (Tovmassian et al. 2005). The group is clearly visible in all diagnostics and extends out to $\sim 182\ \mathrm{kpc}$.

SHK 351 Close and therefore very extended group. A pronounced excess in photometric redshifts and a well defined CMR are evident. It's clearly visible a background extended structure.

SHK 352 Rich and well defined group of compact appearance. The ETS is clearly visible and there is a well pronounced excess of both spectroscopic and photometric redshifts. Radial profile extends out to $\sim 92\ \mathrm{kpc}$.

SHK 355 Small compact and very well defined group of galaxies. Its CMR is evident and its size is $\sim 29\,\mathrm{kpc}$.

SHK 357 It is a rich and compact group clearly detected as both an excess in photometric redshifts distribution and a well defined ETS.

SHK 358 Complex object. First of all its nearness prevent the successful application of our diagnostic tools. There are hints for the existence of a structure but we prefer to reject it from our final analysis.

SHK 359 Poor, nearby structure clearly overlapping with a background one. Its large angular extension makes the diagnostics meaningless.

SHK 360 This group is the central part of the Abell cluster 2113 and has been studied in detail by (Tiersch et al. 2002) who proved that the central galaxies have accordant redshifts and that the brightest objects are interacting. The group is almost entirely composed of early type (E/S0) galaxies. In our data, this group appears to be the richest in the sample with a compact core and a very well defined borders. The photometric redshifts excess appears to be slightly off-centred with respect to what is listed in both Stoll's catalogue and in (Tiersch et al. 2002). As it could be expected, a significant excess is found out to $1\,\mathrm{Mpc}$.

SHK 371 Poor, compact and well defined group, elongated in shape.

SHK 376 It is a peculiar SHK group, exclusively composed by spirals and studied in detail by (Tovmassian et al. 2003).

We refer to this paper for a detailed discussion of membership and interactions. Our data confirm that the group is embedded within a larger structure extending out to $\sim 150\,\mathrm{kpc}$. According to the above discussion we excluded from further analysis the groups SHK 11, SHK 65, SHK 70, SHK 96, SHK 123, SHK 152, SHK 184, SHK 229, SHK 237, SHK 248, SHK 258, SHK 358.

10.6 Global properties

Using the individual properties listed in the previous paragraph together with the parameters in Tables 10.3 and 10.4.7, we can now proceed to an analysis of the global optical properties of the groups in our sample. The first fact to be noticed is that SHKGs, on average much richer in early type systems than both the field and the compact groups in the Hickson lists. In Fig. 10.6 (left panel) we plot the histogram of the early type fraction for the field and the SHKG and HCG samples respectively. As it can be seen, $\sim 95\%$

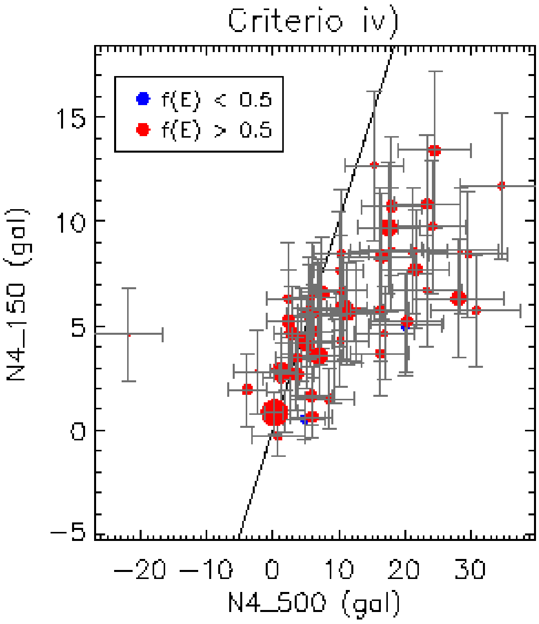


Figure 10.3: *Richness N4 (see text) inside a 150 kpc radius against the same quantity computed within a 500 kpc. The size of the symbols is scaled according to the mean spectroscopic redshift of SHK groups.*

of the SHKGs has $f(E) \geq 0.5$ and this fraction drops to $\sim 50\%$ for the field and $\sim 65\%$ for the HCG sample. Moreover when we consider the early type fraction in the inner, intermediate and background regions shown in Fig. 10.6 (right panel), we note that the dominant morphological type changes when moving outwards. In Figure 10.3 we plot the richnesses $N4$ obtained in the 150 kpc and 500 kpc regions and a clear trend can be observed for increasing richness: the richest groups being less centrally concentrated than poorer ones. In the two plots the sizes of the symbols are scaled according to the early fraction content (Left panel) and to the group redshift (right panel).

10.7 Conclusions

The results of this paper is that approximately 70% of SHK groups reveal the presence of a spatial over-density, an excess in the photometric redshift distribution, compatibility between literature's spectroscopic and our photometric values of mean redshift. Furthermore, most of them present a well-defined CMR if compared with that obtained by Bernardi (Bernardi et al. 2003). For the remaining 30% we can identify several case studies: i) groups are not real, but only projection effects; ii) groups are very close and

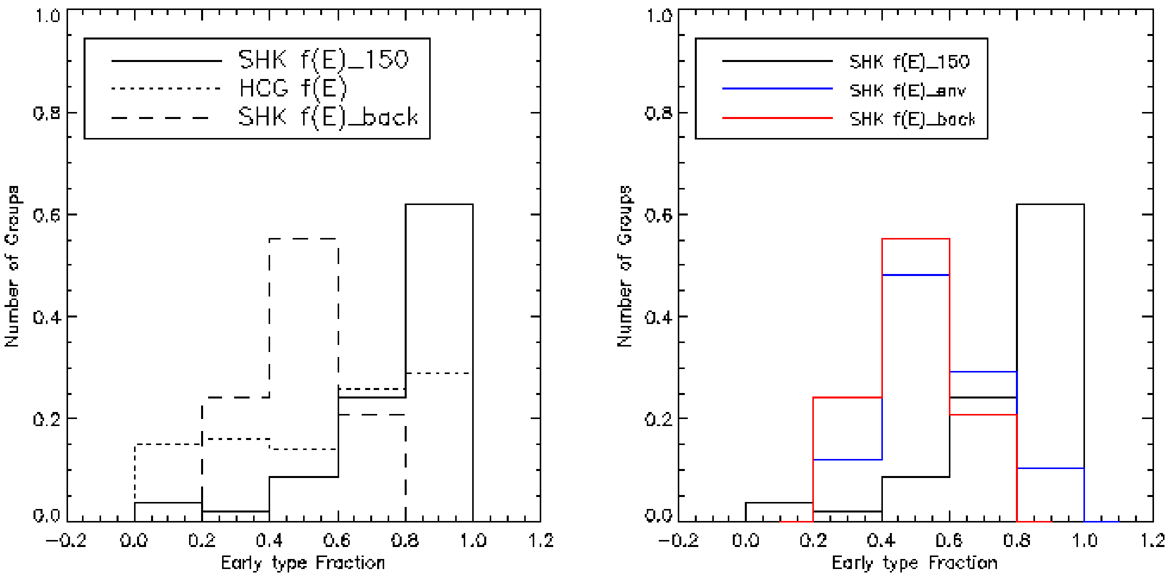


Figure 10.4: *Left Panel: Normalized $f(E)$ distributions for the inner region (continuous line), background (dashed line) and Hickson compact groups taken from literature (dotted line). Right Panel: Normalized $f(E)$ distributions for inner region (black line), environment (blue line) and background (red line).*

largely contaminated by the background, so that, since the low S/N ratio, they cannot be detected; iii) the spectroscopic and photometric measures of redshift are not consistent and there's no structure at the spectroscopic redshift value. In this latter case, we reveal the presence of other structures at a redshift different from the spectroscopic one; iv) there's more than one excess in the photometric redshift distribution, revealing the superposition of different structures at different distances.Richnesses were then estimated according to four different criteria, based on selections on radial distances from the centroids, on apparent magnitude and on photometric redshift. The best of them turned out to be the combination of all the three above mentioned criteria. The SHK richness ranges from 3 to 13 gal, and about 95% of SHK groups show a high content of early-type galaxies ($f(E) \geq 0.5$) (see figures 10.3 and 10.6), much more than revealed for groups of similar multiplicity like the Hickson's ones. Comparing the values of richness obtained in two circular regions having radius $R = 150$ kpc and $R = 500$ kpc respectively, we find hints of the existence of two types of groups, compact and isolated, and embedded inside more extended and disperse structures (see figure 10.3). Additional evidence in this direction comes also by the change we find in the the dominant morphological type when the inner region and the surrounding environment are considered

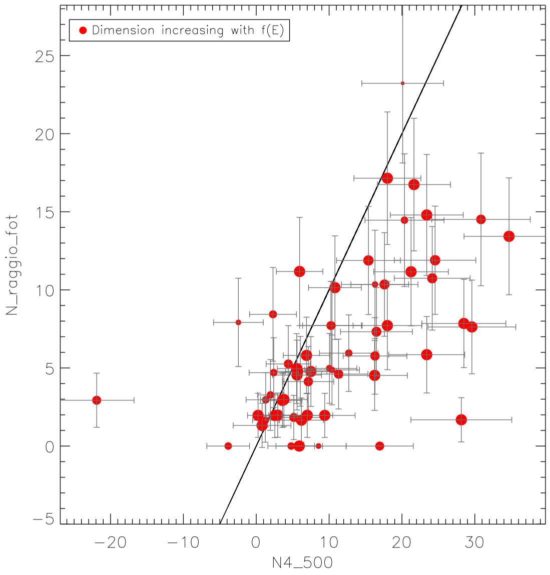


Figure 10.5: *Comparison between richnesses obtained according with the fourth selection criterion for galaxies with radial distance smaller than the estimated group's radius and with* $R < 500\,\mathrm{kpc}$. *Symbols dimensions increase with* $f(E)$.

(see figure 10.6), a behaviour which agrees with Dressler's density-morphology relation (Dressler 1980). In other words, it seems that the more rich and the more extended is the group, the larger is the fraction of early-type galaxies (fig. 10.5) that it contains.The SHK groups that belong to the second class may be going through a first formation phase, following the scenario according to which groups evolve assuming sequentially loose, [halo+core] and compact configurations. This would agree with the Secondary Infall Scenario. In the future we plan to extend the analysis to the other 167 groups contained into the region covered by SDSS and to evaluate the results obtained in this work using a larger sample. Moreover we want to observe the galaxies of the groups spectroscopically, aiming to better analyse their membership. Finally we plan to mine the existing archives in order to investigate the diffuse X-Ray emission from the objects in our sample. This would allow to probe the presence of a hot diffuse gas trapped into the gravitational potential well of the groups and therefore their reality.

Chapter 11

AGNs identification and classification

There once was a note, pure and easy,
Playing so free like a breath rippling by.

Pure and Easy, The Who

11.1 Classification: general definition

While the first step of every classification in a homogeneous sample of objects is the establishment of a relation of order of qualitative nature $R(p_1, ..., p_n)$ depending on n parameters representing observables quantities, the ultimate purpose is to conceive a physical taxonomy, i.e. a categorization of the objects on the base of physical properties and phenomena. In general, a taxonomy can be achieved if and only if the observables chosen as parameters determining the value of the relation of order $R(p_1, ..., p_n)$ reflect as closely as possible one or more physical laws $L_1, ..., L_m$. The validity of a taxonomic classification can be tested evaluating its power of prediction as the capability of formulating new forecasts regarding observables not directly related to the relation of order considered and that can be explained by the same physical laws $L_1, ..., L_m$.

11.1.1 Goodness of a classification

Given the sample of objects $S = \{x_1, x_2, ..., n_n\}$, an ideal classification is a partition $P = \{S_1, S_2, ..., S_n\}$ such that:

$$\bigcup_{j=1}^{m} S_j = P \text{ where } m \ll n \text{ and } \forall j \;\; S_j \neq \phi \tag{11.1}$$

and also the following relation holds:

$$S_j \bigcap_{j \neq k} S_k = \phi \tag{11.2}$$

where the classes S_j are defined by the parameters or predicates $p_{j,l}$ with $l = \{1, ..., L\}$ defined inside S_j:

Requirements on the parameters	Drawbacks
The number of parameters needs to be as little as possible	The use of a large number of parameters reduces the utility of the classification.
The parameters are supposed to be easily measurable from the data	Classification of an object can take longer than a detailed study
Parameters are supposed not to be contradictory	A given object can be assigned to different classes according to the set of parameters considered for the classification
Parameters are requested to reflect at least in first approximations some physical criterion	Classification based on phenotipycal characteristics not reflecting any real physical difference can be misleading

Table 11.1: *Synthetic description of the main requirements for an optimal classification.*

$$x_i \in S_j \Leftrightarrow \forall l, p_{j,l}(x_i) = \text{True} \tag{11.3}$$

A summary of the features of an optimal choice of the parameters used for a classification and the most serious drawbacks that may affect the classification in case these requirements are not met by the chosen parameters can be found in the table 11.1.

11.2 Morphological classification of galaxies

One classical example of classification in astronomy is the morphological classification of galaxies. Since the developments of the first telescopes capable of observing the details of the nearest galaxies, the wealth of different details in the observed appearance has prompted a classification based on the morphology of these interesting objects. During past centuries, all extended objects not resolved in stars were classified under the common name of *nebulae*, while the first attempts of classification in more then one class were based on morphological appearance of objects only. Even if the correctness of any tempted classification depends on the availability of a homogeneous sample composed by a large number of objects whose observable quantities have been measured in the same conditions, it is of paramount importance in particular for astronomical classifications that the wavelength interval within which objects are observed, the observational conditions and the peculiar features of the observational equipment used for gather-ing the data need to be specified correctly and are assumed to be as homogeneous as possible throughout the process of classification. The principal goal of morphological studied has been to obtain insights into galaxy formation and evolution. Fundamental problems, such as the effects of environment, morphological segregation in clusters, the origin of bars, the driving mechanisms for spiral structure, the possibility of evolution of structures within a Hubble time and the underlying factors which determined the various types at the time of galaxy formation all require accurate knowledge of morphology in order to be addressed reliably. The actual effectiveness of morphology to address these problems depends on how well the relationships between the various types of galaxies are established, and the extent to which follow-up observations and theoretical analyses are carried out. The setbacks of morphological classification of galaxies can be placed in a general perspective and summarized as follows:

- Many distinct features or components are present in galaxy aspect: spirals, bulge, arms, rings, lenses, disks, etc. These features, visible in several combinations and inclinations, suggest that galaxy morphology consists in an almost impenetrable and irrational assortment of unique templates.

- Galaxies have a wide range of surface brightnesses, luminosities and other measured properties, so that selection effects are always important since they could affect both the availability of details and the statistical reliability of the samples of objects used to set up the classification.

- Galaxy structure is by and large continuos, as can be seen looking at high dimensional parameter space classifications, where transition cases between discrete and common morphological types always exist.

- Environment is important in determining the galactic form, since the distributions of morphologies in different environments can differ significantly (cf. field and clusters).

- Dynamical events on relatively short time scales (like mergers, interactions, tidal stripping and collisions) can lead, on one hand, to the possible evolution of rare transient forms, but on the other hand, could be also responsible for some of the most common morphological types. In this sense, these events can at the same time both confuse and simplify the complex problem of morphological classification of galaxies.

The goal of any physical classification is to reduce the complexity and to correct the misleading aspects of morphological classification, stepping back form the chaos produced

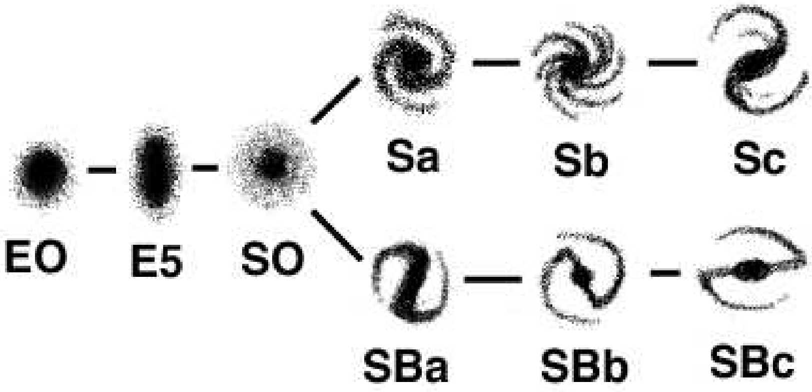


Figure 11.1: *Hubble's morphological classification of galaxies. Letters associated to each model indicate codes used to identify galaxies belonging to different types: E stands for Elliptical (the number is a measure of the ellipticity of the galaxy), S stands for Spiral and SB stands for Spiral Barred (the following lowercase letter expresses the appearance of the arms relative to the central part of the galaxy (the "bulge")).*

by the overwhelming mass of observed forms of galaxies in order to provide relationships between morphology and other observable quantities. If a classification system eventually shed light on these relationships, then it could provide the needed physical insights for, at least partly, addressing the ultimate goals of understanding galaxy formation and evolution.

11.3 Hubble's classification of galaxies

The morphological classification of galaxies proposed by Hubble in (Hubble 1958) represents a template for all morphological classification in astronomy (Hubble 1926), and for this reason, it has been deeply studied and scrutinized during its long history of successful application. Hubble introduced the classification scheme illustrated in the figure 11.3, which separates most galaxies into elliptical, normal spiral, and barred spiral categories, and then sub-classifies these categories with respect to properties such as the amount of flattening for elliptical galaxies and the nature of the arms for spiral galaxies (see figure 11.3).

Galaxies that do not fit into these categories are classified separately as irregular galaxies. Hubble's classification can be interpreted also as a classification of galaxies as a function of the disk/bulge prominency ratio, since this parameter goes from almost 0 in early type galaxies to a value in the interval $[0.5, 1.0]$ going from the early-type to late-type galaxies. This interpretation depends on the wavelength of the observations, and even if holding in the optical bands, it is false in other regions of the electromagnetic spectrum. A drawback of successful Hubble's classification, even if still today widespread through the astronomical community, is that it is strongly influenced by the nature of observations from which the images of galaxies on which the classification was based were taken, and in turn is affected by the same selection effects and biases of the imaging material the data were extracted from. At the same time, Hubble's tuning fork embodies the classical astronomical classification based solely on the morphology of the objects, and as a consequence, reflects the incompleteness introduced by the use of information derived by one specific spectral interval. Classifications based on other photometric and spectroscopic parameters and not affected by the same problems of old classification of galaxies, are complementary to the Hubble's and hopefully will stand by morphological classifications in the future in the effort of a reciprocal improvement and completion.

11.4 An introduction to Active Galactic Nuclei

As active galactic nucleus (AGN) is designed a compact region at the centre of a galaxy which has a much higher luminosity than average galaxies, over some or all of the electromagnetic spectrum (in the radio, infrared, optical, ultra-violet, X-ray and/or gamma ray wavebands). The host galaxy may be visible or not according to its luminosity respect to the luminosity of the AGN, and to the redshift of the object. When visible, a galaxy hosting an AGN is called an active galaxy. According to one of the most accepted model of AGN nature and emission, the radiation from AGN is the result of the transformation in electromagnetic energy of a small fraction of the gravitational energy released by a supermassive black hole residing at the centre of the host galaxy (and of the same AGN, of course) during the accretion of matter from the surrounding region of the host galaxy.

11.4.1 Observational features of AGNs

There is no single observational signature of an AGNs-galaxy system, since several physical mechanisms of emission at different wavelength, interaction between the light from the host galaxy and the AGN and the selection and biasing effect depending on the redshift of the source all collaborate at producing a large assortment of observational scenarios. Some of the historically important features that have allowed systems to be identified as AGNs are listed below:

- Nuclear optical continuum emission, which is clearly visible in the optical as a roughly power-law dependence on wavelength;

- Nuclear infrared emission;
- Broad optical emission lines;
- Narrow optical emission lines;
- Radio continuum emission, showing a spectrum characteristic of synchrotron radiation;
- X-ray continuum emission;
- X-ray line emission;

This rather complex scheme is usually simplified introducing classification based on the characteristics of only one of these properties. As an example, is often convenient to divide AGN into two classes, conventionally called radio-quiet and radio-loud AGNs and diversified according to the strength of the radio emission. Further AGN terminology, whose goal is the produce a more precise scheme of the classification as a function of the observable parameters, is often confusing since distinctions between different types of AGN sometimes reflects historical differences in how objects were discovered or initially classified, rather than real physical differences. A summary of main families of radio-quiet AGNs together with their observational characteristics are reported below:

- Low-ionization nuclear emission-line regions (LINERs). These systems show only weak nuclear emission-line regions, and no other signatures of AGN emission. It is debatable whether all such systems are true AGN (powered by accretion on to a supermassive black hole). If they are, they constitute the lowest-luminosity class of radio-quiet AGN. Some may be radio-quiet analogues of the low-excitation radio galaxies.
- Seyfert galaxies. Seyferts were the earliest distinct class of AGN to be identified. They show optical nuclear continuum emission, narrow and (sometimes) broad emission lines, (sometimes) strong nuclear X-ray emission and sometimes a weak small-scale radio jet. Originally they were divided into two types known as Seyfert 1 and 2: Seyfert 1s show strong broad emission lines while Seyfert 2s do not, and Seyfert 1s are more likely to show strong low-energy X-ray emission. Various forms of elaboration on this scheme exist: for example, Seyfert 1s with relatively narrow broad lines are sometimes referred to as narrow-line Seyfert 1s. The host galaxies of Seyferts are usually spiral or irregular galaxies.

- Radio-quiet quasars/QSOs. These are essentially more luminous versions of Seyfert 1s: the distinction is arbitrary and is usually expressed in terms of a limiting optical magnitude. Quasars were originally 'quasi-stellar' in optical images, and so had optical luminosities that were greater than that of their host galaxy. They always show strong optical continuum emission, X-ray continuum emission, and broad and narrow optical emission lines. Some astronomers use the term QSO (Quasi-Stellar Object) for this class of AGN, reserving 'quasar' for radio-loud objects, while others talk about radio-quiet and radio-loud quasars. The host galaxies of quasars can be spirals, irregulars or ellipticals: there is a correlation between the quasar's luminosity and the mass of its host galaxy, so that the most luminous quasars inhabit the most massive galaxies (ellipticals).

- 'Quasar 2s'. By analogy with Seyfert 2s, these are objects with quasar-like luminosities but without strong optical nuclear continuum emission or broad line emission. They are hard to find in surveys, though a number of possible candidate quasar 2s have been identified.

On the other hand, classification for radio-loud AGNs in terms of their observed properties is the following:

- Radio-loud quasars. These AGNs behave exactly like radio-quiet quasars with the addition of radio emission from an extended region protruding form the inner region of the quasar (jet).These objects show strong optical continuum emission, broad and narrow emission lines, and strong X-ray emission, together with nuclear and often extended radio emission.

- 'Blazars' (BL Lac objects and OVV quasars). These classes are distinguished by rapidly variable, polarized optical, radio and X-ray emission. BL Lac objects show no optical emission lines, broad or narrow, so that their redshifts can only be determined from features in the spectra of their host galaxies. The emission-line features may be intrinsically absent or simply swamped by the additional variable component: in the latter case, it may become visible when the variable component is at a low level (Vermeulen et al. 1995). OVV quasars behave more like standard radio-loud quasars with the addition of a rapidly variable component. In both classes of source, the variable emission is believed to originate in a relativistic jet oriented close to the line of sight. Relativistic effects amplify both the luminosity of the jet and the amplitude of variability.

- Radio galaxies. These objects show nuclear and extended radio emission. Their other AGN properties are heterogeneous. They can broadly be divided into low-excitation and high-excitation classes (Hine and Longair 1979, Laing et al. 1994).

Low-excitation objects show no strong narrow or broad emission lines, and the emission lines they do have may be excited by a different mechanism. Their optical and X-ray nuclear emission is consistent with originating purely in a jet (Chiaberge et al. 2002). They may be the best current candidates for AGN with radiatively inefficient accretion. By contrast, high-excitation objects (narrow-line radio galaxies) have emission-line spectra similar to those of Seyfert 2s. The small class of broad-line radio galaxies, which show relatively strong nuclear optical continuum emission (Grandi and Osterbrock 1978) probably includes some objects that are simply low-luminosity radio-loud quasars. The host galaxies of radio galaxies, whatever their emission-line type, are essentially always ellipticals.

11.4.2 The unified model of AGNs

The AGN phenomenon suggests a unique theoretical model accounting for the spectrum of observations. Although direct proof is still missing, the gathered evidences point towards gravitational accretion of matter by supermassive black holes as being as the primary energy source for AGNs. Gravitational potential energy is converted into radiation via viscous dissipation in an accretion disk surrounding the black hole. AGNs appear to have axial rather than spherical symmetry, and it is hypothesised that all the unresolved AGN components are surrounded by an optically thick obscuring torus that permits the AGN radiation to escape only along the torus axis, which is defined by large scale ionization cones. As an example of partial confirmation of this model, in radio-loud AGNs, the radio axis appears to be aligned with the torus axis. The clear signs of anisotropy in both radio emission and higher-frequency radiation imply that the appearance of a given AGN will depend strongly on the observer's location relative to the axis of symmetry. Indeed the observed characteristics of a particular AGN might be so strongly orientation dependent that the classification of the system is a function of the viewing angle. This is the fundamental notion behind what is called as "unified models" of active galactic nuclei, whose basic motivation is the appeal of the idea that they appeal to our belief that any description of natural phenomena should be made as simple and general as possible in absence of evidences to the contrary. The assumption at the base of the unified model is that there is less intrinsic diversity among AGNs that is observed, and that the wide variety of AGN phenomena which are visible is due to a combination of real differences in a small number of physical parameters coupled with apparent differences which are due to observer-dependent parameters, like the orientation. Unified models of AGN can be characterized as either strong or weak, depending on the number of fundamental parameters allowed. Weak unification models allow more physical diversity and attempt to explain the relationships among a limited number of AGN types. An example of weak unification model is one that allows two

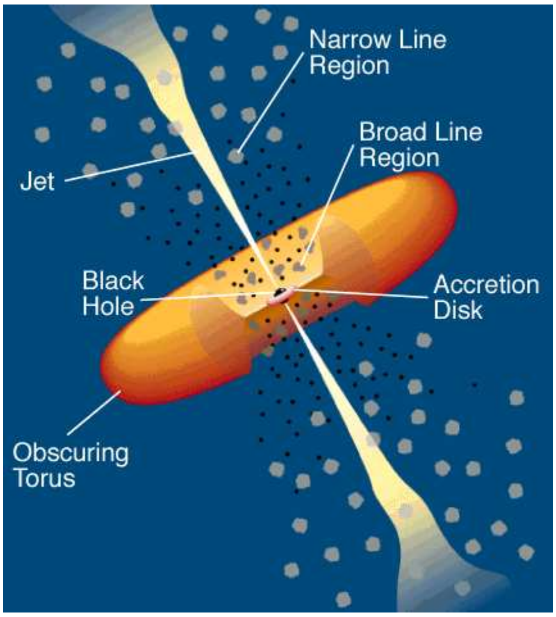


Figure 11.2: *Schematic representation of an AGN with the main emission regions, according to a version of the unified model for active galactic nuclei.*

intrinsic parameters, radio and optical luminosity. In this model there are two basic types of AGNS, radio-quiet and radio-loud. In each kind, a wide range of phenomena that have to do with variations in these two basic parameters plus apparent differences due to the orientation of the system relative to the observer. The complementary strong model, on the other hand, assumes that there is only a single intrinsic parameter, the total luminosity, and that all of the differences observed, including the discrepancies in the optical and radio properties, are ascribable to various orientation effects.

The simplest scheme (see figure 11.2) that can be used to develop unifying models was summarized by (Antonucci 1993). All radio-quiet active galaxies and quasars have regions close to the nucleus that produce broad lines and featureless continuum radiation. This region is surrounded by opaque tori, approximately along the axis of which are located weak radio jets. When the torus in a given object is face-on relative to an observer, so that the line of sight reaches the nuclear region, the broad lines and continuum are seen. Otherwise only narrow lines, which are produced in a region outside the torus, are seen directly. But when the signal to noise ratio is sufficiently high, the nuclear region can be seen in radiation that is reflected into observer's direction because of scattering by electrons, which also produced polarization. The torus has the same geometrical properties, such as an opening angle, in all sources and these determine the relative proportion of broad and narrow line objects. In a minority of cases, twin jets of relativistic particles are present, oriented close to the torus axis. These jets

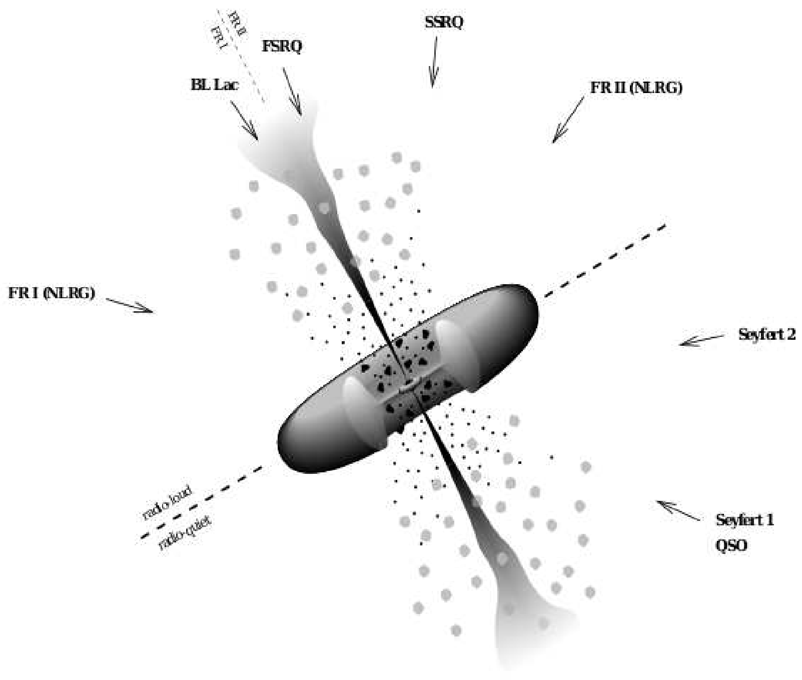


Figure 11.3: *Unification model for AGNs and dependence of the observed type of active galaxy on the orientation angle. Blazars are those AGNs for which the jets are close to line of sight, while a regular quasar or a Seyfert 1 galaxy is observed if the orientation angle is* $\sim 30^\circ$*, where the narrow-line and broad-line regions are visible. At larger angular offsets, the broad-line region will be hidden by the torus, the corresponding class being Seyfert 2 galaxies. Perpendicular to the jet axis, the full extent of the jets may be seen particular at low frequencies, giving rise to a morphology typical of radio galaxies.*

produce powerful radio emission through synchrotron process and their bulk motion is relativistic, at least close to the nucleus, with a Lorenz factor quiet similar in all sources. When the axis of a radio-loud source us close to the line of sight, the observer sees a continuum superimposed with a broad and narrow lines, and a one-sided jet perhaps with superluminal motion. When the orientation is very close to the line of sight, the beamed emission dominates and the object appears to be a blazar (see figure 11.3).

11.5 The classification of AGN in multicolour surveys

Several different selection techniques have been used to determine homogeneous samples of AGNs since their discovery. In particular, ambitious photometric surveys (York et al. 2000) have caused the number of known AGNs and quasars to rise from one to tens of thousands. Yet even in this day of very large surveys and deep digital imaging, the identification of more than 1.6 million quasars with redshift $z < 3$ that are expected to fill the celestial sphere to limiting magnitude g = 21 is very far. The problem lies not in covering enough of the sky to faint enough magnitudes, but rather in the efficient separation of quasars from other astronomical sources. Current algorithms are typically more than 60% efficient for UV-excess (UVX) quasars to relatively bright magnitudes, but the selection efficiency drops toward fainter magnitudes where the photometric errors are largest and most of the observable objects reside. Further complicating the issue is the need to obtain time consuming spectral information for each candidate in order to confirm their AGN nature with reliable identification of spectral continuum and features. For this reason, surveys of quasars would benefit considerably from algorithms with selection efficiencies that mitigate the need for confirming spectra. More in detail, optical surveys for quasars, including the Sloan Digital Sky Survey (SDSS) (Stoughton et al. 2002), typically rely on simple colour cuts in two or more colours to select objects that are likely to be quasars and to reject objects that are of different nature. A promising method to select quasars from imaging data, stemmed from the long lived colour cut method and taking advantage of the birth of multi-band observations covering large areas of the sky, is based on the exploitation of already known AGNs to determine what regions of colour space are occupied by confirmed quasars. Once identified these regions, quasar target selection is carried out simply picking up objects from those regions of colour space that are most likely to yield quasars (or perhaps least likely to yield significant number of contaminants). The success of this candidates selection strategy would have been very difficult until a decade ago given the lack of abundant spectroscopic data, but with the current abundance of imaging data and spectroscopic follow-up provided by current and planned all-sky deep optical mixed surveys, it is now possible to design such high efficiency algorithms.

11.6 Dimensionality reduction techniques

In statistics, dimensionality reduction is the process of reduction of the number of random variables needed to describe a data sample without loss of information, i.e. with the maximum possible global variance. The motivation for dimensionality reduction of multivariate distributions is the fact that sometimes data analysis, such as regression or classification, can be done more accurately in a simpler features space than in the original one. Dimensionality reduction techniques can be divided into two different classes, namely i) feature selection and ii) feature extraction. Feature selection algorithms try to find a subset of the original variables to accomplish the tasks above mentioned. Feature extraction, on the other hand, is the result of the application a mathematical function mapping the multidimensional space into a space of fewer dimensions, i.e. the original feature space is transformed by applying a generic transformation. In the case of a simple linear combination of features, this technique is called Principal Components Analysis (PCA) (see paragraph 11.6.1 for details on the PCA).

11.6.1 Principal Component Analysis

Principal components analysis (PCA) is a member of a family of algorithms, known as multivariate statistics, whose aim is to find, in a sample of N objects with n measured variables x_n, what variables produce primary and secondary correlations via the remaining $(n-2)$ variables. PCA represent the main linear technique for dimensionality reduction and performs a linear mapping of the data to a lower dimensional space in such a way that the variance of the data in the low dimensional representation is maximized. This result can be achieved in practice by constructing of the correlation matrix of the data and calculating the corresponding eigenvectors. The eigenvectors that correspond to the largest eigenvalues (the so called principal components) can be used to reconstruct a large fraction of the variance of the original data, since the eigenvectors determine the transpose matrix T for variable transformation and axis rotation. This rotation diagonalizes the covariance matrix, i.e. in the new vector basis the cross-terms are zero. In a mathematical fashion, the goal of PCA is finding a new set of n variables χ_n that are orthogonal (i.e. independent), each one obtained as a linear combination of the original variables x_n:

$$\chi_i = \sum_{j=1}^{n} a_{ij} x_j \tag{11.4}$$

with values of a_{ij} such that the smallest number of new variables accounts for as much of the variance as possible. In this context, the functions χ_i are the principal components. Moreover, the first few eigenvectors can often be interpreted in terms of the large-scale physical behaviour of the system. The original space (with dimension of the number of points) has been reduced (with data loss, but hopefully retaining the most important variance) to the space spanned by a few eigenvectors. PCA assumes that the covariance matrix suffices to describe the data, which is the case if the data are drawn from a multivariate gaussian distribution, or in general when a simple quadratic form, using the covariance matrix, can describe the distribution of the data.

11.7 Clustering

Clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (or clusters), so that all objects contained in each subset share some common trait, often proximity according to some defined distance measure. Data clustering is a common technique for statistical data analysis, and clustering techniques are widespread in many fields, including machine learning, data mining, pattern recognition and image analysis. Data clustering algorithms can be divided in hierarchical or partitional algorithms. Hierarchical algorithms find successive clusters using previously established clusters, whereas partitional algorithms determine all clusters at once. Hierarchical algorithms can be again split in agglomerative (also called "bottom-up") or divisive ("top-down") methods, where agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters, whereas divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters. Two-way clustering, co-clustering or biclustering are clustering methods where not only the objects are clustered but also the features of the objects. An important step in the definition of any clustering strategy is the selection of a distance measure. In general, a distance, or metrics, is a function $d : X \times X \longrightarrow$ **R**, where **R** is the set of real numbers, and for all $x, y and z$ in X this function is required to satisfy the following conditions:

1. $d(x, y) \geq 0 \text{ (nonnegativity)}$

2. $d(x, y) = 0 \text{ id and only if } x = y \text{ (identity of indiscernibles)}$

3. $d(x, y) = d(y, x) \text{ (symmetry)}$

4. $d(x, z) \leq d(x, y) + d(y, z) \text{ (subadditivity)}$

One or more of these conditions can be relaxed, and, in particular, a function not abiding the third requirement is called a quasi-metric (another important distinction between clustering strategies is represented by whether the adopted definition of distance is symmetric or not). The properties of the distance used will determine how the similarity and proximity of two elements is evaluated, thus influencing also the shape of the clusters, since some elements may be close to one another according to one distance and further away according to a different one distance definition. Common distance definitions usually employed for clustering are showed in the following list, together with a brief description:

- Euclidean distance or squared euclidean distance.

- Manhattan distance or taxicab geometry, is a form of geometry in which the usual metric of euclidean geometry is replaced by a new metric in which the distance between two points is the sum of the (absolute) differences of their coordinates.

- Mahalanobis distance, based on correlations between variables by which different patterns can be identified and analysed. It is useful to determine similarity of an unknown sample set to a known one. It differs from Euclidean distance in that it takes into account the correlations of the data set and is scale-invariant, i.e. not dependent on the scale of measurements.

- Maximum norm, defined as the maximum of the absolute values of the single components of the vector $x = (x_1, ..., x_n)$ associated to the position of an object with respect to the origin of the feature space: $\| x \|_\infty = \max\{\|x_1\|, ..., \|x_n\|\}$.

- Angle between two vectors, which can be used as a distance measure when clustering high dimensional data.

- Hamming distance, which measures the minimum number of substitutions required to change one member into another. Derived from information theory, the Hamming distance between two strings of equal length is the number of positions for which the corresponding symbols are different.

11.7.1 Agglomerative clustering

Agglomerative hierarchical clustering builds a hierarchy of clusters, based on a distance definition between clusters and a linkage strategy (i.d. the rule followed to choose the members of the couples of clusters whose distance is measured and merged together to form a new cluster if a threshold criterion on the distance is met). Possible definitions of distance have already been discussed in the previous paragraph 11.7; in the case of Negative Entropy clustering (NEC) (used in the work described in chapter 12), the distance between clusters has been replaced by a function known as "negentropy" borrowed by information theory, indicating a measure of the distance to normality of a given distribution. Negentropy $J(p_x)$ is an always positive and invariant by any linear invertible change of coordinates function, defined as null if and only if the distribution p_x is gaussian:

$$J(p_x) = S(\phi_x) - S(p_x) \tag{11.5}$$

where $S(\phi_x)$ stands for a standard gaussian density with the same mean and variance as p_x, and $S(P_x)$ is the differential entropy defined as:

$$S(p_x) = -\int p_x(u) \log p_x(u) du \tag{11.6}$$

The linkage strategy adopted in the application in chapter 12, is based on the evaluation of the distance between the clusters after their projection on a one dimensional manifold (i.e. a straight line) at each step of the agglomerative process. The projection is performed applying Fisher's linear discriminant to the means and variances of the multivariate distributions associated to the clusters in the feature space, and selecting couple of clusters whose separation is a minimum. In other words, only contiguous clusters on the line are considered for the computation of the negentropy and for the possible merging. In general, the distance definition between two clusters $\mathcal{A}$ and $\mathcal{B}$ used to define the linkage strategy is one of the following:

- The maximum distance between elements of each cluster (complete linkage):

  $$\max\{d(x, y) : x \in \mathcal{A},\ y \in \mathcal{B}\}$$

- The minimum distance between elements of each cluster (single linkage):

  $$\min\{d(x, y) : x \in \mathcal{A},\ y \in \mathcal{B}\}$$

- The mean distance between elements of each cluster (average linkage):

  $$\frac{1}{\mathrm{card}(\mathcal{A})\mathrm{card}(\mathcal{B})} \sum_{x \in \mathcal{A}} \sum_{y \in \mathcal{B}} d(x, y)$$

  where $\mathrm{card}(\mathcal{A})$ is the cardinality of cluster $\mathcal{A}$;

- The sum of all intra-cluster variance;

- The increase in variance for the cluster being merged (Ward's criterion);

- The probability that candidate clusters spawn from the same distribution function (V-linkage).

The traditional representation of this hierarchy of clusters is a tree (called a dendrogram), with all the initial individual elements at one end and a single cluster containing every element at the other (see figure 11.4). In this picture, every step of the agglomeration is associated at a given value of the distance threshold, for which a couple of clusters is merged; the dendrogram, cut at a given height, will give a clustering at a selected precision and a different number of final clusters.

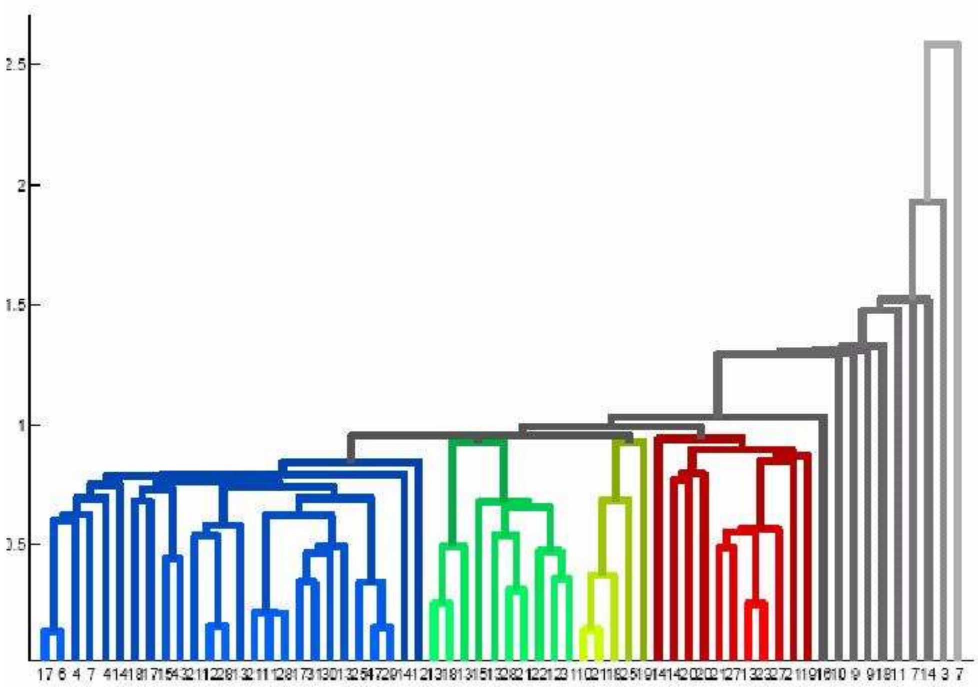


Figure 11.4: *Example of dendrogram. The x-axis contain the initial clusters, on the y axis is the value of the threshold corresponding to different clusterings.*

A common way to implement this type of clustering is to compute a distance matrix for all clusters at every step of the agglomerative process, where the generic element a_{ij} is the distance between the i-th and j-th clusters. As clustering progresses, rows and columns are combined as clusters are merged, and the distances updated. As can been easily understood, each agglomeration occurs at a greater distance between clusters than the previous one, so that it is necessary to decide to stop the clustering process either when the clusters are too far apart to be merged (”distance criterion”) or when there is a sufficiently small number of clusters (”number criterion”). In both cases, a fiducial value for clusters distance or number of final clusters needs to be fixed *a priori*.

11.8 Our work and the VO

The work described in the next chapter 12 is the application of an original unsupervised clustering method for the classification of galaxies to the problem of the selection of candidate quasars from photometric datasets, using spectroscopic base of knowledge. This algorithm is meant to be applied to the more general problem of the physical classification of galaxies, in order to produce a reliable taxonomy of the observed galaxy populations as thoroughly discussed in the chapter 2 and, more precisely, in the paragraph 2.3.

This algorithm based on the concerted use of a dimensionality reduction algorithm and of a particular type of agglomerative clustering method, is one of the statistical tools that will hopefully facilitate and boost the development of a new kind of astronomy based on the Virtual Observatory infrastructure and services. The following work is the first application of this methodology whose relevance for the specific problems of AGNs and obscured optically ″dull″ AGNs problems is being investigated.

Submitted as: D'Abrusco, R., Longo, G., Walton, N. A.. 2007 – *"Quasar candidates selection in the Virtual Observatory era"*, Monthly Notices of the Royal Astronomical Society

Chapter 12

A data-mining approach to quasar candidates selection

Abstract

We present a method for the photometric selection of candidate quasars in multiband surveys. The disentanglement of quasar candidates and stars is performed in the colour space through the combined use of two algorithms, the Probabilistic Principal Surfaces and the Negative Entropy clustering, which are for the first time used in an astronomical context. Both methods have been implemented in the VONeural package on the Astrogrid platform. Even though they belong to the class of the unsupervised clustering tools, the performances of the method are optimized by using the available sample of confirmed quasars and it is therefore possible to learn from any improvement in the available "base of knowledge". The method has been applied and tested on both optical, and optical plus near infrared data extracted from the visible SDSS and infrared UKIDSS–LAS public databases. In all cases, the experiments lead to high values of both efficiency and completeness comparable or better than most methods already known in the literature.

12.1 Introduction

Over the years, serendipitous discoveries and systematic searches have caused the number of confirmed quasars to grow dramatically but we are still far from having discovered even a significant fraction of the $\sim$ 1.6 million QSO's which are expected to populate the universe out to $z \simeq 3$ (e.g. (Richards et al. 2004)).

Due to their high intrinsic brightness, such lack of coverage is not due to limiting fluxes but mainly to the difficulties encountered first in disentangling normal stars from QSO candidates and then in confirming their quasar nature through additional data (such as spectroscopy, radio or X-ray fluxes) which are usually either difficult or very time consuming to provide for statistically significant samples of objects.

This is very f.unfortunate since, as it has been stressed by many authors (Richards et al. 2002, Richards et al. 2004), large samples of quasars covering a broad range of redshifts and selected with uniform and well controlled criteria, are greatly needed to

address many relevant issues such as the evolution of the quasar luminosity function, or the spatial clustering of quasars as a function of the redshift.
The most notable recent efforts at building extensive samples have been the quasars searches in the Sloan Digital Sky Survey (hereafter SDSS, (York et al. 2000)) and its main concurrent project, the 2dF QSO redshift survey (Croom et al. 2001) which will soon be joined by ongoing or planned multiband photometric survey projects such as, for instance, the Palomar Quest (Djorgovski et al. 2004), or the VST (Capaccioli et al. 2003) and VISTA (McPherson et al. 2006) surveys.

From the photometric point of view, all quasar candidates selection algorithms are based on a few simple facts: i) stars have a spectrum that is roughly blackbody in shape, while quasars have spectra that are characterized by featureless blue continua and strong emission lines, thus causing quasars to have colors different from those of stars;

ii) as stressed by (Richards et al. 2001), the overall shape of the continuum of quasars is well approximated by a power law, and since a redshifted power law remains a power law with the same spectral index, quasar colors are only a weak function of redshift as emission lines move in and out of the various filters;

iii) quasar spectra deviate dramatically from power laws at rest wavelengths below 1216 Å, where the Ly_α forest systematically absorbs light from the quasar (Lynds 1971) making the quasar appear increasingly redder with redshift.

In the particular case of infrared wavelengths, the availability of large-field detectors on large telescopes has provided the opportunity to undertake surveys capable of establishing the importance of the main mechanism of reddening, i.e. the extinction by dust, on the observed population of quasars. Since the spectral energy distributions of quasars vary significantly as a function of the wavelength, flux measurements at widely separated wavelengths are used to characterize fully the spectral properties of the quasar population. More precisely, two methods exploiting the differences between the power-law nature of quasar spectra and the convex spectra of stars in ranges have been proposed to select candidate quasars, based on the fact that quasars are significantly brighter than stars at both short wavelengths - the UVX method (Richards et al. 2004) - and long wavelengths - the KX method (Warren et al. 2000).

In order to build large samples of quasars a major goal is to improve the reliability and efficiency of the algorithms used to extract from multiband survey data the list of quasar candidates.

Most current algorithms are typically more than 60% efficient for UV-excess (UVX) quasars to relatively bright magnitudes, but the selection efficiency drops toward fainter magnitudes where the photometric errors are largest and most of the observable objects resides. As shown by (Richards et al. 2004), however, it is possible to build algorithms achieving levels of accuracy and completeness (for a definition of these terms see Section (12.3) which can mitigate the need for confirming spectra.

One additional fact that needs to be noticed is that the efficiency of all quasar candidate selection algorithms depends on some degree on the fine tuning of the algorithm on the *a priori* knowledge (hereafter "Base of Knowledge" or BoK) of what quasars are and on what their main characteristics are. This BoK may be either built out of synthetic spectra or from available quasar samples. The latter approach being based on fewer *a priori* assumptions seems preferable but, on the other end, it keeps all biases introduced by the selection criteria (to be more explicit: if a specific subclass of objects is not present in the BoK, the algorithm will not be able to pick them up). In a near future, however, the large amount of data which will be made available to the community through the Virtual Observatory (Walton 2002), will provide an ever growing (both in size and accuracy) BoK which will allow to overcome at least some of the existing limitations. In this paper we present a method based on unsupervised clustering capable to map the photometric parameter space using the information contained in the BoK and to disentangle stars from candidate quasars. Even though it is applied to the SDSS and/or UKIDSS public data, the method is of general validity and can be easily adapted to any other dataset given that a proper BoK is available.

In section 12.2 we present the main characteristics of the data used for the experiments, and in Section 12.3 we shortly summarize the main methods used so far. In Section 12.4 we introduce the clustering algorithm and the agglomerative method and in Section 12.5 we discuss the results of the experiments performed. The conclusions are drawn in Section 12.7.

In two forthcoming papers (D'Abrusco et al. 2007) and (Cavuoti et al. in prep.) we shall discuss the application of the method to the selection of heavily obscured quasars and to the physical classification of galaxies respectively.

12.2 The data

12.2.1 SDSS data

The Sloan Digital Sky Survey is a digital survey aimed at covering $\sim 10,000$ sq. deg. mainly in the Northern hemisphere (Stoughton et al. 2002) in five specifically designed bands (u, g, r, i, z) (Fukugita et al. 1996) and is complemented by an extensive redshift survey for about 10^6 objects (mainly galaxies and QSO's). The SDSS data are made available to the community through a public archive which at the moment is distributing its

Fifth Data Release (hereafter DR5) (Adelman-McCarthy 2007). SDSS provides the best dataset ever where to mine for photometrically selected subsamples of objects. As such, it has been extensively studied in almost all its aspects and an impressive amount of literature has been produced providing an accurate knowledge of completeness, selection effects etc. ((Adelman-McCarthy 2007)).

As to quasar selection from the SDSS data, it is worth to recall a few facts. The SDSS photometric system does not allow the detection of quasars with $z > 6$ and; with the additional constraint of having the objects detected in at least two bands this limit reduces to $z \sim 5.8$ (Fan et al. 2001, Richards et al. 2002). At the low-redshift end, the design of the u filter and the location of the gap between the u and g filters were chosen to emphasize the difference between objects with power-law spectral energy distributions (SEDs), such as quasars at $z < 2.2$, and objects that are strongly affected by the Balmer decrement, particularly A stars, which are historically the prime contaminants in multicolor optical surveys for low-redshift quasars.

12.2.2 UKIDSS data

The United Kingdom Infrared Deep Sky Survey (UKIDSS) is a near-infrared sky survey that will cover 7500 square degrees of the Northern sky, extending over both high and low Galactic latitudes, in JHK bands down to $K \simeq 18.3$, thus reaching three magnitudes deeper than 2MASS. UKIDSS has been designed and operated to be the true near-infrared counterpart to the SDSS survey. In fact UKIDSS is made up of five separate surveys and includes two deep extra-Galactic elements, one covering 35 square degrees down to $K = 21$, and the other reaching $K = 23$ over 0.77 square degrees of the sky. In this work we make use of the UKIDSS Large Area Survey (hereafter LAS) which aims at covering an area of 4,000 deg^2 overlapping with the SDSS. The LAS is expected to be completed after an observing period of seven years. LAS is surveying the sky in four photometric band $YJHK$ with typical limiting magnitudes $[20.5, 20.0, 18.8, 18.4]$ and astrometric accuracy typically $< 0''.1$. UKIDSS DR1 (Dye et al. 2006) release overlaps a subset of the SDSS northern and southern areas with photometric and astrometric performances similar to the SDSS.

12.2.3 The bases of knowledge

In this work, three different samples of objects have been used as BoKs.

The first sample (hereafter S-A) is formed by candidate quasars selected from the SDSS-DR5 database, classified as unresolved (i.e. belonging to the table "Star"), and for which the spectroscopic classification index "specClass" is available together with a spectroscopic redshift for each object. Such index classifies objects in: $SP = 1$ stars, $SP = 2$

Table 12.1: *"specClass" distribution of the three samples used in this work.*

Sample	SP = 0	SP = 1	SP = 3	SP = 4	SP = 6
S-A	43 (1.6%)	1176 (43.3%)	827 (30.4%)	73 (2.6%)	600 (22.0%)
S-UK	23 (1.0%)	954 (43.5%)	773 (35.3%)	69 (3.1%)	373 (17.0%)
S-S	2609 (2.4%)	22636 (20.4%)	53554 (48.4%)	4661 (4.2%)	10737 (9.67%)

galaxies, $SP = 3$ nearby AGN, $SP = 4$ quasars; $SP = 5$ sky, $SP = 6$ late type stars. Since most SDSS quasars fall into the star-like category, the "specClass" index of the objects in our datasets are only 0, 1, 3, 4 & 6. The objects considered in this sample have been selected inside a roughly rectangular patch of the sky, situated in the equatorial region matching with the area covered by the data release 1 of UKIDSS LAS observations (see next paragraph). For 2519 sources matching the selection criteria, point spread function magnitudes "psfMag" in the five SDSS bands (u, g, r, i, z) have been retrieved.

The second sample (hereafter S-UK) is formed by all objects belonging to the SDSS-DR5 "Star" (containing all star-like photometric sources) table (with spectroscopic classification available) positionally matching with UKIDSS-DR1 LAS objects which have been also classified as stars according to the "mergedClass" classification index (requiring that "mergedClass" = -1). The matching was performed selecting all unresolved sources in LAS "lasSource" database table laying within 10 arc-seconds from the SDSS source. For this sample optical PSF magnitudes from SDSS and near infrared PSF magnitudes in the four LAS UKIDSS (J, Y, H, K) bands have been retrieved. A total of 2192 candidate quasars were successfully selected according to these prescriptions.

The third sample of objects (hereafter S-S) is formed by star-like sources belonging to the SDSS-DR5 "Target" table and selected as candidate quasars according to the algorithm described in ((Richards et al. 2002)). For all these objects spectroscopic classification ("specClass") and redshifts are available. The only additional constraint applied on these objects is the fact that the psf magnitudes need to be correctly measured in all photometric band. The number of objects selected according to these requirements is 94,196. In the figures (12.1) and (12.2), we plot the positions and the redshift distribution for the members of the samples. The S-A and S-UK samples differ only for a few objects for which one of more UKIDSS magnitudes have not been measured and therefore their distributions in redshift are almost identical. The distribution in redshift of S-S objects is characterized by a peak at $z \sim 1.7$. The composition of the samples in terms of spectroscopic classification index "specClass" is given in table (12.1).

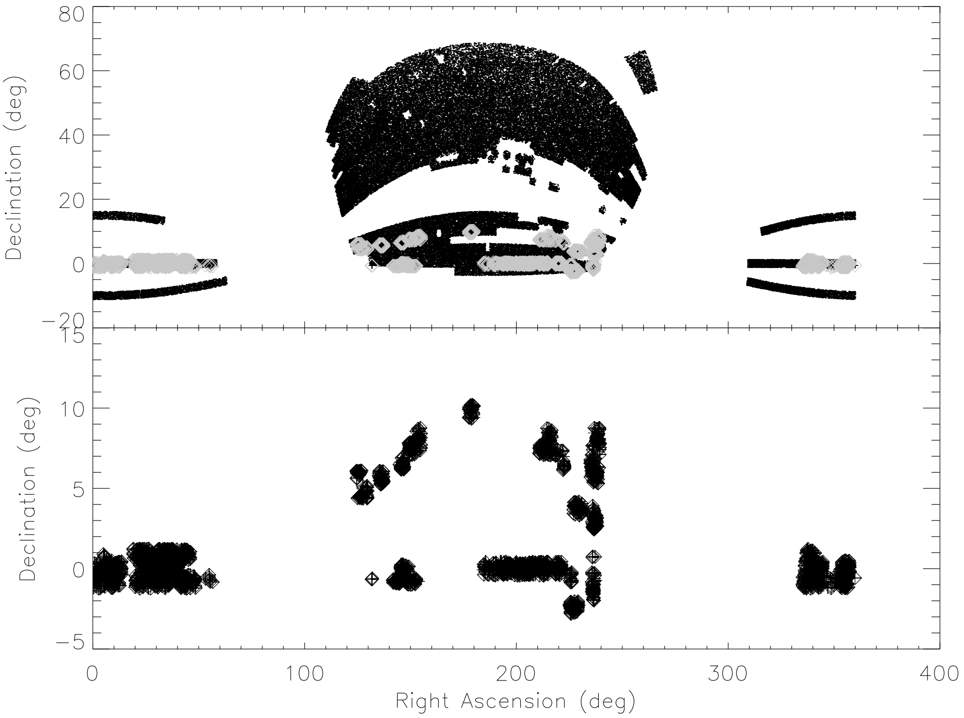


Figure 12.1: *Positions of objects belonging to the three samples used in this work (see text). The upper figure represents the S-S sample, while the lower shows the positions of the S-A sample members as black crosses and those of the S-UK sample members as grey diamonds.*

12.3 Photometric selection of quasars

As a result of their distinct colors, the general idea behind all quasar candidate selection algorithms working in photometric multidimensional parameter spaces is that quasars tend to lay far away from the regions occupied by normal stars (i.e. what we shall call stellar locus). Therefore, from the mathematical point of view, the problem of quasar candidate selection can be regarded as that of properly partitioning the parameters space in order to isolate the regions populated by quasars, minimizing the level of contamination from stars and the number of missed quasars.

Spectroscopic observations of candidate quasars are necessary for quasars confirmation and in order to test the relative performances of the algorithm used to identify candidates. Such performances are usually expressed by two parameters, called respectively ″completeness c″ and ″efficiency e″, and defined as follows:

$$c = \frac{\text{N candidates}}{\text{N \textit{a priori} known QSO's}}; \quad e = \frac{\text{N confirmed}}{\text{N candidates}} \tag{12.1}$$

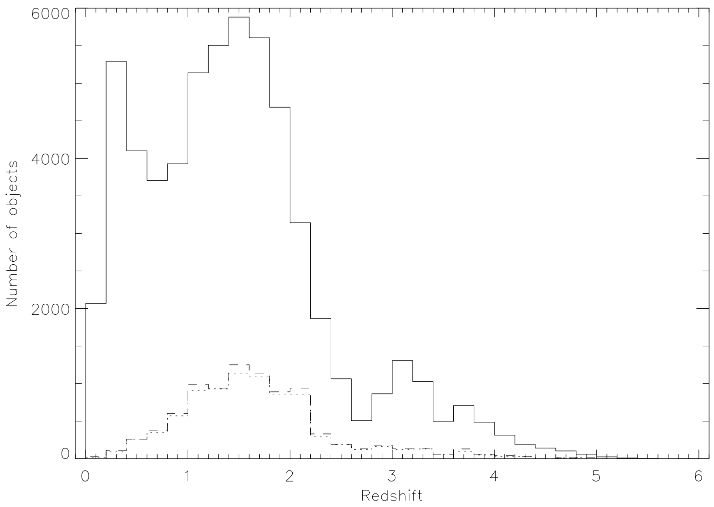


Figure 12.2: *Redshift distribution for the three samples used in this work. S-A (dashed line) and S-UK (dotted line) samples counts have been multiplied by ten to increase their visibility.*

It is apparent that c provides a measure of how good is the method at retrieving all quasars in the sample, while e provides a measure of the contamination in the list of candidates selected by the algorithm. In what follows we shall identify as "base of knowledge" (BoK) the spectroscopically studied objects which can be used to build the samples of *a priori* known and "confirmed" quasars. The optimal balance (if anything like that exists at all...) between completeness and efficiency is a delicate one since stars outnumber quasars by several orders of magnitude and improving the efficiency by rejecting objects in regions of color space in which both stars and quasars lie, necessarily affects the completeness. Due to the unavoidable incompleteness in spectroscopic surveys (and as a matter of fact, of any other type of selection criterion), the BoK is always affected by biases which reflect into the value of c. In other words, in a given photometric catalogue, in order to have an exhaustive list of *a priori* known quasars, all objects should be observed spectroscopically and the same holds true for the list of candidates.

The BoK is also needed to evaluate the errors. For example, the knowledge of the intrinsic spread in quasar spectral indices translates into a lack of knowledge of the intrinsic spread in quasar colors (Richards et al. 2002). The first attempt to produce a list of candidate quasars from multicolor survey data was by (Sandage and Wyndham 1965). This pioneering attempt was soon followed by many others (Koo and Kron 1982, Schmidt and Green 1983, Warren and Hewett 1990, Warren et al. 1991), and more recently by (Hewett et al. 1995, Hall et al. 1996, Croom et al. 2001, Richards et al. 2002, Richards et al. 2004). In what follows we shall shortly summarize some of them, focusing on those which have been tested on the SDSS dataset.

12.3.1 SDSS selection algorithms

The official SDSS quasars candidate selection algorithm (Richards et al. 2002) (hereafter R02) is sensitive to quasars at all redshifts lower than $z \leq 5.8$ (i.e. very close to the theoretical limit predicted for the SDSS), and to atypical AGNs such as broad absorption line quasars and heavily reddened quasars. Performances of this algorithm, as stated in the paper, are completeness $c \sim 90\%$ and efficiency $e \sim 65\%$. The R02 algorithm is less accurate in certain zones in the colours space where degeneracy between colours of quasars in the redshift range $[2.2, 3.0]$ and stars (e.g. Brown Dwarfs) is present due to the Lyα forest crossing the SDSS filters system. Since this redshift range is crucial for the cosmological applications which were the primary target of SDSS (Stoughton et al. 2002), objects falling in these regions were nonetheless selected paying the price of a worse overall efficiency.
Star-like objects are selected in a four dimensional colour space defined by the (u, g, r, i, z) SDSS bands. Non stellar candidates are selected via their colours and by matching unresolved sources to the FIRST radio catalogs. The R02 algorithm can be summarized as it follows: i) objects with spurious and/or problematic fluxes in the imaging data are rejected; ii) extended matches to FIRST radio sources are preferentially targeted without reference to their colors; iii) the sources remaining after the first step are compared to the distribution of normal stars and galaxies in two distinct three-dimensional color spaces, one for low-redshift quasar candidates (based on the $ugri$ colors) and one for high-redshift quasar candidates (based on the $griz$ colors). The two groups are selected down to limiting magnitudes $i^* \sim 19.1$ and ~ 21.2, respectively. Color selection is performed accordingly to the their distance from a modelled, fixed hypersurface containing the stellar locus which, for a given photometric system, has been shown to be rather stable with respect to changes in stellar populations (e.g. (Richards et al. 2002)). No specific line is drawn between quasars and other types of active galactic nuclei.

12.4 Unsupervised quasar selection

Quasars candidates detection can be achieved using unsupervised clustering algorithms on colours space distribution of candidate objects. The method presented in this paper follows a hierarchical approach which, starting from a preliminary clustering performed on the objects inside the parameter space, is followed by a second phase of agglomeration which reduces the initial number of clusters produced in the first step to an *a priori* unknown numbers of final clusters. We than have a phase of what we shall call ”labelling”, based on the existing base of knowledge, i.e. on the objects for which independent spectroscopic confirmation is available. This labelling is used to refine the partition of the parameter space in order to define the stellar and quasar loci. The characterization of the final clusters is then used to select *ex novo* candidate quasars the from photometric datasets.

The unsupervised[1] clustering is accomplished using the Probabilistic Principal Surfaces algorithm which, strictly speaking is not a clustering algorithm but rather a nonlinear generalization of principal components particularly suited for dimensionality reduction purposes. As it will be shown, PPS project the input data onto a lower dimensionality space defined by what we shall call 'latent variables' which act as attractors of input vectors and, therefore, can be interpreted as cluster centroids. The algorithm used for the second step is the so called Negative Entropy Clustering algorithm (hereafter NEC), which has been selected after comparative testing against other similar algorithms among the wide class of unsupervised hierarchical agglomerative clustering algorithms according to its high efficiency and reliability (Ciaramella et al. 2005).

One advantage, which is as well a limitation, of this technique needs to be stressed: the distribution in the parameter space of the objects belonging to each cluster selected by the NEC is approximated by multivariate gaussians. Consequently, the projection of cluster members positions along each axis of the parameter space can be modelled as a one-dimensional gaussian, and common statistics quantities such as the average or the standard deviation can be used to describe the distribution of the members of each cluster over the entire parameter space. On the other end, the assumption of gaussian shape for the clusters requires further discussion (see Section 12.4.3).

12.4.1 Latent variables and the PPS algorithm

The Probabilistic Principal Surfaces model (Chang and Ghosh 2000, Chang and Ghosh 2001, Staiano 2003) belongs to the family of the so called *latent variables* methods (Bishop 1999) and can be regarded as an extension of the Generative Topographic Mapping (Bishop 1998).

The goal of any latent variable model is to express the distribution $p(\mathbf{t})$ of the variable $\mathbf{t} = (t_1, \ldots, t_D) \in \mathbb{R}^D$ in terms of a smaller number of latent variables $\mathbf{x} = (x_1, \ldots, x_Q) \in \mathbb{R}^Q$ where $Q < D$. In order to achieve it, the joint distribution $p(\mathbf{t}, \mathbf{x})$ is decomposed into the product of the marginal distribution $p(\mathbf{x})$ of the latent variables and the conditional distribution $p(\mathbf{t}|\mathbf{x})$ of the data variables given the latent variables. It is convenient to express the conditional distribution as a factorization over the data variables, so that the joint distribution becomes:

[1] PPS, as most other unsupervised algorithms require the number of clusters to be provided by the user. In our approach this limitation can be circumvented by assuming a number of clusters much higher than what could be realistically be present in the data.

$$p(\mathbf{t},\mathbf{x}) = p(\mathbf{x})p(\mathbf{t}|\mathbf{x}) = p(\mathbf{x})\prod_{d=1}^{D} p(t_d l\mathbf{x}) \tag{12.2}$$

The conditional distribution $p(\mathbf{t}|\mathbf{x})$ is then expressed in terms of a mapping from latent variables to data variables, so that

$$\mathbf{t} = \mathbf{y}(\mathbf{x};\mathbf{w}) + \mathbf{u} \tag{12.3}$$

where $\mathbf{y}(\mathbf{x};\mathbf{w})$ is a function of the latent variable $\mathbf{x}$ with parameters $\mathbf{w}$, and $\mathbf{u}$ is an $\mathbf{x}$-independent noise process. If the components of $\mathbf{u}$ are uncorrelated, the conditional distribution for $\mathbf{t}$ will factorize as in (12.2). From the geometrical point of view, the function $\mathbf{y}(\mathbf{x};\mathbf{w})$ defines a manifold in the data space given by the image of the latent space. The definition of the latent variable model needs to be completed by specifying the distribution $p(\mathbf{u})$, the mapping $\mathbf{y}(\mathbf{x};\mathbf{w})$, and the marginal distribution $p(\mathbf{x})$. The type of mapping $\mathbf{y}(\mathbf{x};\mathbf{w})$ determines the specific latent variable model. The desired model for the distribution $p(\mathbf{t})$ of the data is then obtained by marginalizing over the latent variables:

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{x})p(\mathbf{x})d\mathbf{x}. \tag{12.4}$$

This integration will, in general, be analytically intractable except for specific forms of the distributions $p(\mathbf{t}|\mathbf{x})$ and $p(\mathbf{x})$. PPS define a non-linear, parametric mapping $\mathbf{y}(\mathbf{x};\mathbf{W})$, where $\mathbf{y}$ is defined continuous and differentiable, which projects every point in the latent space to a point into the data space. Since the latent space is Q-dimensional, these points will be confined to a Q-dimensional manifold non-linearly embedded into the D-dimensional data space. This implies that data points projecting near a principal surface node (i.e., a Gaussian center of the mixture) have higher influences on that node than points projecting far away from it (cf. Fig. 12.3).

Each of these nodes $\mathbf{y}(\mathbf{x};\mathbf{w})$, $\mathbf{x} \in \{\mathbf{x}_m\}_{m=1}^{M}$ has covariance expressed by:

$$\mathbf{\Sigma}(\mathbf{x}) = \frac{\alpha}{\beta}\sum_{q=1}^{Q}\mathbf{e}_q(\mathbf{x})\mathbf{e}_q^T(\mathbf{x}) + \frac{(D-\alpha Q)}{\beta(D-Q)}\sum_{d=Q+1}^{D}\mathbf{e}_d(\mathbf{x})\mathbf{e}_d^T(\mathbf{x}), \tag{12.5}$$

$$0 < \alpha < \frac{D}{Q}$$

where

- $\{\mathbf{e}_q(\mathbf{x})\}_{q=1}^{Q}$ is the set of orthonormal vectors tangential to the manifold at $\mathbf{y}(\mathbf{x};\mathbf{w})$,
- $\{\mathbf{e}_d(\mathbf{x})\}_{d=Q+1}^{D}$ is the set of orthonormal vectors orthogonal to the manifold in $\mathbf{y}(\mathbf{x};\mathbf{w})$.

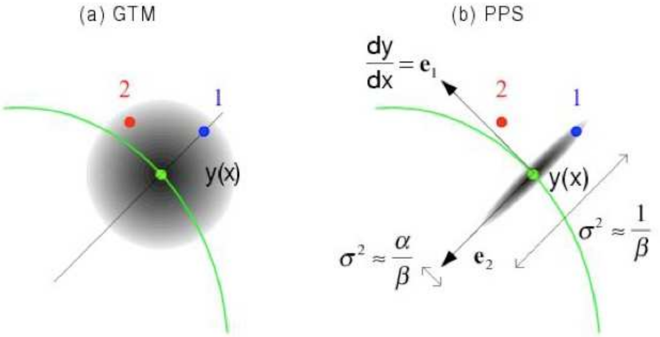

(a) GTM
2
1
y(x)
(b) PPS
$\frac{dy}{dx} = \mathbf{e}_1$
2
1
y(x)
$\sigma^2 \approx \frac{1}{\beta}$
$\sigma^2 \approx \frac{\alpha}{\beta}$
$\mathbf{e}_2$

Figure 12.3: *Graphical representation of fundamental vectors defining the oriented covariance ellipse and the covariance components for the Probabilistic Principal Surfaces algorithm. On the left the same image for the Generative Topographic Mapping algorithm and its spherical covariance is shown.*

The complete set of orthonormal vectors $\{\mathbf{e}_d(\mathbf{x})\}_{d=1}^{D}$ spans $\mathbb{R}^D$ and the parameter α is a clamping factor and determines the orientation of the covariance matrix. The unified *PPS* model reduces to GTM for $\alpha = 1$ and to the manifold-aligned GTM for $\alpha > 1$:

$$\boldsymbol{\Sigma}(\mathbf{x}) = \begin{cases} 0 < \alpha < 1 & \perp \text{ to the manifold} \\ \alpha = 1 & I_D \text{ or spherical} \\ 1 < \alpha < D/Q & \| \text{ to the manifold.} \end{cases}$$

In order to estimate the parameters $\mathbf{W}$ and β we used the Expectation–Maximization (EM) algorithm (Dempster et al. 1977), while the clamping factor is fixed by the user and is assumed to be constant during the EM iterations. In a $3D$ latent space, then, a spherical manifold can be constructed using a PPS with nodes $\{\mathbf{x}_m\}_{m=1}^{M}$ arranged regularly on the surface of a sphere in $\mathbb{R}^3$ latent space, with the latent basis functions evenly distributed on the sphere at a lower density. The motivation behind such a spherical manifold is that spherical PPS are particularly well suited to capture the sparsity and periphery of data in large input spaces (Bishop 1995). In order to better explain this issue let us consider the following low-D analogy first proposed by (Chang and Ghosh 2000): *... imagine fitting a rubber band ($2-D$ spherical manifold) to data distributed uniformly on the surface of a sphere in $\mathbb{R}^3$. Any fit bisecting the sphere into two equal halves will be optimal. On the other hand, consider using a piece of string to fit the same data. The string has a significantly* *lower probability of finding the optimal fit as it is open-ended...* After a spherical PPS model is fitted to the data, the data themselves are projected into the latent space as points onto a sphere (Fig. 12.4).

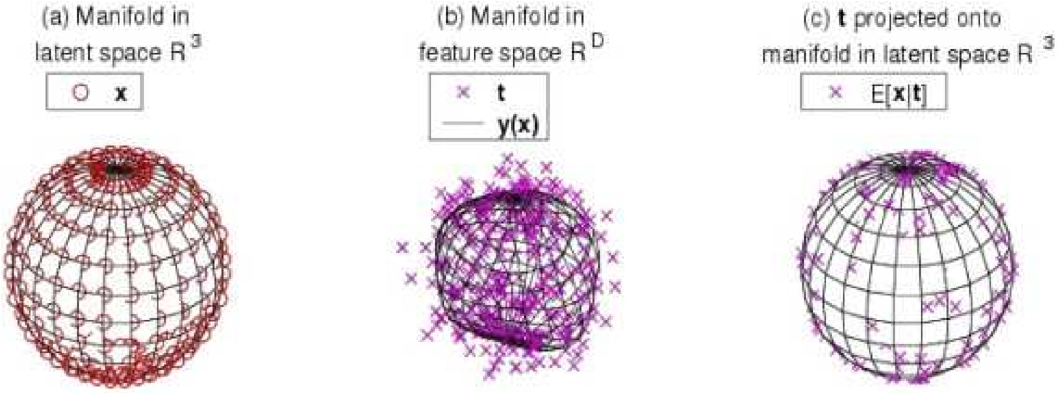


Figure 12.4: *Schematic representation of the spherical manifold in the three dimensional latent space R^3 (a), the same manifold distorted in the feature space R^D together with points associated to data (b), and the projection of the points distribution onto the surface of the spherical manifold embedded in R^3 latent space.*

The latent manifold coordinates $\hat{\mathbf{x}}_n$ of each data point $\mathbf{t}_n$ are computed as:

$$\hat{\mathbf{x}}_n \equiv \langle \mathbf{x} | \mathbf{t}_n \rangle = \int \mathbf{x} p(\mathbf{x}|\mathbf{t}) d\mathbf{x} = \sum_{m=1}^{M} r_{mn} \mathbf{x}_m$$

where r_{mn} are the latent variable responsibilities defined as:

$$llr_{mn} = p(\mathbf{x}_m|\mathbf{t}_n) = \frac{p(\mathbf{t}_n|\mathbf{x}_m)P(\mathbf{x}_m)}{\sum_{m'=1}^{M} p(\mathbf{t}_n|\mathbf{x}_{m'})P(\mathbf{x}_{m'})} \tag{12.6}$$

$$= \frac{p(\mathbf{t}_n|\mathbf{x}_m)}{\sum_{m'=1}^{M} p(\mathbf{t}_n|\mathbf{x}_{m'})} \tag{12.7}$$

Since $\|\mathbf{x}_m\| = 1$ and $\sum_m r_{mn} = 1$, for $n = 1, \ldots, N$, these coordinates lay within a unit sphere, i.e. $\|\hat{\mathbf{x}}_n\| \leq 1$.

An interesting issue is the assessment of the incidence of each input data feature on the latent variables which helps to understand the relation between the features and the clusters found. The feature incidences are computed by evaluating the probability density of the input vector components with respect to each latent variable. More specifically, let $\{\mathbf{t}_n\}_{n=1}^{N}$ be the set of the D-dimensional input data, i.e $\mathbf{t}_n = (t_{n1}, \ldots, t_{nD}) \in \mathbb{R}^D$, and $\{\mathbf{x}_m\}_{m=1}^{M}$ be the set of latent variables with $\mathbf{x}_m \in \mathbb{R}^3$. For each data point $\mathbf{t}_n = (t_{n1}, \ldots, t_{nD})$ we want to evaluate $p(t_{ni}/t_{n1}, \ldots, t_{ni-1}, t_{ni+1}, \ldots, t_{nD}, \mathbf{x}_m)$, for $m = 1, \ldots, M$ and $i = 1, \ldots, D$. In detail:

$$p(t_{ni}/t_{n1},\ldots,t_{ni-1},t_{ni+1},\ldots,t_{nD},\mathbf{x}_m) = \tag{12.8}$$

$$= \frac{p(t_{n1},t_{n2},\ldots,t_{nD},\mathbf{x}_m)}{p(t_{n1},\ldots,t_{ni-1},t_{ni+1},\ldots,t_{nD},\mathbf{x}_m)} = \tag{12.9}$$

$$= \frac{p(t_{n1},\ldots,t_{nD}/\mathbf{x}_m)P(\mathbf{x}_m)}{p(t_{n1},\ldots,t_{ni-1},t_{ni+1},\ldots,t_{nD}/\mathbf{x}_m)P(\mathbf{x}_m)} = \tag{12.10}$$

$$\frac{p(t_{n1},\ldots,t_{nD}/\mathbf{x}_m)}{p(t_{n1},\ldots,t_{ni-1},y_{ni+1},\ldots,t_{nD}/\mathbf{x}_m)}. \tag{12.11}$$

The last term is easily obtained since the numerator is simply the m-th Gaussian component of the mixture computed by the PPS model with mean $y(\mathbf{x}_m;\mathbf{W})$ and oriented variance Σ_m, while the denominator is the same Gaussian component in which the i-th component is missing. Finally the mean of expression (12.8) over the N input data points, for each $\mathbf{x}_m$, is computed. This explains why spherical PPS can be used as a "reference manifold" for classifying high-D data. A reference spherical manifold is computed for each class during the training phase. In the test phase, a data previously unseen by the network is classified to the class of its nearest spherical manifold. Obviously, the concept of "nearest" implies a distance computation between a data point $\mathbf{t}$ and the nodes of the manifold. Before doing this computation, the data point $\mathbf{t}$ must be linearly projected onto the manifold. Since a spherical manifold consists of square and triangular patches, each one defined by three or four manifold nodes, what is computed is an approximation of the distance. The PPS framework provides three approximation methods:

- Nearest Neighbour: finds the minimal square distance to all manifold nodes;
- Grid Projections: finds the shortest projection distance to a manifold grid;
- Nearest Triangulation: finds the nearest projection distance to the possible triangulation;

In what follows we have used the Nearest Neighbour approximation method because it allows to evaluate distances of each data point in the feature space to all nodes embedded in the spherical manifold; even if computationally heavier than the other two methods, the Nearest Neighbour approximation provides the most trustworthy choice of the node (or nodes, in case of multiple nodes at the same distance from a given point) that each data point has to be assigned to. Another way to use PPS as classifiers consists in choosing the class C with the maximum posterior class probability for a given new input $\mathbf{t}$. Formally speaking, let us suppose to have N labelled data points $\{\mathbf{t}_1,\ldots,\mathbf{t}_N\}$, with $\mathbf{t}_i \in \mathcal{R}^D$ and labels *class* in the set $\{1, \ldots, C\}$, then the posterior probabilities may be derived from the class-conditional density $p(\mathbf{t}|\mathit{class})$ via the Bayes theorem:

$$P(\mathit{class}|\mathbf{t}) = \frac{p(\mathbf{t}|\mathit{class})P(\mathit{class})}{p(\mathbf{t})} \propto p(\mathbf{t}|\mathit{class})P(\mathit{class}).$$

In order to approximate the posterior probabilities $P(\mathit{class}|\mathbf{t})$ we estimate $p(\mathbf{t}|\mathit{class})$ and $P(\mathit{class})$ from the training data. Finally, an input $\mathbf{t}$ is assigned to the class with maximum $P(class|\mathbf{t})$. In (Staiano 2003) and (Chang and Ghosh 2000) the effectiveness of PPS classifier is reported. A more detailed exposition of PPS as data mining framework can be found in (Staiano 2003, Staiano 2004).

12.4.2 PPS as a clustering tool

It needs to be explicitly noted that, as already mentioned, even though strictly speaking PPS are not a clustering algorithm, they can be effectively used for clustering purposes. Each latent variable, in fact, defines an attractor for points which are projected near to it and therefore the input space is partitioned in a number of clusters coinciding with the number of latent variables. The number of latent variables can therefore be regarded as the initial 'resolution' of clustering process but, provided that this number is not too low or too high (to avoid respectively a rough and imprecise or sparse clustering), it is found empirically that every reasonable choice leads to consistent results. In fact it suffices to set it to a value higher than the number of clusters realistically expected to be present in the data and then to use an agglomerative algorithm capable to recombine clusters of points artificially split into two or more clusters of smaller size. In our case we used the Negative Entropy Clustering described in the following paragraph.

12.4.3 The hierarchical clustering algorithm

Most unsupervised methods require the number of clusters to be provided *a priori*. This circumstance represents a serious problem when exploring large complex data sets where the number of clusters can be very high or, in any case, largely unpredictable. A simple threshold criterion is not satisfactory in most astronomical applications due to the high degeneracy and noisiness of the data which usually lead to the erroneous agglomeration of data. A classical agglomerative clustering algorithm is completely specified by assigning a definition of distance between clusters and a linkage strategy, i.e. a rule according to which clusters separated by some value of the distance are merged and others are not. Several definitions of distances can be found in the literature (distance between centroids of the clusters, maximum distance between members of the clusters, minimum distance, etc.) and many linkage strategies are used for common tasks (for example simple linkage, average, complete, Ward's, etc.). Successive generations of merging are carried out using updated distances and the resulting structure of clusters can be represented using a tree-like graph, called dendrogram, until some convergence criterion is satisfied (usually, until no other clusters can be merged according to the definition of distance).

The interpretation of distances between clusters in the parameter space can be relaxed in order to generalize this class of algorithm. In this case the "distance" between clusters becomes a generic function of the position in the parameter space of the members of each cluster. We chose an approach to the hierarchical clustering based on the combination of a similarity criterion founded on the notion of 'Negative Entropy' and the use of a dendrogram to investigate the structure of clusters produced by the agglomerative algorithm.

We made use of the Fisher's linear discriminant which is a classification method that first projects high-dimensional data onto a line, and then performs a classification in the projected one-dimensional space (Bishop 1995). The projection is performed in such a way to maximize the distance between the means of the two classes while minimizing the variance within each class. On the other hand, we define the differential entropy H of a random vector

$$\mathbf{y} = (y_1, \ldots, y_n)^T$$

with density $f(.)$ as:

$$H(\mathbf{y}) = \int f(\mathbf{y}) \log f(\mathbf{y}) d\mathbf{y}$$

so that negentropy J can be defined as:

$$J(\mathbf{y}) = J(\mathbf{y}_{Gauss}) - H(\mathbf{y})$$

where $\mathbf{y}_{Gauss}$ is a Gaussian random vector of the same covariance matrix as $\mathbf{y}$.
The Negentropy can be interpreted as a measure of non-gaussianity and, since it is invariant for invertible linear transformations, it is obvious that finding an invertible transformation that minimizes the mutual information is roughly equivalent at finding directions in which the Negentropy is maximized. Our implementation of the method uses an approximation of Negentropy that provides a good compromise between the properties of the two classic non-gaussianity measures given by Kurtosis and Negentropy.

Negentropy clustering algorithm can be used to perform unsupervised agglomeration of the clusters (or "preclusters") found by the PPS algorithm during the first step of our method. The only *a priori* information needed by NEC is a value of the dissimilarity threshold T. We suppose to have n D-dimensional preclusters X_i with $i = 1, \ldots, n$ that have been determined by the PPS; these clusters are passed to the Negentropy Clustering algorithm which, in practice, ascertains whether each couple of contiguous clusters (according to the Fisher's linear discriminant) can or cannot be more efficiently modelled by one single multivariate gaussian distribution. In other words, NEC algorithm determines if two clusters belonging to a given couple can be considered to be substantially distinct or parts of a greater more general data set (i.e. cluster). It needs to be stressed that this method can be easily generalized to other models; we preferred to use D-dimensional Gaussians only because the normal distribution can be considered a good approximation of any reasonably shaped peaked distribution, since the colours of objects belonging to the same observational family of quasars are widespread around a central value due to several physical mechanism (differential scattering, absorption, etc.).

12.4.4 The labelling phase

The results of the agglomeration performed by the NEC algorithm depend crucially on the value of the dissimilarity threshold T. Since to different values of this constant correspond different clusterings of the dataset, it is necessary to apply an objective criterion for the determination of the best (hereafter critical) value of the dissimilarity threshold T_{cr}, i.e the value leading the best performance in terms of the selection algorithm completeness and efficiency. To this aim, we use the BoK to label as "goal-successful" those clusters C_j (with $j \in \{1, ..., N_{cl}\}$ (where N_{cl} is the total number of clusters for a given value of T), for which the following relation is satisfied:

$$sr_j^{(g)} = \frac{\text{Nm confirmed "goal" members in } C_j}{\text{Nm members in } C_j} \geq \widetilde{sr}^{(g)} \tag{12.12}$$

i.e., the fraction $sr^{(g)}$ of "goal" objects contained in the cluster C_j must be higher then a given value $\widetilde{sr}^{(g)}$. At the same time "not-goal-successful" clusters are defined as clusters for which the following relation holds:

$$sr_j^{(ng)} = \frac{\text{Nm confirmed "not-goal" members in } C_j}{\text{Nm members in } C_j} \geq \widetilde{sr}^{(ng)} \tag{12.13}$$

with a similar meaning of the symbols.

In other words, "goal-successful" ("not-goal-successful") clusters are defined as the clusters containing "goal" ("not-goal") objects fractions above a given threshold.

A third type of clusters which we shall simply call "not successful" are those which do not fulfil any of the two above definitions as they are formed by comparable fractions of "goal" and "not-goal" objects. In the specific case addressed here, they will be composed by a mixture of confirmed quasar and other type of objects (mainly stars).

The critical value T_{cr} of the dissimilarity threshold is therefore defined as the one which, given a set of initial clusters provided by the PPS algorithm, produces the maximum number of "goal-successful" clusters, or, in a more quantitative fashion, as the value which maximizes the normalized success ratio NSR:

$$NSR(T) = \frac{\text{Nm successful clusters}}{\text{Nm clusters}} \tag{12.14}$$

Two further requirements are imposed to select T_{cr}: a stability or robustness criterion which translates to the fact that the number of clusters does not change by slightly perturbing T_{cr}, and that the number of clusters produced ranges between 25 $\%$ and 75 $\%$ of the number of initial clusters. This last constraint excludes from selection values of T_{cr} producing unreasonable numbers of clusters, namely an excessive number of poor clusters or few very reach clusters (for a detailed discussion see Section 12.6.3).

The process described hitherto is recursive: once $\widetilde{sr}^{(g)}$ and $\widetilde{sr}^{(ng)}$ have been fixed, the suitable value of the dissimilarity threshold is identified and a first clustering is performed using T_{cr} as an input to the NEC algorithm. All successful clusters produced in this first generation of clustering are 'frozen' and the efficiency e_1 is estimated, and unsuccessful clusters are merged together and form the input data set for the subsequent iteration. After this second iteration, the new successful clusters, if any is found, are retrieved and stored. The procedure is iterated until no other successful cluster is found.

Critical values of the dissimilarity threshold for each generation are fixed accordingly to the same criteria explained above. The efficiency e_{tot} of the selection algorithm is defined as the sum of the efficiencies of each of M generations weighted according to the total number of objects belonging to goal-successful clusters of that generation:

$$e_{tot} = \frac{\sum_{i=1}^{M} n_i e_i}{\sum_{i=1}^{M} n_i} = \frac{\sum_{i=1}^{M} n_i^{(goal)}}{N_{tot}^{(goal)}} \tag{12.15}$$

where n_i is the total number of objects belonging to "goal-successful" clusters of the i_{th} generation such that $\sum_{i=0}^{M} n_i = N_{tot}^{(goal)}$ and $n_i^{(goal)}$ is the number of confirmed goal objects contained in all "goal-successful" clusters selected in the i-th generation.

The total completeness c_{tot} of the process is defined as follows:

$$c_{tot} = \frac{\sum_{i=1}^{M} n_i^{(goal)}}{N_{tot}^{(all)}} \tag{12.16}$$

where $N_{tot}^{(all)}$ is total number of goal objects contained in the dataset used for the experiment. Extensive testing showed that, within the range $[0.65, 0.90]$ the values of the constant thresholds $\widetilde{sr}^{(g)}$ and $\widetilde{sr}^{(ng)}$ for "goal-successful" and "not-goal-successful" clusters respectively, do not affect the final efficiency and completeness of the candidate quasars algorithm, but only the number of generations of the process needed to achieve the final result.

12.4.5 Selection of candidate quasars from photometric samples.

After the labelling phase, which has provided the most suitable partition of the parameter space in terms of selection of ”goal-successful” and ”not-goal-successful” clusters, candidate quasars extraction from a purely photometric dataset (i.e., for which no spectroscopic BoK is available) can be carried out using one of the two different approaches outlined below.

Method I

The first method is based on the assumption that confirmed QSOs in the BoK are tracers of successful clusters containing mainly goal objects even when other objects are added to the sample. The dataset used for the labeling and the photometric sample are merged and the whole process described above is repeated using this extended group of sources. The selection of candidate quasars is then carried out considering as candidates all non-BoK objects belonging to clusters where spectroscopic confirmed quasars (belonging to the former sample used for the labeling phase) are dominant. This simple and, at least in theory, straightforward method unfortunately is applicable only when the non-BoK sample is composed by few objects, namely a small fraction of the number of BoK objects. The reason is that PPS algorithm determines the best projection from the parameter space to the latent space by modelling a probability distribution which is a function of the initial distribution of BoK sample objects inside the initial space. New objects added to the labelling sample modify the shape of the probability density function and the final result of the unsupervised clustering, so that the efficiency and completeness estimated during the labeling phase are not appropriate.

It needs to be stressed that, this method has to be preferred when limited amount of data are added to an existing data set but it cannot be used when the amount of new data to be processed is large.

Method II

The second approach (which is conceptually similar to the one described in ((Richards et al. 2004)), is based on the characterization of the distribution in the parameter space of the objects belonging to each successful cluster. The first step consists in the determination of the parameters (i.e., in geometrical terms the axes of the parameter space) carrying most information about the displacement of the different successful clusters selected by the algorithm after the labelling phase. A linear principal component analysis (PCA) is carried out to find out which parameters are responsible for most of the variance of the distribution of ”goal-successful” clusters members in the n-dimensional parameter space. Then, for each goal-successful cluster, a set of n constraints is identified. The constraints on the most significant parameters selected through PCA are very stringent since on the surface of the hyper–plane defined by these axes, the successful clusters are distinctly visible and well distinguishable. The constraints regarding the other parameters only require objects belonging to the photometric dataset to lay between the minimum and the maximum values of the distribution. This hybrid method is more flexible than the previous one since the selection of candidate quasars can be fine tuned by modifying the cuts applied to the parameters in order to achieve different performance goals. Obviously this implies a trade-off between efficiency and completeness. For instance, loose constraints allow the selection of a wider number of candidates in the outskirts of the clusters, where the contamination from "not-goal" sources is higher, thus resulting in a increased completeness but in a lower efficiency of the overall selection process. On the other hand, tight constraints increase the efficiency of the algorithm at the cost of a lower completeness, by selecting for scrutiny only the central regions of the parameter space occupied by successful clusters.

In the present work two different prescriptions have been used to determine parameters cuts. According to the first prescription, n-haedrons (where n is the dimensionality of the parameter space) whose vertices are fixed by the extremal values of cluster members distribution for each parameter for both "goal-successful" and "not-goal-successful" clusters, have been chosen to delimit the regions of the parameter space containing candidate quasars. More precisely, all photometric objects placed inside the n-haedrons derived from "goal-successful" clusters and not placed inside n-haedrons derived by "not-goal-successful" clusters are selected as candidate quasars. Errors on parameters have been used to estimate the distance of each object from the surfaces of the n-haedrons generated by "goal-successful" clusters, in order to avoid the possible contamination from spurious objects located near to the borders of the "goal-successful" region of the parameter space. Only objects with distance from the internal side of the surfaces defining the n-haedrons larger than 3σ have been retained as candidates. The second prescription is more conservative (in the sense of ensuring higher efficiency) than the first one: in order to minimize the fraction of contaminants selected inside the regions of the parameter space containing the "goal-successful" clusters, the vertices of the n-haedrons describing the positions of "not-goal-successful" clusters are set to the positions $\bar{x}_i \pm \sigma_i$ along each axis of the parameter space, where $\bar{x}_i$ is the average value and σ_i the standard deviation of the distribution of points along the i-th axis. All objects placed inside these n-haedrons are discarded and the remaining are selected as candidate according to the first prescription.

Table 12.2: *Total efficiency and completeness of the experiments.*

Experiment	Sample	e_{tot}	c_{tot}	n_{gen}
1	S-A	81.5%	89.3%	2
2	S-UK	92.3%	91.4%	1
3	S-UK	97.3%	94.3%	1
4	S-S	95.4%	94.7%	3

12.5 The experiments

Four different sets of experiments involving the samples described above have been performed.

- The first experiment is based on the S-A sample, composed by all SDSS star-like objects having spectroscopic classification and falling in the region of overlap with the UKIDSS-DR1 LAS, and their optical magnitudes are exploited to calculate colours.

- The second and third sets of experiments utilize the S-UK sample, *id est* all stellar objects in the SDSS having spectroscopic classification available and having an IR counterpart in the UKIDSS-DR1 LAS. We made use of optical only (second experiment) and optical plus infrared (third experiment) colours.

- The fourth set of experiments explored the distribution in the S-S parameter space formed by all candidate quasars in SDSS–DR5 for which all magnitudes have been measured. Optical colours derived by SDSS photometry have been used to determine the shape of the parameter space.

Colours used for all these experiments were calculated using adjacent bands: $u-g$, $g-r$, $r-i$, $i-z$ for the optical bands, and $Y-J$, $J-H$, $H-K$ for the near infrared ones. The dependence of the results on the choice of the colours will be discussed in Section 12.6.

The final results of these experiments in terms of total efficiency and completeness of the candidate quasars selection are summarized in table (12.5) while the number of successful clusters and the fraction of confirmed quasars and stars inside each "goal-successful" cluster for each experiment are reported in table (12.3).

12.5.1 First experiment

This experiment aimed at comparing the SDSS native selection algorithm to our method. The number of latent variables (i.e. initial clusters) used for the PPS pre-clustering was

62, and the critical value of the dissimilarity threshold was chosen according to the criteria explained above. The normalized success ratio and other statistical indicators of the clustering features are plotted as functions of the dissimilarity threshold in the upper left panel of Fig. (12.5.1). The estimated efficiency and completeness of the selection process are shown in the upper right panel of Fig. (12.5.1) also as functions of the dissimilarity threshold.

The distribution of sources as a function of redshift and spectroscopic classification ″specClass″ in the ″goal-successful″ clusters selected in this experiment is shown in the middle panel of Fig. (12.5.1). Finally, a ″box and whisker″ plot of the distribution of candidate quasars for each ″goal-successful″ cluster selected in this experiment is shown in the bottom panel of Fig. (12.5.1).

The three ″goal-successful″ clusters selected in this experiment are evenly distributed in terms of colours:

- The first cluster is formed by objects having $u - g$ ranging form 2.0 to 2.5 and the other colours averaging between 0.5 and 0 with little dispersions. These objects have redshift around 3 and they are mainly classified as high-redshift QSO's in terms of spectroscopic ″specClass″ classification, with a little contamination by misidentified stars (″specClass″ = 1).

- The second cluster contains the vast majority of objects and is mainly formed by normal QSO's (″specClass″ = 3), with redshift ranging from 0.8 to 2.0, except for a little fraction of stars. In terms of colours distribution, this cluster shows all colours having average values near 0.0 with a higher and asymmetric dispersion, caused by outliers mainly having higher colours values than the average.

- The third ″goal-successful″ cluster members are characterized by a very high $u - g$ values, compared to the average values of the others colours which are similar to those found for the first cluster. It is formed by only high redshift ($z > 3$) QSO's with spectroscopic classification ″specClass″ = 4.

12.5.2 Second experiment

The goal of the second experiment was the selection of candidate quasars inside the optical colour space, using a sample of star-like objects selected in both optical and near infrared catalogues. The number of latent variables used for PPS pre-clustering was again fixed to 62 in order to ease the comparison with the first experiment. The labelling process has been repeated as described in the previous section; the fraction of ″goal-successful″ clusters and other parameters of the clustering are shown in the

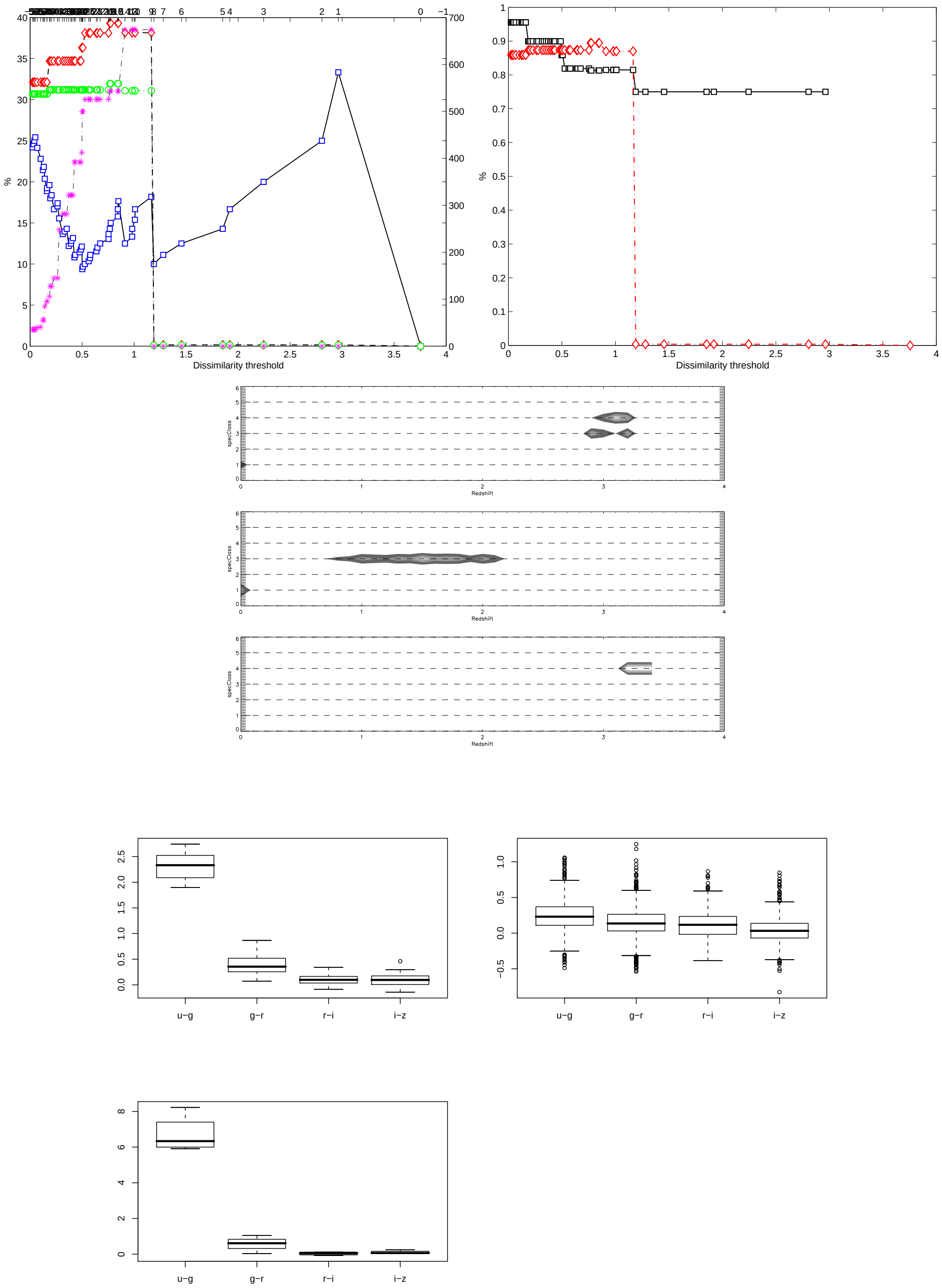


Figure 12.5: *Upper left: In this figure, the fraction of 'goal-successful' clusters (squares), the total percentage of confirmed objects belonging to 'goal-successful' clusters (circles), the total percentage of objects belonging to 'goal-successful' clusters irrespective of their spectroscopic classification (diamonds) and variance of 'goal-successful' clusters (asterisks) are plotted as a function of the dissimilarity threshold for the first experiment. Upper right panel: total estimated efficiency* e_{tot} *(square symbols) and completeness* c_{tot} *(diamond symbols) of the candidate quasars selection process as functions of the dissimilarity threshold for the first experiment. Middle panel: distribution of members of "goal-successful" clusters in the first experiment in the specClass-redshift plane. Lower panel: "box and whisker" plot of the distribution of different clusters produced in the first experiment. The black rectangle indicates the 95% confidence interval for the median while the greater rectangle is contained between the first and third interquartile. The notches extend to* ± 1.58 *of the interquartile range times the inverse of the square root of the number of objects. Outliers are represented as dots.*

upper left panel of figure (12.5.2) as functions of the dissimilarity threshold. The estimated total efficiency and completeness are also plotted as functions of the dissimilarity threshold in the upper right panel of figure 12.5.2. The distribution of sources as a function of redshift and spectroscopic classification ″specClass″ in the ″goal-successful″ clusters selected in this experiment is shown in the middle panel of figure (12.5.2) and the ″box and whisker″ plot of the distribution of candidate quasars for each of ″goal-successful″ clusters is given in the bottom panel of figure (12.5.2). The results of this experiment show that our method can reach a significantly higher efficiency and completeness level respect to the optical-infrared candidate selection algorithms found in the literature (Richards et al. 2002, Richards et al. 2004, Warren et al. 1991), using a base of knowledge formed by both ″specClass″ for SDSS sources with spectroscopic classification but not selected as quasars candidates and spectroscopic classification of quasars external to SDSS.
For what the ″goal-successful″ clusters are concerned, we can notice that:

- The first cluster is composed of sources with optical $u - g$ colour concentrated around 1.0, while the other optical colours $g-r$, $r-i$ and $i-z$ have average values falling from around 0.7 to 0.5 respectively. The infrared colour $Y - J$ averages at about 0.5, while $J - H$ and $H - K$ increase to mean values of about 0.8 and 1.0. The redshift distribution of this cluster members, all ranked as normal QSO's according to the SDSS spectroscopic classification (″specClass″ = 3), shows three different groups of objects situated approximately at $z \sim 0.5, 1.1$ and 2.2.

- The second cluster is characterized by a colour distribution similar to the one of the first cluster, except for the optical $u - g$ colour whose mean value increases to 2.0. This cluster is composed of equal fractions of normal and far QSO's according to the ″specClass″ index, with a distribution spanning about 0.5 in redshift around 3.

- The third cluster is formed by far QSO's only, with a higher redshift ($z \sim 3.3$) mean value and a colour distribution similar to the second cluster except for a much higher value of $u - g$, with well defined mean at approximately 6.1.

- The last cluster is composed mainly by ″specClass″ = 2 sources inside a large redshift interval spanning from ~ 0.7 to 2.2. All colours have mean values included between 0.0 and 0.5.

The clusters from second to fourth appear to be very similar to the ″goal-successful″ clusters selected in the first experiment both in terms of colours and redshift distribution and spectroscopic type composition.

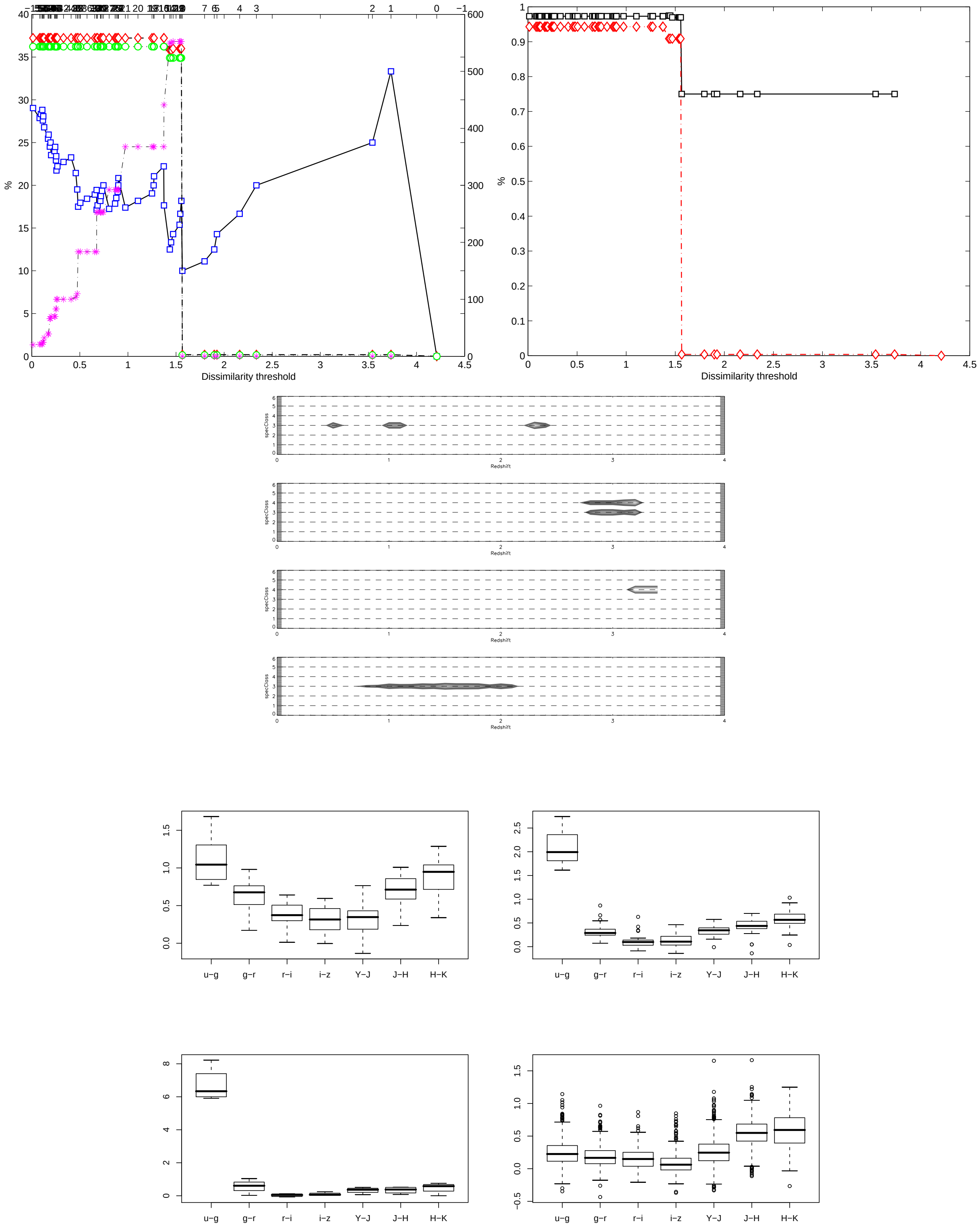


Figure 12.6: *Same as in previous figure but for the second experiment.*

Table 12.3: *Description of the contents of "goal-successful" clusters selected in the experiments described. For each cluster, the total number of members n and the fraction of confirmed quasars are reported.*

Exp.	Cluster	% quasars	n
1	1	75.9%	29
	2	81.5%	957
	3	75.1 %	4
2	1	76.1%	652
	2	83.3%	48
	3	100.0%	4
3	1	92.3%	26
	2	93.5%	31
	3	75.0%	4
	4	97.7%	755
4	1	86.2%	2121
	2	93.1%	52190
	3	77.9%	2433
	4	75.8%	198
	5	78.6%	126
	6	90.6%	171
	7	91.5.%	298
	8	78.9%	90
	9	79.0%	76
	10	86.1%	92

12.5.3 Third experiment

The third experiment was carried out in order to test whether the addition of the near infrared colours to the optical colours already used as parameters for the first two experiments improves or not the total efficiency and completeness of candidate quasars selection. In conformity to the previous experiments, the number of latent variables was fixed to 62, resulting in an equal number of pre-clusters produced. As in the previous experiments, all relevant information are reported in Figure (12.5.3).
The results of this experiment are summarized in the following description of the "goal-successful" clusters selected.

- The mean values of colours distribution of the members of the first cluster range from 0.0 to 0.5 with outliers mainly situated at higher values, while the distribution of normal QSO's in redshift spans from 0.7 to 2.2 and a little contamination

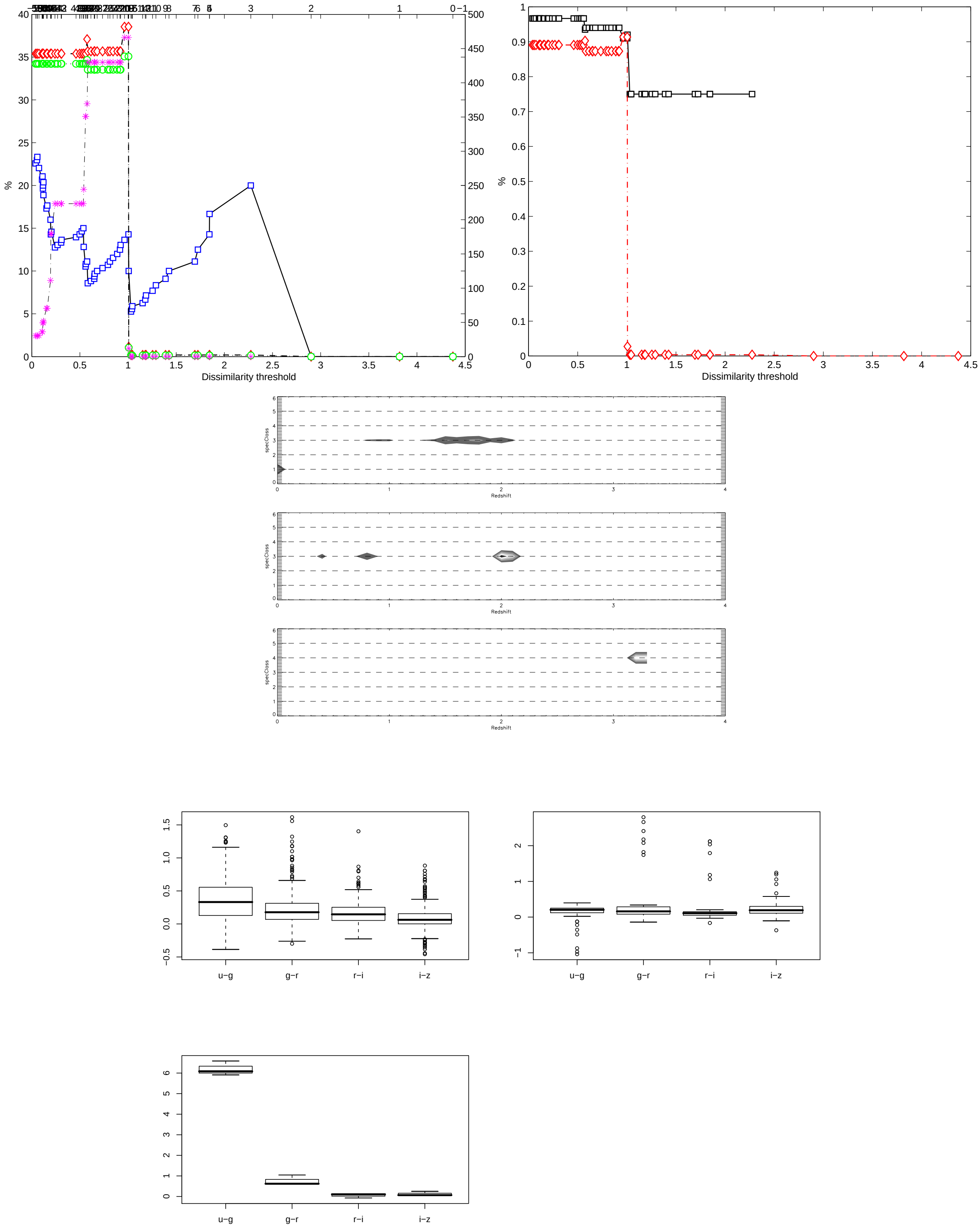


Figure 12.7: *Same as in previous figure but for the Third Experiment.*

from ″specClass = 1″ stars is present.

- The second cluster, entirely composed by ″specClass″ = 2 sources, shows a colours distribution almost identical to the previous cluster, while the distribution in redshift peaks at $z \sim 0.4, 0.7$ and 2.1.
- The third clusters contains sources with a value of $u-g \sim 6$ and $g-r \sim 1$, while the other two colours have distributions consistent with zero. This cluster is formed by far QSO's with redshift higher than 3.0.

Also in this case, the clusters selected are very similar in terms of colours and redshift distributions to the ″goal-successful″ clusters selected in the previous two experiments.

12.5.4 Fourth experiment

The fourth experiment was carried out as an application to the SDSS candidate quasars dataset of the algorithm described in this paper. Also in this case the number of latent variables for the PPS algorithm was fixed to 62. Results are shown in figure (12.5.4).

- The first cluster is composed mainly by normal QSO's situated at low redshift, having $u-g$, $r-i$ and $i-z$ colours with means ranging form 0.0 to 0.5, and the $g-r$ colour with an average value slightly higher and centred on ~ 1.0.
- The distributions of colours of the members of the second cluster all peak at about 0.0 with outliers reaching higher values especially in $r-i$ and $i-z$ colours. The distribution in redshift of the sources of this cluster, formed exclusively by ″specClass″ = 3 normal QSO's, spans the whole interval from ~ 0.0 to 2.2.
- The remaining ″goal-successful″ clusters selected in this experiment show as common feature distributions of $u-g$ with higher means than previous clusters, ranging from $\sim$ 2 to ~ 5.5. The $g-r$ colours are instead distributed between 1.8 and ~ 0.2 and the others colours have similar mean values around 0. All these clusters are formed by a mixture of normal and far QSO's according to the SDSS spectroscopic classification, with redshifts going from 3 to 4.2.

12.6 The selection of the parameters

In this section we shortly discuss how the performances of our candidate quasars selection method depend upon the assumed parameters of the PPS and NEC algorithms respectively, and on the set of features used for the characterization of the distribution of

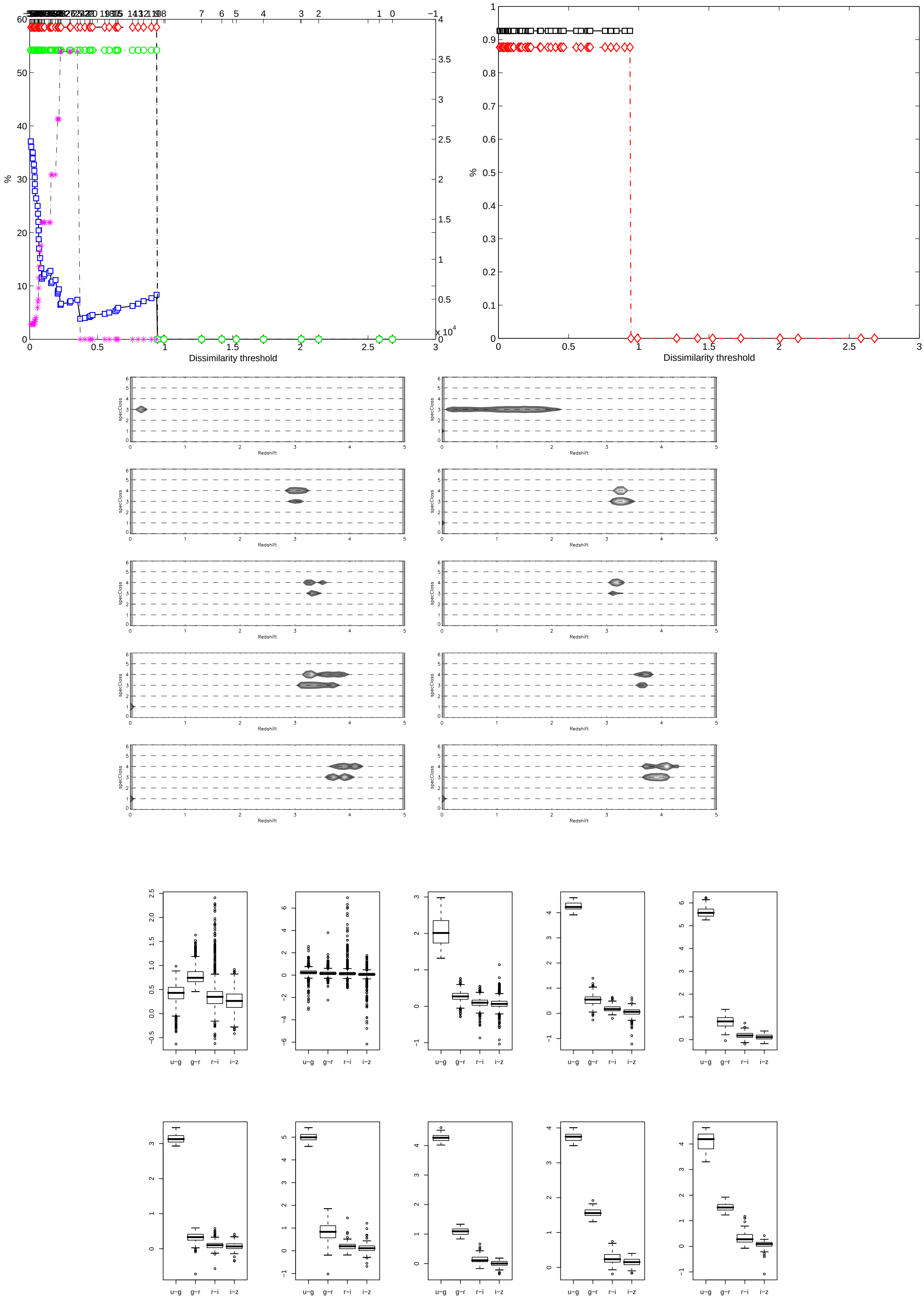


Figure 12.8: *Fourth experiment.*

objects inside the parameter space. To be specific, we have focused our attention on the dependence of the PPS algorithm from the number of latent variables (i.e. the number of preclusters produced by PPS), the dependence of the NEC algorithm from the critical value of the dissimilarity threshold T_{cr}, and the dependence of the overall method from the particular set of colours used to define the parameter space inside which clustering is performed.

12.6.1 Dependence on PPS parameter

As it has been discussed elsewhere, the PPS performances are rather independent on the choice of the parameters (clamping factor, width and orientation of principal surfaces, tolerance and number of iterations), with the exception of the number of latent variables. This number needs to be neither too large nor too small (within rather large boundaries) An excessive number of latent variables produces clusters that are not agglomerated efficiently by the NEC algorithm, while setting the number of latent variable to a low number does not allow a proper separation of different groups of objects in distinct clusters. In absence of theoretical ways to estimate the correct number of latent variables, the only way is to guess it through a trial and error procedure. It needs to be stressed that even though rather demanding in terms of computing time, for a given problem and data set, this procedure needs to be run only once.

12.6.2 Dependence on colours choice

In order to detect any bias introduced by the choice of a particular colours set in the results of our experiments, all experiments have been repeated using colours derived by different combination of magnitudes (table 12.6.2) and keeping unchanged all other parameters. These tests showed that the total efficiency e_{tot} and completeness c_{tot} are robust with respect to the particular parameter space where the distribution of objects is studied. Fluctuations affecting both e_{tot} and c_{tot} for all possible sets of parameters are, in the worst cases, comparable with few percents of the optimal values obtained using the natural combination of colours. This result can be understood from a theoretical point of view reminding that, given an initial set of colours C_0 derived from a certain photometric system, all other possible colours sets C_i can be expressed as linear combinations of the members of C_0, so that each parameter space Σ_i generated by C_i is the result of the application of a rigid rotation to the the parameter space Σ_0 associated to C_0. The transformation applied to the parameter space does not affect the relative positions distances between points of the distribution and, as a consequence, the principal curves or surfaces generated by PPS algorithm in order to determine the best projection from the generic parameter space Σ_i to the latent space remain unchanged.

Experiment	Colours	e_{tot}	c_{tot}
1	natural	81.5%	89.3%
1	$(u-r, g-i, r-z, i-u)$	81.7%	89.5%
1	$(u-i, g-z, r-g, i-r)$	80.8%	89.3%
1	$(u-z, g-u, r-z, i-r)$	82.0%	89.0%
1	$(u-g, g-z, r-i, i-z)$	81.4%	89.7%
2	natural	92.3%	91.4%
2	$(u-r, g-i, r-z, i-u, Y-H, J-K, H-Y)$	92.3%	91.5%
2	$(u-i, g-z, r-g, i-r, Y-K, J-Y, H-J)$	92.7%	91.8%
2	$(u-z, g-u, r-z, i-r, Y-H, J-K, H-Y)$	91.9%	90.9%
2	$(u-Y, g-H, r-J, i-K, z-u, Y-g, H-r)$	91.0%	91.0%
2	$(u-H, g-J, r-K, i-u, z-g, Y-z, H-i)$	90.9%	91.2%
2	$(u-J, g-K, r-u, i-g, z-r, Y-i, H-z)$	92.2%	91.5%
2	$(u-z, g-K, H-J, z-Y, r-u, z-i, i-H)$	92.6%	91.4%
3	natural	97.3%	94.3 %
3	$(u-r, g-i, r-z, i-u)$	97.1%	94.8 %
3	$(u-i, g-z, r-g, i-r)$	97.0%	93.9 %
3	$(u-z, g-u, r-g, i-r)$	97.3%	94.0 %
3	$(u-g, g-z, r-i, i-z)$	96.9%	94.9%
4	natural		
4	$(u-r, g-i, r-z, i-u)$	95.2%	93.9%
4	$(u-i, g-z, r-g, i-r)$	95.0%	94.0%
4	$(u-z, g-u, r-g, i-r)$	95.4%	94.4%
4	$(u-g, g-z, r-i, i-z)$	95.7%	94.6%

Table 12.4: *Efficiency and completeness for the experiments evaluated using sets of different colours. The natural combinations of colours corresponds, by definition, to ($u-g$, $g-r$, $r-i$, $i-z$) for the experiments making use of only optical colours, and ($u-g$, $g-r$, $r-i$, $i-z$, $Y-J$, $J-H$, $H-K$) for the experiment making use of both optical and infrared colours. Only a selection of all permutations is shown.*

12.6.3 Dependence on clustering algorithm parameter

The choice of the thresholds $\widetilde{sr}^{(g)}$ and $\widetilde{sr}^{(ng)}$ introduced in the definitions of "goal-successful" and "not-goal-successful" clusters, respectively, only affects the total time required by algorithm in order to converge, i. e. the number of generations needed to select all possible candidate quasars in a given sample. The only requirement is that these thresholds need to be included in an interval of reasonable values (in this case, we have tested values of this parameter in the range [0.65, 0.90]). Lower values of the thresholds would allow the algorithm to select many clusters with a very high level of contamination from ″not-goal″ objects (″goal″ objects) at the cost of a reduction of the overall efficiency, while higher values would make practically impossible the selection of a major fraction of candidates, since only few clusters composed almost by only ″goal″ (″not-goal″) sources would conform to the definition.

12.7 Conclusions

We have presented a new unsupervised method to perform quasar candidate selection, based on the clustering of data in the parametric space defined by the photometric colours. The method requires a suitably large base of knowledge (BoK) which is used only for labelling purposes. The method consists of three steps: i) an unsupervised clustering performed by Probabilistic Principal Surfaces algorithm; ii) an agglomerative process driven by a measure of negative entropy, and iii) a fine tuning of the clustering performances through the exploitation of the information contained in the BoK. In the case discussed here, the BoK consists of the spectroscopic classification provided by the ″specClass″ SDSS flag which is available for a relatively small fraction of the objects. Extensive testing on different datasets (experiments) has shown that the method can achieve better performances than most methods published in the literature so far. The results obtained for the various experiments can be summarized as it follows:

- First experiment (optical data): the S-A sample composed by unresolved objects belonging to the ″Star″ table of the SDSS database and placed in the overlapping regions between SDSS DR1 and UKIDSS-DR1 LAS surveys, together with a BoK represented by the ″specClass″ spectroscopic classification index are used. Optical colours are employed as parameters. The best performance is reached with a total efficiency $e_{tot} = 81.5\%$ and a completeness $c_{tot} = 89.3\%$, within $n_{gen} = 2$ generations of clustering.
- Second experiment (optical + near infrared data): the S-UK sample composed by matching objects observed in both SDSS and UKIDSS-DR1 LAS surveys and classified as stars in both surveys, together with a BoK embodied by the SDSS ″specClass″ spectroscopic classification index are used. Optical colours are employed as parameters. The best performance is reached with a total efficiency $e_{tot} = 92.3\%$ and a completeness $c_{tot} = 91.4$, within $n_{gen} = 1$ generation of clustering. This experiment shows a significant improvement of the total efficiency and a slight improvement of the total completeness.
- Third experiment (optical + near infrared data): the same sample and BoK of the previous experiment are used. Optical and near infrared colours are employed as

parameters. The best performance is reached with a total efficiency $e_{tot} = 97.2\%$ and a completeness $c_{tot} = 94.3\%$, within $n_{gen} = 1$ generation of clustering. The addition of infrared photometric information notably improves both efficiency and completeness of the candidate quasars selection.

- Fourth experiment (optical data): the S-S sample composed by all the candidate quasars according to the native SDSS candidate quasars selection algorithm in the whole DR5 database "Target" table, and the same BoK of the previous experiments are used. Only optical colours are employed as parameters. The best performance is reached with a total efficiency $e_{tot} = 95.4\%$ and a completeness $c_{tot} = 94.7\%$, within $n_{gen} = 3$ generations of clustering.

If we consider the case of the extraction of a QSO candidate list from multi-band optical surveys data, the most significant experiment are the second and third ones, which take into account also the so called "optically dull" quasars, i. e. showing very little signature in the optical bands and are usually selected as quasars making use of additional information of spectroscopic nature or, like in this case, of photometric nature (near infrared colours). The method has also been applied to the SDSS-DR6 datasets using the setting of the fourth experiment in order to produce a list of candidate quasars which will be made available at the web page *http://people.na.infn.it/voneural/science/qsocandidates/*. We want to emphasize that, even though the method of unsupervised clustering in astronomical parameter spaces described in this paper has been adjusted to the case of quasars selection, such fine tuning depends only on the information contained in the BoK and it can be applied to any similar case provided that a suitably large and complete base of knowledge is available. In this context, the role played by the Virtual Observatory for the future evolution of the data mining approach to classification/selection problems in astronomy, will be extremely important because the VO will allow the construction of BoKs of unprecedented accuracy and completeness by federating and standardizing the information contained in most (if not all) astronomical databases worldwide, thus providing the most natural environment for the further development and exploitation of similar techniques. The application of the method described above to the old problem posed by the physical classification of galaxies using the BoK provided by SDSS is in progress and will be discussed elsewhere.

Acknowledgements: This work was funded through the Ph.D. program of the University Federico II in Napoli, the European VOTECH consortium and the PON-SCOPE (financed by the Italian MIUR). R.d'A. acknowledges the Marie Curie EARA-EST action for financial support during his visit at the Institute of Astronomy, Cambridge (UK). The authors also wish to thank M. Brescia, M. Paolillo, E. De Filippis, S. Cavuoti and

D. Capozzi for many useful discussions. A special thank goes to A. Staiano and A. Ciaramella for making available to us their implementations of the PPS and NEC codes, respectively.

Chapter 13

Where do we go next?

I think the end is the start.
Begin to feel very glad now:
All things are a part
All things are apart
All things are a part.

A plague of the Lighthouse Keeper, Van Der Graaf Generator

13.1 Future research interests

It has to be expected that the combination of a modern data-mining approach with the forthcoming deep surveys providing complete sets of both large and narrow filters in optical and infrared bands will further improve the selection performances and will require further refinements of techniques based on parameter space characterization. Hopefully, the epoch of almost unitary efficiency for quasars selection is just behind the corner, and for my previous experiences and preferences I am inclined to advance with my work this field of research. In general, I think that the basic idea behind the design of the algorithm described in chapter 12 will play a significant role in the future expansion of the data mining approach to the more general problem of classification of astronomical sources, which is embodied, on one hand, by the selection of particular families of objects, and on the other one to the challenge of the physical objective classification of galaxies. At the moment, in collaboration with other researchers from the Astronomy Department in Naples and using the tools and skills developed during my Ph.D., I am investigating several aspects related to the large scale distribution of galaxies in the nearby Universe. In particular, I am working on: i) the identification of galaxy groups and clusters in the SDSS, using a modified 3-dimensional "Friends of friends" algorithm capable to deal with photometric redshifts (Guglielmo et al. 2008, in prep.) ii) on the statistical description of clustering of galaxies and galaxy structures in redshift space through the evaluation of the multiplicity, luminosity and correlation functions of a sample of galaxies extracted from the 5th Data Release of SDSS using both spectroscopic and photometric redshifts; iii) on the characterization of compact group

of galaxies in the SDSS photometric data set by using our own estimates of photometric redshifts (Capozzi et al. 2007, submitted).
In consideration of my current expertise and research interests, I am highly motivated to keep working on the subject of the statistical characterization of the distribution of galaxies in the Universe. I need to stress, however, that the issue which I consider to be the most interesting and is mostly attracting my attention at the moment and on which I would like to keep working during the coming years is the study of obscured AGNs. The most efficient way of selecting obscured AGNs is through deep X-ray observations rather than optical or infrared ones (Treister et al. 2004). Deep X-ray surveys, however, are challenging and extremely time-consuming, since they cover only small areas of the sky, possibly affected by cosmic variance problems. Moreover, a fraction of heavily obscured quasars predicted by cosmological models of galactic formation and evolution does not show up in X-ray surveys and can be only detected in very deep infrared observations (Jones et al. 2006), covering small regions of the sky, and for this reason, the available samples of such objects are extremely incomplete and unsatisfactory. In light of this considerations, the search for counterparts of X-ray and infrared selected ″optically dull″ AGNs in other regions of the electromagnetic spectrum (mainly near infrared and visible bands) is important and complementary to X-ray and deep IR observations. The adaptation of the unsupervised clustering method described in chapter 12 to the selection of candidate obscured AGNs (from optical and infrared data set using a base of knowledge derived by deep X-ray observations) could considerably increase the number of presently known sources of this kind, allowing a better statistical characterization of such objects and a large number of follow-up studies regarding their spatial and luminosity distributions and SEDs properties. In more detail, the key fact is that the actual ratio of obscured to optically visible AGNs is still uncertain, and this uncertainty affects the determination of luminosity distribution of AGNs, their correlation function and the connection with the properties of their host galaxies. The discovery of larger samples of ″obscured″ AGNs in multi-wavelength data sets could help to address several cosmological and astrophysical questions still unanswered or poorly defined, and improve the reconstruction of an aspect of the phylogenetic interpretation of the evolution of galaxies and AGNs, sketched in paragraph 2.3, and whose knowledge is one of ultimate goals of extragalactic astronomy and observational cosmology. In particular, I am interested in the determination of bivariate luminosity functions and clustering properties of AGNs and in comparing the properties of such sample of selected objects with those of the purely optically selected ones. An improvement of the comprehension of the way AGNs are placed in both redshift and luminosity spaces will provide insights on the accretion history of the Universe, posing severe constraints on the cosmic ″downsizing″ model of AGNs evolution and providing information about on correlations between infrared/optical/X-ray properties of active galactic nuclei as a function

of the age of the Universe (Treister and Urry 2006). One of the most valuable consequences of an efficient photometric classification of Active Galactic Nuclei would be the characterization of the AGN spectral energy distribution, and consequently, a spectral classification of sources without the need for spectroscopy.

I wish to stress that one aspect which needs to be explored in this context and which is currently attracting my attention is the possible use of different clustering techniques which have been successfully implemented and tested in other fields of human endeavour but never applied to astronomical data sets. Some of these techniques (cf. Bregman co-clustering (Banerjee et al. 2004), which performs simultaneously the dimensionality reduction and the unsupervised clustering of the objects) are, in theory, also capable of solving the main problem affecting the data sets described above, namely the incompleteness (the presence of missing data, i.e. objects detected only in few bands). These algorithms have been so far limited by their computational cost but the advent of distributed computing seems to open new and challenging opportunities.

Bibliography

Abbasi, R. U. e. a.: 2004, Measurement of the Flux of Ultrahigh Energy Cosmic Rays from Monocular Observations by the High Resolution Fly's Eye Experiment, *Physical Review Letters* **92**(15), 151101–+.

Abu-Zayyad, T. et al.: 2005, Measurement of the spectrum of UHE cosmic rays by the FADC detector of the HiRes experiment, *Astropart. Phys.* **23**, 157–174.

Adelman-McCarthy, J. K., Agüeros, M. A., Allam, S. S., Anderson, K. S. J., Anderson, S. F., Annis, J., Bahcall, N. A., Baldry, I. K., Barentine, J. C., Berlind, A., Bernardi, M., Blanton, M. R., Boroski, W. N., Brewington, H. J., Brinchmann, J., Brinkmann, J., Brunner, R. J., Budavári, T., Carey, L. N., Carr, M. A., Castander, F. J., Connolly, A. J., Csabai, I., Czarapata, P. C., Dalcanton, J. J., Doi, M., Dong, F., Eisenstein, D. J., Evans, M. L., Fan, X., Finkbeiner, D. P., Friedman, S. D., Frieman, J. A., Fukugita, M., Gillespie, B., Glazebrook, K., Gray, J., Grebel, E. K., Gunn, J. E., Gurbani, V. K., de Haas, E., Hall, P. B., Harris, F. H., Harvanek, M., Hawley, S. L., Hayes, J., Hendry, J. S., Hennessy, G. S., Hindsley, R. B., Hirata, C. M., Hogan, C. J., Hogg, D. W., Holmgren, D. J., Holtzman, J. A., Ichikawa, S.-i., Ivezić, Ž., Jester, S., Johnston, D. E., Jorgensen, A. M., Jurić, M., Kent, S. M., Kleinman, S. J., Knapp, G. R., Kniazev, A. Y., Kron, R. G., Krzesinski, J., Kuropatkin, N., Lamb, D. Q., Lampeitl, H., Lee, B. C., Leger, R. F., Lin, H., Long, D. C., Loveday, J., Lupton, R. H., Margon, B., Martínez-Delgado, D., Mandelbaum, R., Matsubara, T., McGehee, P. M., McKay, T. A., Meiksin, A., Munn, J. A., Nakajima, R., Nash, T., Neilsen, Jr., E. H., Newberg, H. J., Newman, P. R., Nichol, R. C., Nicinski, T., Nieto-Santisteban, M., Nitta, A., O'Mullane, W., Okamura, S., Owen, R., Padmanabhan, N., Pauls, G., Peoples, J. J., Pier, J. R., Pope, A. C., Pourbaix, D., Quinn, T. R., Richards, G. T., Richmond, M. W., Rockosi, C. M., Schlegel, D. J., Schneider, D. P., Schroeder, J., Scranton, R., Seljak, U., Sheldon, E., Shimasaku, K., Smith, J. A., Smolčić, V., Snedden, S. A., Stoughton, C., Strauss, M. A., SubbaRao, M., Szalay, A. S., Szapudi, I., Szkody, P., Tegmark, M., Thakar, A. R., Tucker, D. L., Uomoto, A., Vanden Berk, D. E., Vandenberg, J., Vogeley, M. S., Voges, W., Vogt, N. P., Walkowicz, L. M., Weinberg, D. H., West, A. A., White, S. D. M., Xu, Y., Yanny, B., Yocum, D. R., York, D. G., Zehavi, I., Zibetti,

S. and Zucker, D. B.: 2006, The Fourth Data Release of the Sloan Digital Sky Survey, *ApJS* **162**, 38–48.

Adelman-McCarthy, J. K. f.: 2007, The Sixth Data Release of the Sloan Digital Sky Survey, *ArXiv e-prints* **707**.

AGASA Collaboration: N. Hayashida, Honda, K., Inoue, N., Kadota, K., Kakimoto, F., Kamata, K., Kawaguchi, S., Kawasaki, Y., Kawasumi, N., Kitamura, H., Kusano, E., Matsubara, Y., Murakami, K., Nagano, M., Nishikawa, D., Ohoka, H., Sakaki, N., Sasaki, M., Shinozaki, K., Souma, N., Takeda, M., Teshima, M., Torii, R., Tsushima, I., Uchihori, Y., Yamamoto, T., Yoshida, S. and Yoshii, H.: 1998, The Anisotropy of Cosmic Ray Arrival Directions around 10^{18} eV, *ArXiv Astrophysics e-prints* .

Alvarez-Muñiz, J., Engel, R., Gaisser, T. K., Ortiz, J. A. and Stanev, T.: 2004, Influence of shower fluctuations and primary composition on studies of the shower longitudinal development, *Phys. Rev. D* **69**(10), 103003–+.

Alvarez-Muñiz, J. and Halzen, F.: 2001, 10^{20} eV Cosmic Ray and Particle Physics with IceCube, *in* D. Saltzberg and P. Gorham (eds), *Radio Detection of High Energy Particles*, Vol. 579 of *American Institute of Physics Conference Series*, pp. 305–+.

Amendola, L., Frieman, J. A. and Waga, I.: 1999, Weak gravitational lensing by voids, *MNRAS* **309**, 465–473.

Anchordoqui, L. A., Dova, M. T., Epele, L. N. and Sciutto, S. J.: 1999, Hadronic interaction models beyond collider energies, *Phys. Rev. D* **59**(9), 094003–+.

Anchordoqui, L. A., Dova, M. T., Epele, L. N. and Swain, J. D.: 1997, Effect of the 3 K background radiation on ultrahigh energy cosmic rays, *Phys. Rev. D* **55**, 7356–7360.

Antoni, T., Apel, W. D., Badea, A. F., Bekk, K., Bercuci, A., Blümer, H., Bozdog, H., Brancus, I. M., Büttner, C., Daumiller, K., Doll, P., Engel, R., Engler, J., Fessler, F., Gils, H. J., Glasstetter, R., Haungs, A., Heck, D., Hörandel, J. R., Kampert, K.-H., Klages, H. O., Maier, G., Mathes, H. J., Mayer, H. J., Milke, J., Müller, M., Obenland, R., Oehlschläger, J., Ostapchenko, S., Petcu, M., Rebel, H., Risse, A., Risse, M., Roth, M., Schatz, G., Schieler, H., Scholz, J., Thouw, T., Ulrich, H., van Buren, J., Vardanyan, A., Weindl, A., Wochele, J. and Zabierowski, J.: 2004, Large-Scale Cosmic-Ray Anisotropy KASCADE, *ApJ* **604**, 687–692.

Antoni, T., Apel, W. D., Badea, A. F., Bekk, K., Bercuci, A., Blümer, J., Bozdog, H., Brancus, I. M., Chilingarian, A., Daumiller, K., Doll, P., Engel, R., Engler, J., Feßler, F., Gils, H. J., Glasstetter, R., Haungs, A., Heck, D., Hörandel, J. R., Kampert, K.-H., Klages, H. O., Maier, G., Mathes, H. J., Mayer, H. J., Milke, J., Müller, M., Obenland, R., Oehlschläger, J., Ostapchenko, S., Petcu, M., Rebel, H., Risse, A., Risse, M., Roth, M., Schatz, G., Schieler, H., Scholz, J., Thouw, T., Ulrich, H., van Buren, J., Vardanyan, A., Weindl, A., Wochele, J. and Zabierowski, J.: 2005, KASCADE measurements of energy spectra for elemental groups of cosmic rays: Results and open problems, *Astroparticle Physics* **24**, 1–2.

Antonucci, R.: 1993, Unified models for active galactic nuclei and quasars, *ARA&A* **31**, 473–521.

Ave, M., Hinton, J. A., Vázquez, R. A., Watson, A. A. and Zas, E.: 2000, New Constraints from Haverah Park Data on the Photon and Iron Fluxes of Ultrahigh-Energy Cosmic Rays, *Physical Review Letters* **85**, 2244–2247.

Balbinot, R., Bergamini, R. and Comastri, A.: 1988, Solution of the Einstein-Strauss problem with a Λ term, *Phys. Rev. D* **38**, 2415–2418.

Ball, N. M., Brunner, R. J. and Myers, A. D.: 2007, Robust Machine Learning Applied to Terascale Astronomical Datasets, *ArXiv e-prints* **710**.

Ball, N. M. and Loveday, J.: 2004, Galaxy Types and Luminosity Functions in the Sloan Digital Sky Survey using Artificial Neural Networks, *in* D. L. Block, I. Puerari, K. C. Freeman, R. Groess and E. K. Block (eds), *Penetrating Bars Through Masks of Cosmic Dust*, Vol. 319 of *Astrophysics and Space Science Library*, pp. 771–+.

Banerjee, A., Dhillon, I., Ghosh, J., Merugu, S. and Modha, A. D.: 2004, A generalized maximum entropy approach to bregman co-clustering and matrix approximation, *Technical Report UTCS TR04-24* .

Barnes, J.: 1985, The dynamical state of groups of galaxies, *MNRAS* **215**, 517–536.

Barnes, J. E.: 1989, Evolution of compact groups and the formation of elliptical galaxies, *Nature* **338**, 123–126.

Baum, W. A.: 1962, Photoelectric Magnitudes and Red-Shifts, *in* G. C. McVittie (ed.), *Problems of Extra-Galactic Research*, Vol. 15 of *IAU Symposium*, pp. 390–+.

Bazell, D. and Miller, D. J.: 2005, Class Discovery in Galaxy Classification, *ApJ* **618**, 723–732.

Beck, R., Brandenburg, A., Moss, D., Shukurov, A. and Sokoloff, D.: 1996, Galactic Magnetism: Recent Developments and Perspectives, *ARA&A* **34**, 155–206.

Benítez, N.: 2000, Bayesian Photometric Redshift Estimation, *ApJ* **536**, 571–583.

Benson, A. J., Hoyle, F., Torres, F. and Vogeley, M. S.: 2003, Galaxy voids in cold dark matter universes, *MNRAS* **340**, 160–174.

Berezinskii, V. S. and Grigor'eva, S. I.: 1988, A bump in the ultra-high energy cosmic ray spectrum, *A&A* **199**, 1–2.

Bernardi, M., Sheth, R. K., Annis, J., Burles, S., Finkbeiner, D. P., Lupton, R. H., Schlegel, D. J., SubbaRao, M., Bahcall, N. A., Blakeslee, J. P., Brinkmann, J., Castander, F. J., Connolly, A. J., Csabai, I., Doi, M., Fukugita, M., Frieman, J., Heckman, T., Hennessy, G. S., Ivezić, Ž., Knapp, G. R., Lamb, D. Q., McKay, T., Munn, J. A., Nichol, R., Okamura, S., Schneider, D. P., Thakar, A. R. and York, D. G.: 2003, Early-Type Galaxies in the Sloan Digital Sky Survey. IV. Colors and Chemical Evolution, *AJ* **125**, 1882–1896.

Bernstein, G. and Jain, B.: 2004, Dark Energy Constraints from Weak-Lensing Cross-Correlation Cosmography, *ApJ* **600**, 17–25.

Best, P. N., Kauffmann, G., Heckman, T. M., Brinchmann, J., Charlot, S., Ivezić, Ž. and White, S. D. M.: 2005, The host galaxies of radio-loud active galactic nuclei: mass dependences, gas cooling and active galactic nuclei feedback, *MNRAS* **362**, 25–40.

Best, P. N., Kauffmann, G., Heckman, T. M. and Ivezić, Ž.: 2005, A sample of radio-loud active galactic nuclei in the Sloan Digital Sky Survey, *MNRAS* **362**, 9–24.

Bird, D. J., Corbato, S. C., Dai, H. Y., Dawson, B. R., Elbert, J. W., Emerson, B. L., Green, K. D., Huang, M. A., Kieda, D. B., Luo, M., Ko, S., Larsen, C. G., Loh, E. C., Salamon, M. H., Smith, J. D., Sokolsky, P., Sommers, P., Tang, J. K. K. and Thomas, S. B.: 1994, The cosmic-ray energy spectrum observed by the Fly's Eye, *ApJ* **424**, 491–502.

Bird, D. J., Corbató, S. C., Dai, H. Y., Dawson, B. R., Elbert, J. W., Gaisser, T. K., Green, K. D., Huang, M. A., Kieda, D. B., Ko, S., Larsen, C. G., Loh, E. C., Luo, M., Salamon, M. H., Smith, D., Sokolsky, P., Sommers, P., Stanev, T., Tang, J. K., Thomas, S. B. and Tilav, S.: 1993, Evidence for correlated changes in the spectrum and composition of cosmic rays at extremely high energies, *Physical Review Letters* **71**, 3401–3404.

Bird, D. J., Corbato, S. C., Dai, H. Y., Elbert, J. W., Green, K. D., Huang, M. A., Kieda, D. B., Ko, S., Larsen, C. G., Loh, E. C., Luo, M. Z., Salamon, M. H., Smith, J. D., Sokolsky, P., Sommers, P., Tang, J. K. K. and Thomas, S. B.: 1995, Detection of a cosmic ray with measured energy well beyond the expected spectral cutoff due to cosmic microwave radiation, *ApJ* **441**, 144–150.

Bird, D. J., Dai, H. Y., Dawson, B. R., Elbert, J. W., Huang, M. A., Kieda, D. B., Ko, S., Loh, E. C., Luo, M., Smith, J. D., Sokolsky, P., Sommers, P. and Thomas, S. B.: 1999, Study of Broad-Scale Anisotropy of Cosmic-Ray Arrival Directions from 2×10^{17} to 10^{20} Electron Volts from Fly's Eye Data, *ApJ* **511**, 739–749.

Bishop, C.: 1995, *Neural Networks for Pattern Recognition*, Oxford University Press, Inc., New York, NY, USA.

Bishop, C. M.: 1999, *Learning in Graphical Models*.

Bishop, C. M. e. a.: 1998, Developments of the Generative Topographic Mapping, *Neurocomputing* **21**, 203–224.

Blake, C. and Bridle, S.: 2005, Cosmology with photometric redshift surveys, *MNRAS* **363**, 1329–1348.

Blanton, M., Blasi, P. and Olinto, A. V.: 2001, The Greisen-Zatzepin-Kuzmin feature in our neighborhood of the universe, *Astroparticle Physics* **15**, 275–286.

Blanton, M. R., Hogg, D. W., Bahcall, N. A., Brinkmann, J., Britton, M., Connolly, A. J., Csabai, I., Fukugita, M., Loveday, J., Meiksin, A., Munn, J. A., Nichol, R. C., Okamura, S., Quinn, T., Schneider, D. P., Shimasaku, K., Strauss, M. A., Tegmark, M., Vogeley, M. S. and Weinberg,

D. H.: 2003, The Galaxy Luminosity Function and Luminosity Density at Redshift z = 0.1, *ApJ* **592**, 819–838.

Bode, P. W., Cohn, H. N. and Lugger, P. M.: 1993, Simulations of Compact Groups of Galaxies: The Effect of the Dark Matter Distribution, *ApJ* **416**, 17–+.

Bolzonella, M., Miralles, J.-M. and Pelló, R.: 2000, Photometric redshifts based on standard SED fitting procedures, *A&A* **363**, 476–492.

Bolzonella, M., Pelló, R. and Maccagni, D.: 2002, Luminosity functions beyond the spectroscopic limit. I. Method and near-infrared LFs in the HDF-N and HDF-S, *A&A* **395**, 443–463.

Brodwin, M., Brown, M. J. I., Ashby, M. L. N., Bian, C., Brand, K., Dey, A., Eisenhardt, P. R., Eisenstein, D. J., Gonzalez, A. H., Huang, J.-S., Jannuzi, B. T., Kochanek, C. S., McKenzie, E., Murray, S. S., Pahre, M. A., Smith, H. A., Soifer, B. T., Stanford, S. A., Stern, D. and Elston, R. J.: 2006, Photometric Redshifts in the IRAC Shallow Survey, *ApJ* **651**, 791–803.

Brunner, R. J., Connolly, A. J. and Szalay, A. S.: 1999, The Statistical Approach to Quantifying Galaxy Evolution, *ApJ* **516**, 563–581.

Brunner, R. J., Connolly, A. J., Szalay, A. S. and Bershady, M. A.: 1997, Toward More Precise Photometric Redshifts: Calibration Via CCD Photometry, *ApJL* **482**, L21+.

Bruzual A., G. and Charlot, S.: 1993, Spectral evolution of stellar populations using isochrone synthesis, *ApJ* **405**, 538–553.

Budavári, T., Connolly, A. J., Szalay, A. S., Szapudi, I., Csabai, I., Scranton, R., Bahcall, N. A., Brinkmann, J., Eisenstein, D. J., Frieman, J. A., Fukugita, M., Gunn, J. E., Johnston, D., Kent, S., Loveday, J. N., Lupton, R. H., Tegmark, M., Thakar, A. R., Yanny, B., York, D. G. and Zehavi, I.: 2003, Angular Clustering with Photometric Redshifts in the Sloan Digital Sky Survey: Bimodality in the Clustering Properties of Galaxies, *ApJ* **595**, 59–70.

Budavári, T., Szalay, A. S., Charlot, S., Seibert, M., Wyder, T. K., Arnouts, S., Barlow, T. A., Bianchi, L., Byun, Y.-I., Donas, J., Forster, K., Friedman, P. G., Heckman, T. M., Jelinsky, P. N., Lee, Y.-W., Madore, B. F., Malina, R. F., Martin, D. C., Milliard, B., Morrissey, P., Neff, S. G., Rich, R. M., Schiminovich, D., Siegmund, O. H. W., Small, T., Treyer, M. A. and Welsh, B.: 2005, The Ultraviolet Luminosity Function of GALEX Galaxies at Photometric Redshifts between 0.07 and 0.25, *ApJL* **619**, L31–L34.

Butchins, S. A.: 1981, Predicted redshifts of galaxies by broadband photometry, *A&A* **97**, 407–409.

Capaccioli, M., Mancini, D. and Sedmak, G.: 2003, VST: The VLT Survey Telescope, *Memorie della Societa Astronomica Italiana* **74**, 450–+.

Capozziello, S., Funaro, M. and Stornaiolo, C.: 2004, Cosmological black holes as seeds of voids in the galaxy distribution, *A&A* **420**, 847–851.

Chang, K. Y. and Ghosh, J.: 2000, Three-Dimensional Model-Based Object Recognition and Pose Estimation Using Probabilistic Principal Surfaces, *SPIE: Applications of Artificial Neural Networks in Image*, pp. 192–203.

Chang, K. Y. and Ghosh, J.: 2001, A unified Model for Probabilistic Principal Surfaces, *IEE Transactions on Pattern Analysis and Machine intelligence*, Vol. 23, pp. 22–41.

Chiaberge, M., Capetti, A. and Celotti, A.: 2002, Understanding the nature of FR II optical nuclei: A new diagnostic plane for radio galaxies, *A&A* **394**, 791–800.

Chodorowski, M. J., Zdziarski, A. A. and Sikora, M.: 1992, Reaction rate and energy-loss rate for photopair production by relativistic nuclei, *ApJ* **400**, 181–185.

Choudhury, T. R. and Padmanabhan, T.: 2005, Cosmological parameters from supernova observations: A critical comparison of three data sets, *A&A* **429**, 807–818.

Ciaramella, A., Longo, G., Staiano, A. and Tagliaferri, R.: 2005, NEC: A Hierarchical Agglomerative Clustering Based on Fisher and Negentropy Information, *WIRN/NAIS*, pp. 49–56.

Colberg, J. M.: 2007, Quantifying cosmic superstructures, *MNRAS* **375**, 337–347.

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

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

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

Conroy, C., Coil, A. L., White, M., Newman, J. A., Yan, R., Cooper, M. C., Gerke, B. F., Davis, M. and Koo, D. C.: 2005, The DEEP2 Galaxy Redshift Survey: The Evolution of Void Statistics from z ˜ 1 to z ˜ 0, *ApJ* **635**, 990–1005.

Cooray, A., Hu, W., Huterer, D. and Joffre, M.: 2001, Measuring Angular Diameter Distances through Halo Clustering, *ApJL* **557**, L7–L10.

Cronin, J. W.: 1992, Summary of the workshop, *Nuclear Physics B Proceedings Supplements* **28**, 213–225.

Croom, S. M., Warren, S. J. and Glazebrook, K.: 2001, A small-area faint KX redshift survey for QSOs in the ESO Imaging Survey Chandra Deep Field South, *MNRAS* **328**, 150–158.

Croton, D. J., Colless, M., Gaztañaga, E., Baugh, C. M., Norberg, P., Baldry, I. K., Bland-Hawthorn, J., Bridges, T., Cannon, R., Cole, S., Collins, C., Couch, W., Dalton, G., de Propris, R., Driver, S. P., Efstathiou, G., Ellis, R. S., Frenk, C. S., Glazebrook, K., Jackson, C., Lahav, O., Lewis, I., Lumsden, S., Maddox, S., Madgwick, D., Peacock, J. A., Peterson, B. A., Sutherland, W. and Taylor, K.: 2004, The 2dF Galaxy Redshift Survey: voids and hierarchical scaling models, *MNRAS* **352**, 828–836.

Csabai, I., Budavári, T., Connolly, A. J., Szalay, A. S., Győry, Z., Benítez, N., Annis, J., Brinkmann, J., Eisenstein, D., Fukugita, M., Gunn, J., Kent, S., Lupton, R., Nichol, R. C. and Stoughton, C.: 2003, The Application of Photometric Redshifts to the SDSS Early Data Release, *AJ* **125**, 580–592.

D'Abrusco, R., Staiano, A., Longo, G., Brescia, M., Paolillo, M., De Filippis, E. and Tagliaferri, R.: 2007, Mining the SDSS Archive. I. Photometric Redshifts in the Nearby Universe, *ApJ* **663**, 752–764.

Davis, M., Faber, S. M., Newman, J., Phillips, A. C., Ellis, R. S., Steidel, C. C., Conselice, C., Coil, A. L., Finkbeiner, D. P., Koo, D. C., Guhathakurta, P., Weiner, B., Schiavon, R., Willmer, C., Kaiser, N., Luppino, G. A., Wirth, G., Connolly, A., Eisenhardt, P., Cooper, M. and Gerke, B.: 2003, Science Objectives and Early Results of the DEEP2 Redshift Survey, *in* P. Guhathakurta (ed.), *Discoveries and Research Prospects from 6- to 10-Meter-Class Telescopes II. Edited by Guhathakurta, Puragra. Proceedings of the SPIE, Volume 4834, pp. 161-172 (2003).*, Vol. 4834 of *Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference*, pp. 161–172.

Davis, M., Meiksin, A., Strauss, M. A., da Costa, L. N. and Yahil, A.: 1988, On the universality of the two-point galaxy correlation function, *ApJL* **333**, L9–L12.

Davis, M., Newman, J. A., Faber, S. M. and Phillips, A. C.: 2001, The DEEP2 Redshift Survey, *in* S. Cristiani, A. Renzini and R. E. Williams (eds), *Deep Fields*, pp. 241–+.

de Lapparent, V., Geller, M. J. and Huchra, J. P.: 1986, A slice of the universe, *ApJL* **302**, L1–L5.

Dempster, A. P., Laird, N. M. and Rubin, D. B.: 1977, Maximum-Likelihood from Incomplete Data Via the EM Algorithm, *J. Royal Statistical Soc.* **39**(1).

Diaferio, A., Geller, M. J. and Ramella, M.: 1994, The formation of compact groups of galaxies. I: Optical properties, *AJ* **107**, 868–879.

Diaferio, A., Ramella, M., Geller, M. J. and Ferrari, A.: 1993, Are groups of galaxies virialized systems?, *AJ* **105**, 2035–2046.

Dietterich, T. G.: 2002, *Ensemble Learning*.

Djorgovski, G. S.: 1992, Galaxy Manifolds and Galaxy Formation, *in* G. Longo, M. Capaccioli and G. Busarello (eds), *Morphological and Physical Classification of Galaxies*, Vol. 178 of *Astrophysics and Space Science Library*, pp. 337–+.

Djorgovski, S. G., Baltay, C., Mahabal, A., Graham, M., Williams, R., Bogosavljevic, M., Rabinowitz, D., Bauer, A., Ellman, N., Lauer, R., Duffau, S., Andrews, P., Rengstorf, A., Brunner, R., Musser, J., Gebhard, M., Mufson, S. and PQ: 2004, The Palomar-Quest Survey, *Bulletin of the American Astronomical Society*, Vol. 36 of *Bulletin of the American Astronomical Society*, pp. 1487–+.

Dolag, K., Grasso, D., Springel, V. and Tkachev, I.: 2004, Mapping Deflections of Ultrahigh Energy Cosmic Rays in Constrained Simulations of Extraglactic Magnetic Fields, *Soviet Journal of Experimental and Theoretical Physics Letters* **79**, 583–587.

Dova, M. T., Mancenido, M. E., Mariazzi, A. G., McCauley, T. P. and Watson, A. A.: 2003, New constraints on the mass composition of cosmic rays above 10^{17} eV from Volcano Ranch measurements, *ArXiv Astrophysics e-prints* .

Dressler, A.: 1980, Galaxy morphology in rich clusters - Implications for the formation and evolution of galaxies, *ApJ* **236**, 351–365.

Dye, S., Warren, S. J., Hambly, N. C., Cross, N. J. G., Hodgkin, S. T., Irwin, M. J., Lawrence, A., Adamson, A. J., Almaini, O., Edge, A. C., Hirst, P., Jameson, R. F., Lucas, P. W., van Breukelen, C., Bryant, J., Casali, M., Collins, R. S., Dalton, G. B., Davies, J. I., Davis, C. J., Emerson, J. P., Evans, D. W., Foucaud, S., Gonzales-Solares, E. A., Hewett, P. C., Kendall, T. R., Kerr, T. H., Leggett, S. K., Lodieu, N., Loveday, J., Lewis, J. R., Mann, R. G., McMahon, R. G., Mortlock, D. J., Nakajima, Y., Pinfield, D. J., Rawlings, M. G., Read, M. A., Riello, M., Sekiguchi, K., Smith, A. J., Sutorius, E. T. W., Varricatt, W., Walton, N. A. and Weatherley, S. J.: 2006, The UKIRT Infrared Deep Sky Survey Early Data Release, *MNRAS* **372**, 1227–1252.

Edge, D. M., Pollock, A. M. T., Reid, R. J. O., Watson, A. A. and Wilson, J. G.: 1978, A study of the arrival direction distribution of high-energy particles as observed from the Northern Hemisphere, *Journal of Physics G Nuclear Physics* **4**, 133–157.

Edmondson, E. M., Miller, L. and Wolf, C.: 2006, Bayesian photometric redshifts for weak-lensing applications, *MNRAS* **371**, 1693–1704.

Efimov, N. N. and et al.: 1991, The Energy Spectrum and Anisotropy of Primary Cosmic Rays at Energy $E_0 > 10^{17}$ eV Observed in Yakutsk, *in* M. Nagano and F. Takahara (eds), *Astrophysical Aspects of the Most Energetic Cosmic Rays*, pp. 20–+.

Efstathiou, G., Ellis, R. S. and Peterson, B. A.: 1988, Analysis of a complete galaxy redshift survey. II - The field-galaxy luminosity function, *MNRAS* **232**, 431–461.

Einasto, J. e. a.: 2007a, Superclusters of galaxies from the 2dF redshift survey. I. The catalogue, *A&A* **462**, 811–825.

Einasto, J. e. a.: 2007b, Superclusters of galaxies from the 2dF redshift survey. II. Comparison with simulations, *A&A* **462**, 397–410.

Einstein, A. and Straus, E. G.: 1946, Corrections and Additional Remarks to our Paper: The Influence of the Expansion of Space on the Gravitation Fields Surrounding the Individual Stars, *Reviews of Modern Physics* **18**, 148–149.

Eisenstein, D. J., Annis, J., Gunn, J. E., Szalay, A. S., Connolly, A. J., Nichol, R. C., Bahcall, N. A., Bernardi, M., Burles, S., Castander, F. J., Fukugita, M., Hogg, D. W., Ivezić, Ž., Knapp, G. R., Lupton, R. H., Narayanan, V., Postman, M., Reichart, D. E., Richmond, M., Schneider, D. P., Schlegel, D. J., Strauss, M. A., SubbaRao, M., Tucker, D. L., Vanden Berk, D., Vogeley, M. S., Weinberg, D. H. and Yanny, B.: 2001, Spectroscopic Target Selection for the Sloan Digital Sky Survey: The Luminous Red Galaxy Sample, *AJ* **122**, 2267–2280.

Eisenstein, D. J., Hogg, D. W., Fukugita, M., Nakamura, O., Bernardi, M., Finkbeiner, D. P., Schlegel, D. J., Brinkmann, J., Connolly, A. J., Csabai, I., Gunn, J. E., Ivezić, Ž., Lamb, D. Q., Loveday, J., Munn, J. A., Nichol, R. C., Schneider, D. P., Strauss, M. A., Szalay, A. and York, D. G.: 2003, Average Spectra of Massive Galaxies in the Sloan Digital Sky Survey, *ApJ* **585**, 694–713.

El-Ad, H. and Piran, T.: 1997, Voids in the Large-Scale Structure, *ApJ* **491**, 421–+.

El-Ad, H. and Piran, T.: 2000, A case devoid of bias: Optical Redshift Survey voids versus IRAS voids, *MNRAS* **313**, 553–558.

Evans, N. W., Ferrer, F. and Sarkar, S.: 2002, The anisotropy of the ultra-high energy cosmic rays, *Astroparticle Physics* **17**, 319–340.

Evans, N. W., Ferrer, F. and Sarkar, S.: 2003, Clustering of ultrahigh energy cosmic rays and their sources, *Phys. Rev. D* **67**(10), 103005–+.

Faber, S. M., Courteau, S., Dekel, A., Dressler, A., Kollatt, T., Willick, J. A. and Yahil, A.: 1994, Cosmic Velocity Flows, *JRASC* **88**, 92–+.

Fan, X., Strauss, M. A., Schneider, D. P., Gunn, J. E., Lupton, R. H., Becker, R. H., Davis, M., Newman, J. A., Richards, G. T., White, R. L., Anderson, Jr., J. E., Annis, J., Bahcall, N. A., Brunner, R. J., Csabai, I., Hennessy, G. S., Hindsley, R. B., Fukugita, M., Kunszt, P. Z., Ivezić, Ž., Knapp, G. R., McKay, T. A., Munn, J. A., Pier, J. R., Szalay, A. S. and York, D. G.: 2001, High-Redshift Quasars Found in Sloan Digital Sky Survey Commissioning Data. IV. Luminosity Function from the Fall Equatorial Stripe Sample, *AJ* **121**, 54–65.

Farrar, G. R. and Biermann, P. L.: 1998, Correlation between Compact Radio Quasars and Ultra-high Energy Cosmic Rays, *Physical Review Letters* **81**, 3579–3582.

Fasano, G. and Bettoni, D.: 1994, Morphology of early-type galaxies in compact groups, 2., *AJ* **107**, 1649–1667.

Fernández-Soto, A., Lanzetta, K. M., Chen, H.-W., Levine, B. and Yahata, N.: 2002, Error analysis of the photometric redshift technique, *MNRAS* **330**, 889–894.

Fernández-Soto, A., Lanzetta, K. M., Chen, H.-W., Pascarelle, S. M. and Yahata, N.: 2001, On the Compared Accuracy and Reliability of Spectroscopic and Photometric Redshift Measurements, *ApJS* **135**, 41–61.

Fernández-Soto, A., Lanzetta, K. M. and Yahil, A.: 1999, A New Catalog of Photometric Redshifts in the Hubble Deep Field, *ApJ* **513**, 34–50.

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

Fisher, K. B., Huchra, J. P., Strauss, M. A., Davis, M., Yahil, A. and Schlegel, D.: 1995, The IRAS 1.2 Jy Survey: Redshift Data, *ApJS* **100**, 69–+.

Freund, Y, S. R. E.: 1996, Experiments with a new boosting algorithm, *Proceedings 13th International Conference on Machine Learning*.

Friedmann, Y. and Piran, T.: 2001, A Model of Void Formation, *ApJ* **548**, 1–6.

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

Furlanetto, S. R. and Piran, T.: 2006, The evidence of absence: galaxy voids in the excursion set formalism, *MNRAS* **366**, 467–479.

Gladders, M. D. and Yee, H. K. C.: 2000, A New Method For Galaxy Cluster Detection. I. The Algorithm, *AJ* **120**, 2148–2162.

Goldberg, D. M., Jones, T. D., Hoyle, F., Rojas, R. R., Vogeley, M. S. and Blanton, M. R.: 2005, The Mass Function of Void Galaxies in the Sloan Digital Sky Survey Data Release 2, *ApJ* **621**, 643–650.

Goto, T., Yamauchi, C., Fujita, Y., Okamura, S., Sekiguchi, M., Smail, I., Bernardi, M. and Gomez, P. L.: 2003, The morphology-density relation in the Sloan Digital Sky Survey, *MNRAS* **346**, 601–614.

Governato, F., Tozzi, P. and Cavaliere, A.: 1996, Small Groups of Galaxies: A Clue to a Critical Universe, *ApJ* **458**, 18–+.

Grandi, S. A. and Osterbrock, D. E.: 1978, Optical spectra of radio galaxies, *ApJ* **220**, 783–789.

Gregory, S. A., Thompson, L. A. and Tifft, W. G.: 1978, The Perseus/Pisces Supercluster, *Bulletin of the American Astronomical Society*, Vol. 10 of *Bulletin of the American Astronomical Society*, pp. 622–+.

Greisen, K.: 1966, End to the Cosmic-Ray Spectrum?, *Physical Review Letters* **16**, 748–750.

Gunn, J. E. and Gott, J. R. I.: 1972, On the Infall of Matter Into Clusters of Galaxies and Some Effects on Their Evolution, *ApJ* **176**, 1–+.

Hahn, O., Carollo, C. M., Porciani, C. and Dekel, A.: 2007, The evolution of dark matter halo properties in clusters, filaments, sheets and voids, *MNRAS* **381**, 41–51.

Hall, P. B., Osmer, P. S., Green, R. F., Porter, A. C. and Warren, S. J.: 1996, A Deep Multicolor Survey. II. Initial Spectroscopy and Comparison with Expected Quasar Number Counts, *ApJ* **462**, 614–+.

Hamilton, D.: 1985, The spectral evolution of galaxies. I - an observational approach, *ApJ* **297**, 371–389.

Hayashida, N., Honda, K., Honda, M., Inoue, N., Kadota, K., Kakimoto, F., Kamata, K., Kawaguchi, S., Kawasumi, N., Matsubara, Y., Murakami, K., Nagano, M., Ohoka, H., Sakaki, N., Souma, N., Takeda, M., Teshima, M., Tsushima, I., Uchihori, Y., Yoshida, S. and Yoshii, H.: 1996, Possible Clustering of the Most Energetic Cosmic Rays within a Limited Space Angle Observed by the Akeno Giant Air Shower Array, *Physical Review Letters* **77**, 1000–1003.

Hayashida, N., Honda, K., Inoue, N., Kadota, K., Kakimoto, F., Kakizawa, S., Kamata, K., Kawaguchi, S., Kawasaki, Y., Kawasumi, N., Kusano, E., Mahrous, A. M., Mase, K., Minagawa, T., Nagano, M., Nishikawa, D., Ohoka, H., Osone, S., Sakaki, N., Sasaki, M., Shinozaki, K., Takeda, M., Teshima, M., Torii, R., Tsushima, I., Uchihori, Y., Yamamoto, T., Yoshida, S. and Yoshii, H.: 2000, Updated AGASA event list above 4×10^{19} eV, *ArXiv Astrophysics e-prints* .

Haykin, S.: 1999, *Neural networks: a comprehensive foundation*, second edn, Prentice Hall.

Hewett, P. C., Foltz, C. B. and Chaffee, F. H.: 1995, The large bright quasar survey. 6: Quasar catalog and survey parameters, *AJ* **109**, 1498–1521.

Hickson, P.: 1997, Compact Groups of Galaxies, *ARA&A* **35**, 357–388.

Hickson, P., Mendes de Oliveira, C., Huchra, J. P. and Palumbo, G. G.: 1992, Dynamical properties of compact groups of galaxies, *ApJ* **399**, 353–367.

High Resolution Fly'S Eye Collaboration, Abbasi, R. U., Abu-Zayyad, T., Amann, J. F., Archbold, G., Atkins, R., Bellido, J. A., Belov, K., Belz, J. W., Benzvi, S., Bergman, D. R., Burt, G. W., Cao, Z., Clay, R. W., Connolly, B., Dawson, B. R., Deng, W., Fedorova, Y., Findlay, J., Finley, C. B., Hanlon, W. F., Hoffman, C. M., Holzscheiter, M. H., Hughes, G. A., Hüntemeyer, P., Jui, C. C. H., Kim, K., Kirn, M. A., Loh, E. C., Maestas, M. M., Manago, N., Marek, L. J., Martens, K., Matthews, J. A. J., Matthews, J. N., O'Neill, A., Painter, C. A., Perera, L., Reil, K., Riehle, R., Roberts, M., Sasaki, M., Schnetzer, S. R., Simpson, K. M., Sinnis, G., Smith, J. D., Snow, R., Sokolsky, P., Song, C., Springer, R. W., Stokes, B. T., Thomas, J. R., Thomas, S. B., Thomson, G. B., Tupa, D., Westerhoff, S., Wiencke, L. R. and Zech, A.: 2004, A search for arrival direction clustering in the HiRes-I monocular data above $10^{19.5}$ eV, *Astroparticle Physics* **22**, 139–149.

Hine, R. G. and Longair, M. S.: 1979, Optical spectra of 3CR radio galaxies, *MNRAS* **188**, 111–130.

Hoyle, F. and Vogeley, M. S.: 2004, Voids in the Two-Degree Field Galaxy Redshift Survey, *ApJ* **607**, 751–764.

Hu, W. and Sugiyama, N.: 1996, Small-Scale Cosmological Perturbations: an Analytic Approach, *ApJ* **471**, 542–+.

Hubble, E. P.: 1926, Extragalactic nebulae., *ApJ* **64**, 321–369.

Hubble, E. P.: 1958, *The realm of the nebulae*, New York: Dover, 1958.

Huchra, J. P., Geller, M. J., de Lapparent, V. and Burg, R.: 1988, The CFA Redshift Survey, *in* J. Audouze, M.-C. Pelletan and S. Szalay (eds), *Large Scale Structures of the Universe*, Vol. 130 of *IAU Symposium*, pp. 105–+.

Huterer, D., Kim, A., Krauss, L. M. and Broderick, T.: 2004, Redshift Accuracy Requirements for Future Supernova and Number Count Surveys, *ApJ* **615**, 595–602.

Huterer, D., Takada, M., Bernstein, G. and Jain, B.: 2006, Systematic errors in future weak-lensing surveys: requirements and prospects for self-calibration, *MNRAS* **366**, 101–114.

Iovino, A. and Hickson, P.: 1997, Discordant redshifts in compact groups, *MNRAS* **287**, 21–25.

Ishak, M.: 2005, Probing decisive answers to dark energy questions from cosmic complementarity and lensing tomography, *MNRAS* **363**, 469–478.

Jarrett, T.: 2004, Large Scale Structure in the Local Universe - The 2MASS Galaxy Catalog, *Publications of the Astronomical Society of Australia* **21**, 396–403.

Jarrett, T. H., Chester, T., Cutri, R., Schneider, S., Skrutskie, M. and Huchra, J. P.: 2000, 2MASS Extended Source Catalog: Overview and Algorithms, *AJ* **119**, 2498–2531.

Jones, C., Hickox, R., Murray, S., Forman, W., Brodwin, M., XBootes, Shallow Survey, I., NDWFS and AGES Teams: 2006, A Large Population of Infrared-Selected, Obscured AGN in the Bootes Field, *Bulletin of the American Astronomical Society*, Vol. 38 of *Bulletin of the American Astronomical Society*, pp. 1124–+.

Kachelrieß, M., Serpico, P. D. and Teshima, M.: 2007, The Galactic magnetic field as spectrograph for ultra-high energy cosmic rays, *Astroparticle Physics* **26**, 378–386.

Kaiser, N., Tonry, J. L. and Luppino, G. A.: 2000, A New Strategy for Deep Wide-Field High-Resolution Optical Imaging, *PASP* **112**, 768–800.

Kantowski, R.: 1969, Corrections in the Luminosity-Redshift Relations of the Homogeneous Fried-Mann Models, *ApJ* **155**, 89–+.

Kantowski, R. and Thomas, R. C.: 2001, Distance-Redshift in Inhomogeneous $\Omega_0 = 1$ Friedmann-Lemaître-Robertson-Walker Cosmology, *ApJ* **561**, 491–495.

Kasahara, K.: 2005, The Current Status and Prospect of the Ta Experiment, *ArXiv Astrophysics e-prints* .

Kerscher, M., Szapudi, I. and Szalay, A. S.: 2000, A Comparison of Estimators for the Two-Point Correlation Function, *ApJL* **535**, L13–L16.

Kirshner, R. P., Oemler, Jr., A., Schechter, P. L. and Shectman, S. A.: 1981, A million cubic megaparsec void in Bootes, *ApJL* **248**, L57–L60.

Kleinmann, S. G., Young, J. S., Claussen, M. J., Rubin, V. C. and Scoville, N. Z.: 1987, Optical and Millimeter-Wave Studies of NGC 2148, *Bulletin of the American Astronomical Society*, Vol. 19 of *Bulletin of the American Astronomical Society*, pp. 681–+.

Knapp, J., Heck, D., Sciutto, S. J., Dova, M. T. and Risse, M.: 2003, Extensive air shower simulations at the highest energies, *Astroparticle Physics* **19**, 77–99.

Kodaira, K., Doi, M., Ichikawa, S.-I. and Okamura, S.: 1990, An observational study of Shakhbazyan's compact groups of galaxies. II - SCGG 202, 205, 223, 245, and 348, *Publications of the National Astronomical Observatory of Japan* **1**, 283–295.

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

Koo, D. C. and Kron, R. G.: 1982, QSO counts - A complete survey of stellar objects to B = 23, *A&A* **105**, 107–119.

Laing, R. A., Jenkins, C. R., Wall, J. V. and Unger, S. W.: 1994, Spectrophotometry of a Complete Sample of 3CR Radio Sources: Implications for Unified Models, *in* G. V. Bicknell, M. A. Dopita and P. J. Quinn (eds), *The Physics of Active Galaxies*, Vol. 54 of *Astronomical Society of the Pacific Conference Series*, pp. 201–+.

Landy, S. D. and Szalay, A. S.: 1993, Bias and variance of angular correlation functions, *ApJ* **412**, 64–71.

Lanzetta, K. M., Yahil, A. and Fernández-Soto, A.: 1996, Star-forming galaxies at very high redshifts, *Nature* **381**, 759–763.

Lawrence, M. A., Reid, R. J. O. and Watson, A. A.: 1991, The cosmic ray energy spectrum above 4×10^{17} eV as measured by the Haverah Park array, *Journal of Physics G Nuclear Physics* **17**, 733–757.

Lee, J. and Shandarin, S. F.: 1998, Large-Scale Biasing and the Primordial Gravitational Potential, *ApJL* **505**, L75–L78.

Li, L., Zhang, Y., Zhao, Y. and Yang, D.: 2006, Multi-parameter estimating photometric redshifts with artificial neural networks, *ArXiv Astrophysics e-prints* .

Linsley, J.: 1963, Evidence for a Primary Cosmic-Ray Particle with Energy 10^{20} eV, *Physical Review Letters* **10**, 146–148.

Linsley, J. and Watson, A. A.: 1981, Validity of scaling to 10^{20} eV and high-energy cosmic-ray composition, *Physical Review Letters* **46**, 459–463.

Lundmark, K.: 1925, Nebulæ, The motions and the distances of spiral, *MNRAS* **85**, 865–+.

Lynds, R.: 1971, The Absorption-Line Spectrum of 4c 05.34, *ApJL* **164**, L73+.

Ma, Z., Hu, W. and Huterer, D.: 2006, Effects of Photometric Redshift Uncertainties on Weak-Lensing Tomography, *ApJ* **636**, 21–29.

Madsen, S., Doroshkevich, A. G., Gottlober, S. and Müller, V.: 1998, The cross correlation between the gravitational potential and the large scale matter distribution, *A&A* **329**, 1–13.

Mamon, G.: 1996, The Dynamics of Groups and Clusters of Galaxies and Links to Cosmology, *in* H. J. de Vega and N. Sánchez (eds), *Third Paris Cosmology Colloquium*, pp. 95–+.

Mamon, G. A.: 1986, Are compact groups of galaxies physically dense?, *ApJ* **307**, 426–430.

Mamon, G. A.: 1999, Understanding low and high velocity dispersion compact groups, *ArXiv Astrophysics e-prints* .

Mamon, G. A.: 2006, The evolution of galaxy groups and of galaxies therein, *ArXiv Astrophysics e-prints* .

Massarotti, M., Iovino, A. and Buzzoni, A.: 2001, A critical appraisal of the SED fitting method to estimate photometric redshifts, *A&A* **368**, 74–85.

Massarotti, M., Iovino, A., Buzzoni, A. and Valls-Gabaud, D.: 2001, New insights on the accuracy of photometric redshift measurements, *A&A* **380**, 425–434.

McPherson, A. M., Born, A., Sutherland, W., Emerson, J., Little, B., Jeffers, P., Stewart, M., Murray, J. and Ward, K.: 2006, VISTA: project status, *Ground-based and Airborne Telescopes. Edited by Stepp, Larry M.. Proceedings of the SPIE, Volume 6267, pp. 626707 (2006).*, Vol. 6267 of *Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference*.

Montoya, M. L., Dominguez-Tenreiro, R., Gonzalez-Casado, G., Mamon, G. A. and Salvador-Sole, E.: 1996, The Surface Density Profiles and Lensing Characteristics of Hickson Compact Groups of Galaxies, *ApJL* **473**, L83+.

Müller, V., Arbabi-Bidgoli, S., Einasto, J. and Tucker, D.: 2000, Voids in the Las Campanas Redshift Survey versus cold dark matter models, *MNRAS* **318**, 280–288.

Oleak, H., Stoll, D., Tiersch, H. and MacGillivray, H. T.: 1995, On the ellipticity of the Shakhbazian compact groups of galaxies, *AJ* **109**, 1485–1489.

Padmanabhan, N., Budavári, T., Schlegel, D. J., Bridges, T., Brinkmann, J., Cannon, R., Connolly, A. J., Croom, S. M., Csabai, I., Drinkwater, M., Eisenstein, D. J., Hewett, P. C., Loveday, J., Nichol, R. C., Pimbblet, K. A., De Propris, R., Schneider, D. P., Scranton, R., Seljak, U., Shanks, T., Szapudi, I., Szalay, A. S. and Wake, D.: 2005, Calibrating photometric redshifts of luminous red galaxies, *MNRAS* **359**, 237–250.

Palumbo, G. G. C., Saracco, P., Hickson, P. and Mendes de Oliveira, C.: 1995, Environment of compact groups of galaxies, *AJ* **109**, 1476–1484.

Patiri, S. G., Prada, F., Holtzman, J., Klypin, A. and Betancort-Rijo, J.: 2006, The properties of galaxies in voids, *MNRAS* **372**, 1710–1720.

Peebles, P. J. E.: 1980, *The large-scale structure of the universe*, Research supported by the National Science Foundation. Princeton, N.J., Princeton University Press, 1980. 435 p.

Peebles, P. J. E.: 2001, The Void Phenomenon, *ApJ* **557**, 495–504.

Percival, W. J., Baugh, C. M., Bland-Hawthorn, J., Bridges, T., Cannon, R., Cole, S., Colless, M., Collins, C., Couch, W., Dalton, G., De Propris, R., Driver, S. P., Efstathiou, G., Ellis, R. S., Frenk, C. S., Glazebrook, K., Jackson, C., Lahav, O., Lewis, I., Lumsden, S., Maddox, S., Moody, S., Norberg, P., Peacock, J. A., Peterson, B. A., Sutherland, W. and Taylor, K.: 2001, The 2dF Galaxy Redshift Survey: the power spectrum and the matter content of the Universe, *MNRAS* **327**, 1297–1306.

Ranft, J.: 1999, Cosmic ray particle production., *Nuclear Physics B Proceedings Supplements* **71**, 228–237.

Richards, G. T., Fan, X., Newberg, H. J., Strauss, M. A., Vanden Berk, D. E., Schneider, D. P., Yanny, B., Boucher, A., Burles, S., Frieman, J. A., Gunn, J. E., Hall, P. B., Ivezić, Ž., Kent, S., Loveday, J., Lupton, R. H., Rockosi, C. M., Schlegel, D. J., Stoughton, C., SubbaRao, M. and York, D. G.: 2002, Spectroscopic Target Selection in the Sloan Digital Sky Survey: The Quasar Sample, *AJ* **123**, 2945–2975.

Richards, G. T., Fan, X., Schneider, D. P., Vanden Berk, D. E., Strauss, M. A., York, D. G., Anderson, Jr., J. E., Anderson, S. F., Annis, J., Bahcall, N. A., Bernardi, M., Briggs, J. W., Brinkmann, J., Brunner, R., Burles, S., Carey, L., Castander, F. J., Connolly, A. J., Crocker, J. H., Csabai, I., Doi, M., Finkbeiner, D., Friedman, S. D., Frieman, J. A., Fukugita, M., Gunn, J. E., Hindsley, R. B., Ivezić, Ž., Kent, S., Knapp, G. R., Lamb, D. Q., Leger, R. F., Long, D. C., Loveday, J., Lupton, R. H., McKay, T. A., Meiksin, A., Merrelli, A., Munn, J. A., Newberg, H. J., Newcomb, M., Nichol, R. C., Owen, R., Pier, J. R., Pope, A., Richmond, M. W., Rockosi, C. M., Schlegel, D. J., Siegmund, W. A., Smee, S., Snir, Y., Stoughton, C., Stubbs, C., SubbaRao, M., Szalay, A. S., Szokoly, G. P., Tremonti, C., Uomoto, A., Waddell, P., Yanny, B. and Zheng, W.: 2001, Colors of 2625 Quasars at $0 < z < 5$ Measured in the Sloan Digital Sky Survey Photometric System, *AJ* **121**, 2308–2330.

Richards, G. T., Nichol, R. C., Gray, A. G., Brunner, R. J., Lupton, R. H., Vanden Berk, D. E., Chong, S. S., Weinstein, M. A., Schneider, D. P., Anderson, S. F., Munn, J. A., Harris, H. C., Strauss, M. A., Fan, X., Gunn, J. E., Ivezić, Ž., York, D. G., Brinkmann, J. and Moore, A. W.: 2004, Efficient Photometric Selection of Quasars from the Sloan Digital Sky Survey: 100,000 $z < 3$ Quasars from Data Release One, *ApJS* **155**, 257–269.

R.O. Duda, P. H. and Stork, D.: 2001, *Pattern classification*, Wiley.

Rojas, R. R., Vogeley, M. S., Hoyle, F. and Brinkmann, J.: 2004, Photometric Properties of Void Galaxies in the Sloan Digital Sky Survey, *ApJ* **617**, 50–63.

Rojas, R. R., Vogeley, M. S., Hoyle, F. and Brinkmann, J.: 2005, Spectroscopic Properties of Void Galaxies in the Sloan Digital Sky Survey, *ApJ* **624**, 571–585.

Roulet, E.: 2004, Astroparticle Theory:, *International Journal of Modern Physics A* **19**, 1133–1141.

Sadler, E. M., Jackson, C. A., Cannon, R. D., McIntyre, V. J., Murphy, T., Bland-Hawthorn, J., Bridges, T., Cole, S., Colless, M., Collins, C., Couch, W., Dalton, G., de Propris, R., Driver, S. P., Efstathiou, G., Ellis, R. S., Frenk, C. S., Glazebrook, K., Lahav, O., Lewis, I., Lumsden, S., Maddox, S., Madgwick, D., Norberg, P., Peacock, J. A., Peterson, B. A., Sutherland, W. and Taylor, K.: 2002, 2dF Galaxy Redshift Survey. II. (Sadler+, 2002), *VizieR Online Data Catalog* **732**, 90227–+.

Sandage, A., Tammann, G. A. and Yahil, A.: 1979, The velocity field of bright nearby galaxies. I - The variation of mean absolute magnitude with redshift for galaxies in a magnitude-limited sample, *ApJ* **232**, 352–364.

Sandage, A. and Wyndham, J. D.: 1965, On the Optical Identification of Eleven New Quasi-Stellar Radio Sources., *ApJ* **141**, 328–+.

Saunders, W., Sutherland, W. J., Maddox, S. J., Keeble, O., Oliver, S. J., Rowan-Robinson, M., McMahon, R. G., Efstathiou, G. P., Tadros, H., White, S. D. M., Frenk, C. S., Carramiñana, A. and Hawkins, M. R. S.: 2000, The PSCz catalogue, *MNRAS* **317**, 55–63.

Schmidt, M. and Green, R. F.: 1983, Quasar evolution derived from the Palomar bright quasar survey and other complete quasar surveys, *ApJ* **269**, 352–374.

Schneider, D. P., Gunn, J. E. and Hoessel, J. G.: 1983, CCD photometry of Abell clusters. I - Magnitudes and redshifts for 84 brightest cluster galaxies, *ApJ* **264**, 337–355.

Schneider, M., Knox, L., Zhan, H. and Connolly, A.: 2006, Using Galaxy Two-Point Correlation Functions to Determine the Redshift Distributions of Galaxies Binned by Photometric Redshift, *ApJ* **651**, 14–23.

Seo, H.-J. and Eisenstein, D. J.: 2003, Probing Dark Energy with Baryonic Acoustic Oscillations from Future Large Galaxy Redshift Surveys, *ApJ* **598**, 720–740.

Shimasaku, K., Fukugita, M., Doi, M., Hamabe, M., Ichikawa, T., Okamura, S., Sekiguchi, M., Yasuda, N., Brinkmann, J., Csabai, I., Ichikawa, S.-I., Ivezić, Z., Kunszt, P. Z., Schneider, D. P., Szokoly, G. P., Watanabe, M. and York, D. G.: 2001, Statistical Properties of Bright Galaxies in the Sloan Digital Sky Survey Photometric System, *AJ* **122**, 1238–1250.

Sigl, G., Miniati, F. and Enßlin, T. A.: 2004, Ultrahigh energy cosmic ray probes of large scale structure and magnetic fields, *Phys. Rev. D* **70**(4), 043007–+.

Singh, S., Ma, C.-P. and Arons, J.: 2004, Gamma-ray bursts and magnetars as possible sources of ultrahigh energy cosmic rays: Correlation of cosmic ray event positions with IRAS galaxies, *Phys. Rev. D* **69**(6), 063003–+.

Smialkowski, A., Giller, M. and Michalak, W.: 2002, Luminous infrared galaxies as possible sources of UHE cosmic rays, *Journal of Physics G Nuclear Physics* **28**, 1359–1374.

Sommers, P.: 2001, Cosmic ray anisotropy analysis with a full-sky observatory, *Astroparticle Physics* **14**, 271–286.

Spergel, D. N., Bean, R., Doré, O., Nolta, M. R., Bennett, C. L., Dunkley, J., Hinshaw, G., Jarosik, N., Komatsu, E., Page, L., Peiris, H. V., Verde, L., Halpern, M., Hill, R. S., Kogut, A., Limon, M., Meyer, S. S., Odegard, N., Tucker, G. S., Weiland, J. L., Wollack, E. and Wright, E. L.: 2007, Three-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Implications for Cosmology, *ApJS* **170**, 377–408.

Spergel, D. N., Verde, L., Peiris, H. V., Komatsu, E., Nolta, M. R., Bennett, C. L., Halpern, M., Hinshaw, G., Jarosik, N., Kogut, A., Limon, M., Meyer, S. S., Page, L., Tucker, G. S., Weiland, J. L., Wollack, E. and Wright, E. L.: 2003, First-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Determination of Cosmological Parameters, *ApJS* **148**, 175–194.

Staiano, A.: 2003, *Unsupervised Neural Networks for the Extraction of Scientific Information from Astronomical Data*, PhD thesis, University of Salerno.

Staiano, A. e. a.: 2004, Probabilistic principal surfaces for yeast gene microarray data-mining, *ICDM'04 - Fourth IEEE International Conference on Data Mining*, pp. 202–209.

Stornaiolo, C.: 2002, Cosmological Black Holes, *General Relativity and Gravitation* **34**, 2089–2099.

Stoughton, C., Lupton, R. H., Bernardi, M., Blanton, M. R., Burles, S., Castander, F. J., Connolly, A. J., Eisenstein, D. J., Frieman, J. A., Hennessy, G. S., Hindsley, R. B., Ivezić, Ž., Kent, S., Kunszt, P. Z., Lee, B. C., Meiksin, A., Munn, J. A., Newberg, H. J., Nichol, R. C., Nicinski, T., Pier, J. R., Richards, G. T., Richmond, M. W., Schlegel, D. J., Smith, J. A., Strauss, M. A., SubbaRao, M., Szalay, A. S., Thakar, A. R., Tucker, D. L., Vanden Berk, D. E., Yanny, B., Adelman, J. K., Anderson, Jr., J. E., Anderson, S. F., Annis, J., Bahcall, N. A., Bakken, J. A., Bartelmann, M., Bastian, S., Bauer, A., Berman, E., Böhringer, H., Boroski, W. N., Bracker, S., Briegel, C., Briggs, J. W., Brinkmann, J., Brunner, R., Carey, L., Carr, M. A., Chen, B., Christian, D., Colestock, P. L., Crocker, J. H., Csabai, I., Czarapata, P. C., Dalcanton, J., Davidsen, A. F., Davis, J. E., Dehnen, W., Dodelson, S., Doi, M., Dombeck, T., Donahue, M., Ellman, N., Elms, B. R., Evans, M. L., Eyer, L., Fan, X., Federwitz, G. R., Friedman, S., Fukugita, M., Gal, R., Gillespie, B., Glazebrook, K., Gray, J., Grebel, E. K., Greenawalt, B., Greene, G., Gunn, J. E., de Haas, E., Haiman, Z., Haldeman, M., Hall, P. B., Hamabe, M., Hansen, B., Harris, F. H., Harris, H., Harvanek, M., Hawley, S. L., Hayes, J. J. E., Heckman, T. M., Helmi, A., Henden, A., Hogan, C. J., Hogg, D. W., Holmgren, D. J., Holtzman, J., Huang, C.-H., Hull, C., Ichikawa, S.-I., Ichikawa, T., Johnston, D. E., Kauffmann, G., Kim, R. S. J., Kimball, T., Kinney, E., Klaene, M., Kleinman, S. J., Klypin, A., Knapp, G. R., Korienek, J., Krolik, J., Kron, R. G., Krzesiński, J., Lamb, D. Q., Leger, R. F., Limmongkol, S., Lindenmeyer, C., Long, D. C., Loomis, C., Loveday, J., MacKinnon, B., Mannery, E. J., Mantsch, P. M., Margon, B., McGehee, P., McKay, T. A., McLean, B., Menou, K., Merelli, A., Mo, H. J., Monet, D. G., Nakamura, O., Narayanan, V. K., Nash, T., Neilsen, Jr., E. H., Newman, P. R., Nitta, A., Odenkirchen, M., Okada, N., Okamura, S., Ostriker, J. P., Owen, R., Pauls, A. G., Peoples, J., Peterson, R. S., Petravick, D., Pope, A., Pordes, R., Postman, M., Prosapio, A., Quinn, T. R., Rechenmacher, R., Rivetta, C. H., Rix, H.-W., Rockosi, C. M., Rosner, R., Ruthmansdorfer, K., Sandford, D., Schneider, D. P., Scranton, R., Sekiguchi, M., Sergey, G., Sheth, R., Shimasaku, K., Smee, S., Snedden, S. A., Stebbins, A., Stubbs, C., Szapudi, I., Szkody, P., Szokoly, G. P., Tabachnik, S., Tsvetanov, Z., Uomoto, A., Vogeley, M. S., Voges, W., Waddell, P., Walterbos, R., Wang, S.-i., Watanabe, M., Weinberg, D. H., White, R. L., White, S. D. M., Wilhite, B., Wolfe, D., Yasuda, N., York, D. G., Zehavi, I. and Zheng, W.: 2002, Sloan Digital Sky Survey: Early Data Release, *AJ* **123**, 485–548.

Strateva, I., Ivezić, Ž., Knapp, G. R., Narayanan, V. K., Strauss, M. A., Gunn, J. E., Lupton, R. H., Schlegel, D., Bahcall, N. A., Brinkmann, J., Brunner, R. J., Budavári, T., Csabai, I., Castander, F. J., Doi, M., Fukugita, M., Győry, Z., Hamabe, M., Hennessy, G., Ichikawa, T., Kunszt, P. Z., Lamb, D. Q., McKay, T. A., Okamura, S., Racusin, J., Sekiguchi, M., Schneider, D. P., Shimasaku, K. and York, D.: 2001, Color Separation of Galaxy Types in the Sloan Digital Sky Survey Imaging Data, *AJ* **122**, 1861–1874.

Suchkov, A. A., Hanisch, R. J. and Margon, B.: 2005, A Census of Object Types and Redshift Estimates in the SDSS Photometric Catalog from a Trained Decision Tree Classifier, *AJ* **130**, 2439–2452.

Tagliaferri, R. e. a.: 2003a, Neural Networks, *"Neural Network Analysis of Complex Scientific Data: Astronomy and Geosciences"*, Vol. 16, pp. 297–321.

Tagliaferri, R. e. a.: 2003b, Neural networks for photometric redshift evaluation, *Lecture Notes in Computer Science, WIRN Vietri 03*, Vol. LNCS 2859, pp. 226–234.

Tagliaferri, R., Longo, G., Andreon, S., Capozziello, S., Donalek, C. and Giordano, G.: 2002, Neural Networks and Photometric Redshifts, *ArXiv Astrophysics e-prints* .

Takeda, M., Hayashida, N., Honda, K., Inoue, N., Kadota, K., Kakimoto, F., Kamata, K., Kawaguchi, S., Kawasaki, Y., Kawasumi, N., Kitamura, H., Kusano, E., Matsubara, Y., Murakami, K., Nagano, M., Nishikawa, D., Ohoka, H., Sakaki, N., Sasaki, M., Shinozaki, K., Souma, N., Teshima, M., Torii, R., Tsushima, I., Uchihori, Y., Yamamoto, T., Yoshida, S. and Yoshii, H.: 1998, Extension of the Cosmic-Ray Energy Spectrum beyond the Predicted Greisen-Zatsepin-Kuz'min Cutoff, *Physical Review Letters* **81**, 1163–1166.

Takeda, M., Hayashida, N., Honda, K., Inoue, N., Kadota, K., Kakimoto, F., Kamata, K., Kawaguchi, S., Kawasaki, Y., Kawasumi, N., Kusano, E., Matsubara, Y., Murakami, K., Nagano, M., Nishikawa, D., Ohoka, H., Osone, S., Sakaki, N., Sasaki, M., Shinozaki, K., Souma, N., Teshima, M., Torii, R., Tsushima, I., Uchihori, Y., Yamamoto, T., Yoshida, S. and Yoshii, H.: 1999, Small-Scale Anisotropy of Cosmic Rays above 10^{19} eV Observed with the Akeno Giant Air Shower Array, *ApJ* **522**, 225–237.

Tegmark, M., Eisenstein, D. J., Strauss, M. A., Weinberg, D. H., Blanton, M. R., Frieman, J. A., Fukugita, M., Gunn, J. E., Hamilton, A. J. S., Knapp, G. R., Nichol, R. C., Ostriker, J. P., Padmanabhan, N., Percival, W. J., Schlegel, D. J., Schneider, D. P., Scoccimarro, R., Seljak, U., Seo, H.-J., Swanson, M., Szalay, A. S., Vogeley, M. S., Yoo, J., Zehavi, I., Abazajian, K., Anderson, S. F., Annis, J., Bahcall, N. A., Bassett, B., Berlind, A., Brinkmann, J., Budavari, T., Castander, F., Connolly, A., Csabai, I., Doi, M., Finkbeiner, D. P., Gillespie, B., Glazebrook, K., Hennessy, G. S., Hogg, D. W., Ivezić, Ž., Jain, B., Johnston, D., Kent, S., Lamb, D. Q., Lee, B. C., Lin, H., Loveday, J., Lupton, R. H., Munn, J. A., Pan, K., Park, C., Peoples, J., Pier, J. R., Pope, A., Richmond, M., Rockosi, C., Scranton, R., Sheth, R. K., Stebbins, A., Stoughton, C., Szapudi, I., Tucker, D. L., Berk, D. E. V., Yanny, B. and York, D. G.: 2006, Cosmological constraints from the SDSS luminous red galaxies, *Phys. Rev. D* **74**(12), 123507–+.

Tegmark, M., Strauss, M. A., Blanton, M. R., Abazajian, K., Dodelson, S., Sandvik, H., Wang, X., Weinberg, D. H., Zehavi, I., Bahcall, N. A., Hoyle, F., Schlegel, D., Scoccimarro, R., Vogeley, M. S., Berlind, A., Budavari, T., Connolly, A., Eisenstein, D. J., Finkbeiner, D., Frieman, J. A., Gunn, J. E., Hui, L., Jain, B., Johnston, D., Kent, S., Lin, H., Nakajima, R., Nichol, R. C., Ostriker, J. P., Pope, A., Scranton, R., Seljak, U., Sheth, R. K., Stebbins, A., Szalay, A. S., Szapudi, I., Xu, Y., Annis, J., Brinkmann, J., Burles, S., Castander, F. J., Csabai, I., Loveday, J., Doi, M., Fukugita, M., Gillespie, B., Hennessy, G., Hogg, D. W., Ivezić, Ž., Knapp, G. R., Lamb, D. Q., Lee, B. C., Lupton, R. H., McKay, T. A., Kunszt, P., Munn, J. A., O'Connell, L., Peoples, J., Pier, J. R., Richmond, M., Rockosi, C., Schneider, D. P., Stoughton, C., Tucker, D. L., vanden Berk, D. E., Yanny, B. and York, D. G.: 2004, Cosmological parameters from SDSS and WMAP, *Physical Review D* **69**(10), 103501–+.

Tegmark, M. and Zaldarriaga, M.: 2002, Separating the early universe from the late universe: Cosmological parameter estimation beyond the black box, *Phys. Rev. D* **66**(10), 103508–+.

The Pierre Auger Collaboration: 2005a, First Estimate of the Primary Cosmic Ray Energy Spectrum above 3 EeV from the Pierre Auger Observatory, *ArXiv Astrophysics e-prints* .

The Pierre Auger Collaboration: 2005b, Performance of the Pierre Auger Observatory Surface Array, *ArXiv Astrophysics e-prints* .

The Pierre Auger Collaboration: 2007, Correlation of the highest energy cosmic rays with nearby extragalactic objects, *ArXiv e-prints* **711**.

Tiersch, H., Stoll, D., Neizvestny, S., Amirkhanian, A. S. and Egikian, A. G.: 1999, Emission-Line Galaxies in Shahbazian Compact Groups, *in* Y. Terzian, E. Khachikian and D. Weedman (eds), *Activity in Galaxies and Related Phenomena*, Vol. 194 of *IAU Symposium*, pp. 394–+.

Tiersch, H., Tovmassian, H. M., Stoll, D., Amirkhanian, A. S., Neizvestny, S., Böhringer, H. and MacGillivray, H. T.: 2002, Shakhbazian compact galaxy groups. I. Photometric, spectroscopic and X-ray study of ShCG 154, ShCG 166, ShCG 328, ShCG 360, *A&A* **392**, 33–52.

Tinyakov, P. G. and Tkachev, I. I.: 2001, BL Lacertae are Probable Sources of the Observed Ultrahigh Energy Cosmic Rays, *Soviet Journal of Experimental and Theoretical Physics Letters* **74**, 445–+.

Tovmassian, H. e. a.: 2005a, Shakhbazian compact galaxy groups. IV. Photometric and spectroscopic study of ShCG 8, ShCG 14 ShCG 19, ShCG 22, *A&A* **439**, 973–979.

Tovmassian, H. M., Chavushyan, V. H., Verkhodanov, O. V. and Tiersch, H.: 1999, Radio Emission of Shakhbazian Compact Galaxy Groups, *ApJ* **523**, 87–99.

Tovmassian, H. M. e. a.: 2005b, Spectroscopy and photometry of ShCG 191 - Abell 1097, *Astronomische Nachrichten* **326**, 362–369.

Tovmassian, H. M., Tiersch, H., Navarro, S. G., Chavushyan, V. H., Tovmassian, G. H., Amirkhanian, A. S. and Neizvestny, S.: 2003, Photometric and Spectroscopic Study of the Shakhbazian Compact Galaxy Groups ShCG 31 ShCG 38, ShCG 43, and ShCG 282, *Revista Mexicana de Astronomia y Astrofisica* **39**, 275–289.

Tovmassian, H. M., Tiersch, H., Navarro, S. G., Chavushyan, V. H., Tovmassian, G. H. and Neizvestny, S.: 2004, Shakhbazian compact galaxy groups. III. Photometric and spectroscopic study of ShCG 181, ShCG 344,ShCG 361, and ShCG 362, *A&A* **415**, 803–811.

Tovmassian, H. M., Tiersch, H., Tovmassian, G. H., Chavushyan, V. H., Navarro, S. G., Neizvestny, S. and Torres-Papaqui, J. P.: 2005, Photometric and Spectroscopic Study of the Shakhbazian Compact Galaxy Groups ShCG74 ShCG188, ShCG251, and ShCG348, *Revista Mexicana de Astronomia y Astrofisica* **41**, 3–16.

Treister, E. and Urry, C. M.: 2006, The Evolution of Obscuration in Active Galactic Nuclei, *ApJL* **652**, L79–L82.

Treister, E., Urry, C. M., Chatzichristou, E., Bauer, F., Alexander, D. M., Koekemoer, A., Van Duyne, J., Brandt, W. N., Bergeron, J., Stern, D., Moustakas, L. A., Chary, R.-R., Conselice, C., Cristiani, S. and Grogin, N.: 2004, Obscured Active Galactic Nuclei and the X-Ray, Optical, and Far-Infrared Number Counts of Active Galactic Nuclei in the GOODS Fields, *ApJ* **616**, 123–135.

Tully, R. B.: 1987, Nearby groups of galaxies. II - an all-sky survey within 3000 kilometers per second, *ApJ* **321**, 280–304.

Vanzella, E., Cristiani, S., Fontana, A., Nonino, M., Arnouts, S., Giallongo, E., Grazian, A., Fasano, G., Popesso, P., Saracco, P. and Zaggia, S.: 2004, Photometric redshifts with the Multilayer Perceptron Neural Network: Application to the HDF-S and SDSS, *A&A* **423**, 761–776.

Vennik, J., Richter, G. M. and Longo, G.: 1993, The Neighbourhoods of the Nearest Hickson Groups, *Astronomische Nachrichten* **314**, 393–+.

Vermeulen, R. C., Browne, I. W. A., Cohen, M. H., Goodrich, R. W., Ogle, P. M., Readhead, A. C. S. and Tran, H. D.: 1995, BL Lacertae, *IAU Circ.* **6176**, 2–+.

Vogeley, M. S., Geller, M. J., Park, C. and Huchra, J. P.: 1994, Voids and constraints on nonlinear clustering of galaxies, *AJ* **108**, 745–758.

Wadadekar, Y.: 2005, Estimating Photometric Redshifts Using Support Vector Machines, *PASP* **117**, 79–85.

Walker, R. and Watson, A. A.: 1982, Measurement of the fluctuations in the depth of maximum of showers produced by primary particles of energy greater than 1.5×10^{17} eV, *Journal of Physics G Nuclear Physics* **8**, 1131–1140.

Walton, N. A.: 2002, AstroGrid: Powering the virtual universe, *Astronomy and Geophysics* **43**, 30–1.

Walton, N. A., Richards, A. M. S., Padovani, P. and Allen, M. G.: 2006, The Virtual Observatories: a major new facility for astronomy: linking ELTs, great observatories and the science community, *in* P. Whitelock, M. Dennefeld and B. Leibundgut (eds), *The Scientific Requirements for Extremely Large Telescopes*, Vol. 232 of *IAU Symposium*, pp. 398–403.

Wang, Y., Bahcall, N. and Turner, E. L.: 1998, A Catalog of Color-based Redshift Estimates for $z \leq 4$ Galaxies in the Hubble Deep Field, *AJ* **116**, 2081–2085.

Warren, S. J. and Hewett, P. C.: 1990, The detection of high-redshift quasars., *Reports of Progress in Physics* **53**, 1095–1135.

Warren, S. J., Hewett, P. C. and Foltz, C. B.: 2000, The KX method for producing K-band flux-limited samples of quasars, *MNRAS* **312**, 827–832.

Warren, S. J., Hewett, P. C. and Osmer, P. S.: 1991, A wide-field multicolor survey for high-redshift quasars, Z above 2.2. II - The quasar catalog, *ApJS* **76**, 23–54.

Watson, A. M., Gallagher, III, J. S., Holtzman, J. A., Hester, J. J., Mould, J. R., Ballester, G. E., Burrows, C. J., Casertano, S., Clarke, J. T., Crisp, D., Evans, R., Griffiths, R. E., Hoessel, J. G., Scowen, P. A., Stapelfeldt, K. R., Trauger, J. T. and Westphtptphal, J. A.: 1996, The Discovery of Young, Luminous, Compact Stellar Clusters in the Starburst Galaxy NGC 253, *AJ* **112**, 534–+.

Waxman, E., Fisher, K. B. and Piran, T.: 1997, The Signature of a Correlation between Cosmic-Ray Sources above 10 19 eV and Large-Scale Structure, *ApJ* **483**, 1–+.

Way, M. J. and Srivastava, A. N.: 2006, Novel Methods for Predicting Photometric Redshifts from Broadband Photometry Using Virtual Sensors, *ApJ* **647**, 102–115.

Wdowczyk, J. and Wolfendale, A. W.: 1979, Diffusion of the highest energy cosmic rays from Virgo, *Nature* **281**, 356–+.

White, M., van Waerbeke, L. and Mackey, J.: 2002, Completeness in Weak-Lensing Searches for Clusters, *ApJ* **575**, 640–649.

Willis, J. P., Hewett, P. C. and Warren, S. J.: 2001, Luminous early-type field galaxies at $z \sim 0.4$ - I. Observations and redshift catalogue of 581 galaxies, *MNRAS* **325**, 1002–1016.

Winn, M. M., Ulrichs, J., Peak, L. S., McCusker, C. B. A. and Horton, L.: 1986, The cosmic-ray energy spectrum above 10^{17} eV., *Journal of Physics G Nuclear Physics* **12**, 653–674.

Winter, C., Jeffery, C. S. and Drilling, J. S.: 2004, Automatic Classification of Subdwarf Spectra using a Neural Network, *Ap&SS* **291**, 375–378.

Yasuda, N., Fukugita, M., Narayanan, V. K., Lupton, R. H., Strateva, I., Strauss, M. A., Ivezić, Ž., Kim, R. S. J., Hogg, D. W., Weinberg, D. H., Shimasaku, K., Loveday, J., Annis, J., Bahcall, N. A., Blanton, M., Brinkmann, J., Brunner, R. J., Connolly, A. J., Csabai, I., Doi, M., Hamabe, M., Ichikawa, S.-I., Ichikawa, T., Johnston, D. E., Knapp, G. R., Kunszt, P. Z., Lamb, D. Q., McKay, T. A., Munn, J. A., Nichol, R. C., Okamura, S., Schneider, D. P., Szokoly, G. P., Vogeley, M. S., Watanabe, M. and York, D. G.: 2001, Galaxy Number Counts from the Sloan Digital Sky Survey Commissioning Data, *AJ* **122**, 1104–1124.

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

Zatsepin, G. T. and Kuz'min, V. A.: 1966, Upper Limit of the Spectrum of Cosmic Rays, *Soviet Journal of Experimental and Theoretical Physics Letters* **4**, 78–+.

Zhan, H.: 2006, Cosmic tomographies: baryon acoustic oscillations and weak lensing, *Journal of Cosmology and Astro-Particle Physics* **8**, 8–+.

Zhan, H. and Knox, L.: 2006, Baryon Oscillations and Consistency Tests for Photometrically Determined Redshifts of Very Faint Galaxies, *ApJ* **644**, 663–670.