The Astrophysical Journal, 663:752–764, 2007 July 10
© 2007. The American Astronomical Society. All rights reserved. Printed in U.S.A.

# MINING THE SDSS ARCHIVE. I. PHOTOMETRIC REDSHIFTS IN THE NEARBY UNIVERSE

Raffaele D'Abrusco,[1,2] Antonino Staiano,[3] Giuseppe Longo,[1,4,5] Massimo Brescia,[5,4] Maurizio Paolillo,[1,4] Elisabetta De Filippis,[5,1] and Roberto Tagliaferri[6,4]

*Received 2006 October 11; accepted 2007 March 2*

## ABSTRACT

We present a supervised neural network approach to the determination of photometric redshifts. The method was fine-tuned to match the characteristics of the Sloan Digital Sky Survey, and as base of "a priori" knowledge, it exploits the rich wealth of spectroscopic redshifts provided by this survey. In order to train, validate, and test the networks, we used two galaxy samples drawn from the SDSS spectroscopic data set, namely, the general galaxy sample (GG) and the luminous red galaxy subsample (LRG). The method consists of a two-step approach. In the first step, objects are classified as nearby ($z < 0.25$) and distant ($0.25 < z < 0.50$), with an accuracy estimated as 97.52%. In the second step, two different networks are separately trained on objects belonging to the two redshift ranges. Using a standard multilayer perceptron operated in a Bayesian framework, the optimal architectures were found to require one hidden layer of 24 (24) and 24 (25) neurons for the GG (LRG) sample. The final results on the GG data set give a robust $\sigma_z \simeq 0.0208$ over the redshift range $[0.01, 0.48]$ and $\sigma_z \simeq 0.0197$ and $\simeq 0.0238$ 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 $\sigma_z \simeq 0.0160$ and $\simeq 0.0183$ 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 circa 30 million galaxies having $z < 0.5$ were derived. A catalog containing redshifts for the LRG subsample was also produced.

*Subject headings:* galaxies: distances and redshifts — galaxies: photometry — large-scale structure of universe

*Online material:* color figures

## 1. INTRODUCTION

After the pioneering work by the Belgian astronomer Vandererkhoven, who in the late 1930s 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 experimentally test the idea that redshift could be obtained from multiband aperture photometry by sampling, at different wavelengths, the galaxy spectral energy distribution (SED). After a period of relative lack of interest, the "photometric redshifts" technique was resurrected in the 1980s (Butchins 1981), when it became clear that it could prove useful in two similar but methodologically very different fields of application:

1. As a method to evaluate distances when spectroscopic estimates become impossible due either to poor signal-to-noise ratio, to instrumental systematics, or to the fact that the objects under study are beyond the spectroscopic limit (cf. Bolzonella et al. 2002).
2. 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 wealth of data produced by a new generation of digital surveys, consisting of accurate multiband photometric data for tens and even hundreds of millions of extragalactic objects, have 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 for deriving a three-dimensional map of the universe, a map which is crucial for a variety of applications, to name just a few: studying large-scale structure (Brodwin et al. 2006), constraining the cosmological constants and models (Blake & Bridle 2005, and references therein; Budavári et al. 2003; Tegmark et al. 2006), and mapping matter distribution using weak lensing (Edmondson et al. 2003, 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 catalogs of photometric redshifts, one for ~30 million objects extracted form the SDSS DR5 main GALAXY data set and a second one for a luminous red galaxy sample.

The paper is structured as follows. In §§ 2 and 3, we briefly summarize the various methods for the determination of photometric redshifts and the theory behind the adopted model of neural network. In § 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 § 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 § 7

[1] Department of Physical Sciences, University of Napoli Federico II, via Cinthia 9, 80126 Napoli, Italy.
[2] Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB4 0HA, UK.
[3] Department of Applied Science, University of Napoli Parthenope, via A. De Gasperi 5, 80133 Napoli, Italy.
[4] INFN-Napoli Unit, Department of Physical Sciences, via Cinthia 9, 80126, Napoli, Italy.
[5] INAF-Italian National Institute of Astrophysics, via del Parco Mellini, Rome, Italy
[6] Department of Mathematics and Applications, University of Salerno, Fisciano, Italy.

we therefore present a discussion of both systematic and random errors and propose a possible strategy to correct for systematic errors (§ 6). In § 8 we briefly describe the two catalogs. Finally, in § 9, we discuss the results and present our conclusions.

This paper is the first in a series of three. In the second one (M. Brescia et al. 2007, in preparation), we shall present the catalog 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 (R. D'Abrusco et al. 2007, in preparation) we shall complement the information contained in the above-quoted catalogs by discussing the statistical clustering of objects in the photometric parameter space.

## 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 & 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 pros and cons widely discussed in the literature (cf. Koo 1999; but see also Fernández-Soto et al. 2001; Massarotti et al. 2001a, 2001b; 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 hand, empirical templates are necessarily derived from nearby and bright galaxies and therefore may not be 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 data set are in progress and will surely prove useful in the near 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, i.e., the empirical ones, can be applied only to "mixed surveys," i.e., to data sets 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 colors 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 color range in several intervals in order to optimize the interpolation. Csabai et al. (2003) used instead an improved nearest neighbor method consisting of 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.

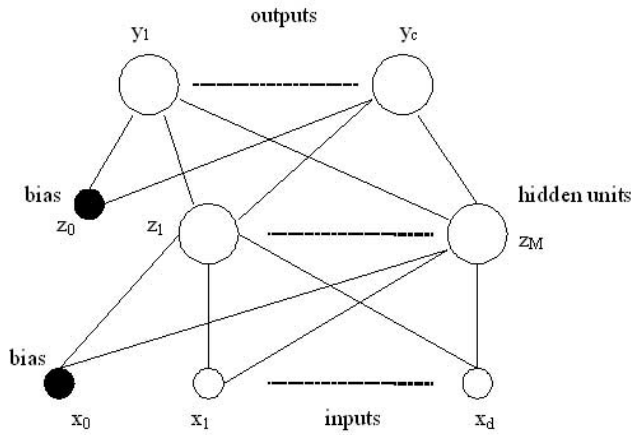


FIG. 1.—Schematic representation of the multilayer perceptron.

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 cases where a suitable base of knowledge is missing.

## 3. THE MULTILAYER PERCEPTRON

Neural Networks (hereafter NNs) have long been known to be excellent tools for interpolating data and for extracting patterns and trends, and in the last few years they have also found their way into the astronomical community and are used in a variety of applications (see the reviews by Tagliaferri et al. 2003a, 2003b, and references therein), ranging from star-galaxy separation (Donalek 2007), spectral classification (Winter et al. 2004), and photometric redshift 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 nonlinear (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 discussed below, we used a very simple neural model known as *multilayer 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. 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 Figure 1 would be called a 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.

### 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 $\boldsymbol{x} = \{x_1, x_2, \ldots, x_d\}$. The first layer of the network forms a linear combination of these inputs to give a set of intermediate activation variables,

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

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 bias 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

$$z_j = \tanh\left(a_j^{(1)}\right), \quad j = 1, \ldots, M. \tag{2}$$

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

$$a_k^{(2)} = \sum_{j=1}^{M} w_{kj}^{(1)} z_j + b_k^{(2)}, \quad k = 1, \ldots, c, \tag{3}$$

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

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

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

### 3.2. *MLP Training Phase*

The basic learning algorithm for MLPs is the so-called back-propagation and is based on the error-correction learning rule. In essence, back-propagation 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 $\boldsymbol{t} = \{t_1, t_2, \ldots, t_c\}$) 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,

1. For regression problems we adopted the sum-of-squares error function:

$$E = \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{c} \left[y_k(\boldsymbol{x}^n; \boldsymbol{w}) - t_k^n\right]^2.$$

2. For classification problems we used the cross-entropy error function:

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

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 computationally more efficient variant of the back-propagation algorithm, namely, the quasi-Newtonian 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 weights and biases. The $\nu$ 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-layer 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 layers of weights can approximate any decision boundary to arbitrary accuracy. Thus, such networks also provide universal nonlinear discriminant functions. More generally, the capability of such networks of approximating 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 (Duda & Hart 1973) 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 was just mentioned, it is also possible to train NNs in a Bayesian framework, which allows us to find the most efficient ones 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 $\alpha$ and $\beta$. The parameter $\alpha$ is related to the weights of the network and allows us to estimate the relative importance of the different inputs and the selection of the input parameters which are most relevant to a given task (*automatic relevance determination*; Bishop 1995). In fact, a larger value for a component of $\alpha$ implies a less meaningful corresponding weight. The parameter $\beta$ is instead related to the variance of the noise (a smaller value corresponding to a larger value of the noise)

TABLE 1
List of the Parameters Extracted from the SDSS Database and Used in the Experiments

| Parameter (1) | N (2) | Explanation (3) | F/L (4) |
|---|---|---|---|
| objID | ... | SDSS identification code | ... |
| R.A. | ... | Right ascension (J2000.0) | ... |
| decl. | ... | Declination (J2000.0) | ... |
| $\texttt{petroR50}_i$ | 1 | 50% of Petrosian radius in the $i$th band, $i = u, g, r, i, z$ | $F$ |
| $\texttt{petroR90}_i$ | 2 | 90% of Petrosian radius in the $i$th band, $i = u, g, r, i, z$ | $F$ |
| $\texttt{dered}_i$ | 3 | Dereddened magnitude in the $i$th band, $i = u, g, r, i, z$ | $F$ |
| $\texttt{lnLDeV}_r$ | 4 | log likelihood for De Vaucouleurs profile, $r$ band | $F$ |
| $\texttt{lnLExp}_r$ | 5 | log likelihood for exponential profile, $r$ band | $F$ |
| $\texttt{lnLStar}_r$ | 6 | log likelihood for PSF profile, $r$ band | $F$ |
| Spectroscopic redshift | $z$ | ... | $L$ |
| Spectral classification index | specClass | ... | $L$ |

Notes.—Col. (1): SDSS code. Col. (2): Running number for features only. Col. (3): Short explanation. Col. (4): Type of parameter, either feature ($F$) or label ($L$).

and therefore to a lower reliability of the network. The implementation of a Bayesian framework requires several steps: initialization of weights and hyperparameters and training the network via a nonlinear optimization algorithm in order to minimize the total error function. Every few cycles of the algorithm, the hyperparameters are reestimated and eventually the cycles are reiterated.

## 4. THE DATA AND THE "BASE OF KNOWLEDGE"

The Sloan Digital Sky Survey (hereafter SDSS) is an ongoing survey to image approximately $\pi$ sr 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.[7] The existence of such a spectroscopic subset (SpS), together with the accurate characterization of biases and errors, renders the SDSS a unique and ideal playground 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 criteria and use the GALAXY table membership. The data used in this work were therefore extracted from the SDSS catalogs. In particular, the spectroscopic subsample (SpS), used for training and testing purposes, was extracted from Data Release 4 (DR4; e.g., Adelman-McCarthy et al. 2006). While this work was in progress, Data Release 5 (DR5) was made publicly available. Thus, the photometric data used to produce the final catalogs were derived from the latter data. We wish to stress that this extension of the data set 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:

1. *General Galaxy Sample or GG.*—Composed of 445,933 objects with $z < 0.5$ matching the following selection criteria: dereddened magnitude in $r$ band, $r < 21$, and mode $= 1$, which corresponds to primary objects only in the case of deblended sources.

2. *Luminous Red Galaxy sample or LRG.*—Composed of 97,475 Luminous Red Galaxy candidates having spectroscopic redshift $<0.5$.

The SDSS spectroscopic survey (Eisenstein et al. 2001) was planned in order to favor the observation of the so-called luminous red galaxies or LRGs, which are expected to represent a more homogeneous population of luminous elliptical galaxies that 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 us 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, e.g., Hamilton 1985; Gladders & Yee 2000; Eisenstein et al. 2001; Willis et al. 2001; 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 magnitudes or colors alone, estimates of photometric redshifts should become more accurate. Such an effect was, for instance, found by Tagliaferri et al. (2002) and 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 1. These parameters are of two types; those which we call "features" (marked as $F$ in Table 1) are parameters which potentially may carry some useful information capable of improving 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 the magnitudes concerned, and at in contrast with other groups who used `modelMag`, we used the so-called dereddened magnitudes (`dered`), corrected for the best available estimate of the SDSS photometric zero points:

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

[7] See the SDSS Web pages at http://www.sdss.org/ for further details.

TABLE 2
RESULTS OF THE FEATURE SIGNIFICANCE ESTIMATION

| Parameters (1) | $\sigma_3$ (2) |
|---|---|
| All | 0.0202 |
| All but 1 | 0.0209 |
| All but 2 | 0.0213 |
| All but 4 and 5 | 0.214 |
| All but 6 | 0.215 |
| Only magnitudes | 0.0199 |

NOTES.—Col. (1): Features used. Features are numbered as in Table 1. Col. (2): Robust $\sigma$ of the residuals.

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 all data sets equally.

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.

### 4.1. *Feature 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 multilayer perceptron with one hidden layer and 24 neurons. In each experiment, the training, validation, and test sets were constructed by randomly extracting from the overall data set 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 $i = 1, \ldots, N$. For each experiment, following Csabai et al. (2003), we used the test set to evaluate the robust variance $\sigma_3$ obtained by excluding all points whose dispersion is larger then 3 $\sigma$ (see § 7). The values are listed in Table 2.

As can be seen, the most significant parameters are the magnitudes (or the colors). 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. Despite what was found in Vanzella et al. (2004) and Tagliaferri et al. (2002), the fact that 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 color parameter space is (on average, but see below) better mapped.

## 5. THE EVALUATION OF PHOTOMETRIC REDSHIFTS

One preliminary consideration: as was first pointed out by Connolly et al. (1995), when working in the near and intermediate redshift universe ($z < 1$), the most relevant broadband 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.

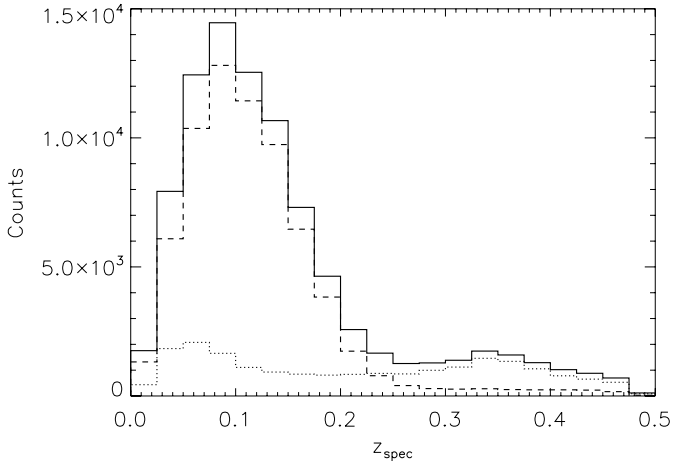


FIG. 2.—Distribution of redshifts in the SpS sample. *Solid line*: GG sample. *Dashed line*: Non-LRG sample. *Dotted line*: LRG sample (see text for details). Notice the sharp drop at $z \sim 0.25$.

One additional reason comes from the redshift distribution of the objects in the SpS DR4 shown in Figure 2 (*solid line*). As 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 $z > 0.5$.

In Figure 2 we also, plot as a dotted line, the redshift distribution of the galaxies in the SpS data set which match the LRG photometric selection criteria. As can be seen, within the tail at $z > 0.25$ only a very small fraction (11.4%) of the objects do not match the LRG selection criteria. In Figure 3 we plot the redshift of objects belonging to the GG sample against their luminosity in the $r$ band; black dots represent those galaxies which have been a posteriori identified as LRG. As is clearly seen, the overall distribution at redshift $\leq 0.25$ 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 dishomogeneity in the density and nature of training data poses severe constraints on any empirical method, since

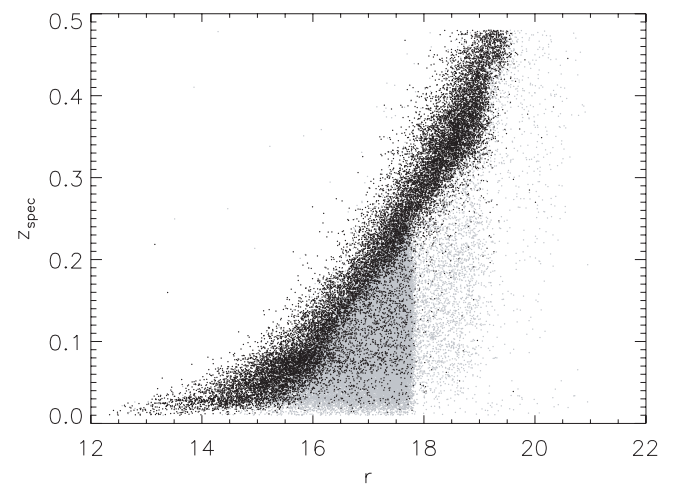


FIG. 3.—Distribution of the objects in the GG sample vs. the $r$ magnitude (*gray dots*). We plot the LRG objects as black dots. [*See the electronic edition of the Journal for a color version of this figure* .]

TABLE 3
CONFUSION MATRIX FOR THE "NEARBY-DISTANT" TEST SET

| | SDSS nearby | SDSS far |
|---|---|---|
| NN nearby..... | 76498 | 1096 |
| NN far..... | 1135 | 11145 |

the different weights of samples extracted in the different redshift bins would lead either to overfitting in the densest region or to 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 (note that in this paper we use the word "cluster" in the statistical sense, i.e., 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 we will show in Paper III (R. D'Abrusco et al. 2007, in preparation), 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.

### 5.1. *The Nearby and Intermediate Redshifts Samples*

In order to tackle the above-mentioned problems, we adopted a two-step approach; first, we trained a network to recognize nearby (i.e., 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 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 DR4 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 one 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 3. As can be seen, this first NN is capable of separating 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%).

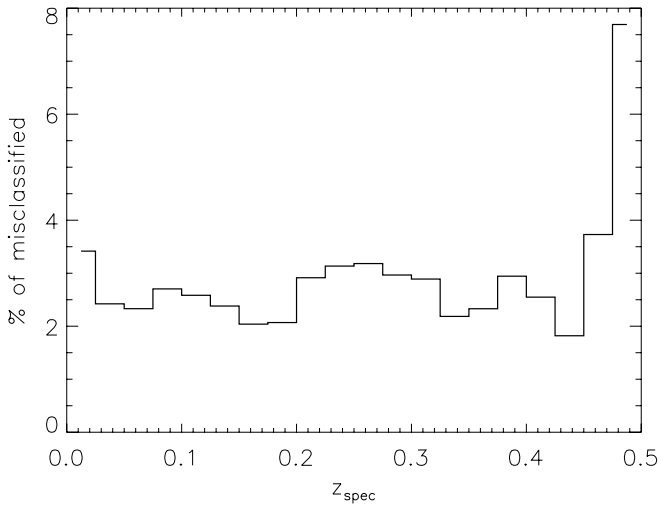


FIG. 4.—Percentage distribution of misclassified objects of GG sample normalized to the total number of galaxies in each redshift bin.

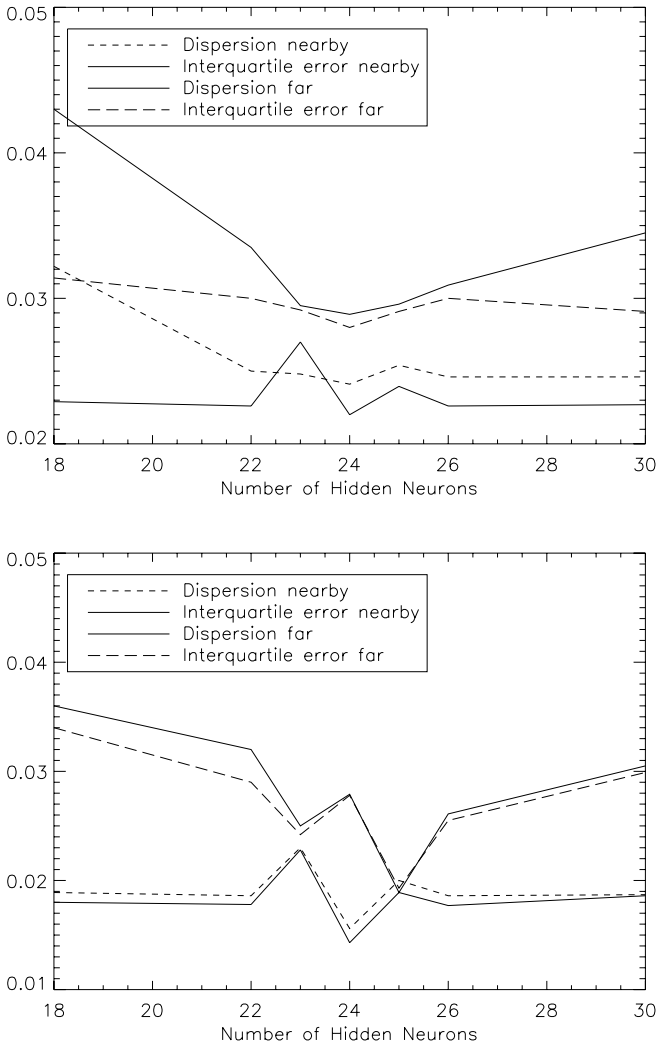


FIG. 5.—*Top*: GG sample, trend of the interquartile error and of the robust $\sigma$ as a function of the number N of the neurons in the hidden layer. The nearby and distant samples are plotted separately. *Bottom*: Same as above, but for the LRG sample.

In Figure 4 we plot against the redshift the percentage (calculated binning over the redshifts) for the objects in the test set which were misclassified (i.e., objects belonging to the nearby sample which were erroneously attributed to the distant one and vice versa). The distribution appears fairly constant from $z_{spec} \sim 0.05$ to $\sim 0.45$, while higher (but still negligible respect to the total number of objects in the sample) percentages are found at the extremes.

Note that when using photometric data alone, the absence of training data for $z > 0.5$ does not allow us 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.

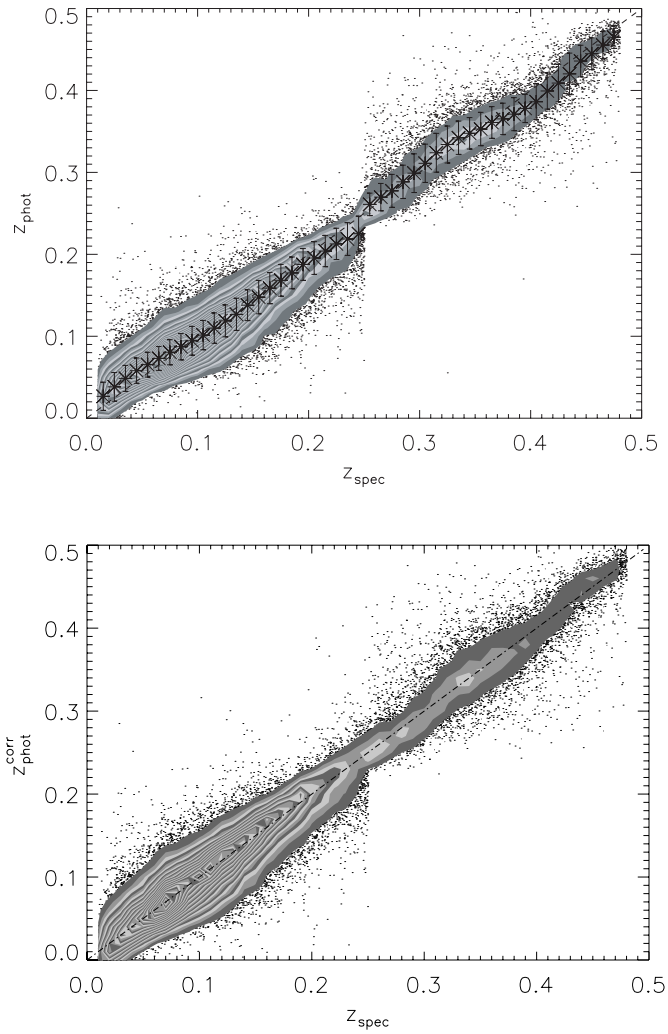


FIG. 6.—*Top*: Photometric vs. spectroscopic redshifts for the objects in the GG test set. The continuous lines are isodensity contours increasing with a step of 2% of the maximum density. The crosses mark the average value of photometric redshifts in a specific spectroscopic redshift bin (see text), while the error bars give the robust variance $\sigma_3$. *Bottom*: Same as above, but after the correction for the systematic trends via interpolation (see text).

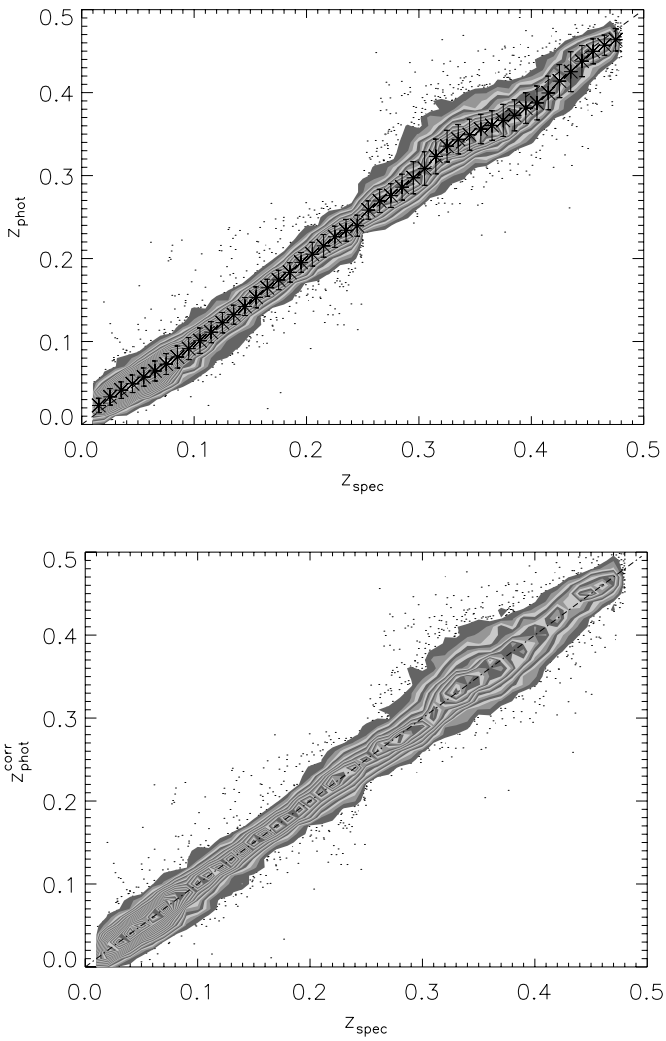


FIG. 7.—Same as Fig. 6, but for the LRG sample.

### 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 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 Figure 5 we give the trends as a function of the number of hidden neurons, of the interquartile errors, and of 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 the LRG sample, we obtained $\sigma_3 \simeq 0.0163$ over the whole range, and $\sigma_3 \simeq 0.0154$ and $\simeq 0.0189$ for the nearby and distant samples, respectively. In the upper panels of Figures 6 and 7 we plot the spectroscopic versus the photometric redshifts for the GG and the LRG samples, respectively. Due to the huge number of points that 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).

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 Figure 8 and in Figure 9, where we plot for each 0.05 redshift bin the average value of the photometric redshifts and the robust $\sigma$ of the residuals.

## 6. INTERPOLATIVE CORRECTION

The most significant deviations, as expected (Connolly et al. 1995), are clearly visible in the nearby sample for $z < 0.1$ and in

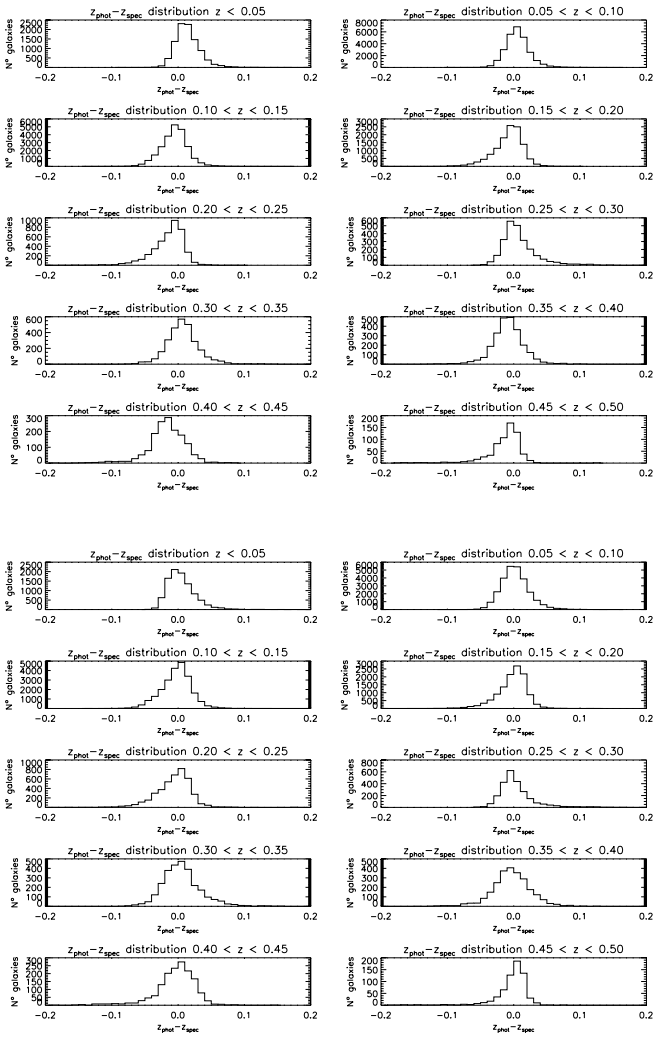


FIG. 8.—Histograms of residuals for the GG sample in slices of redshift. *Upper panels*: Before the correction. *Lower panels*: After the correction.

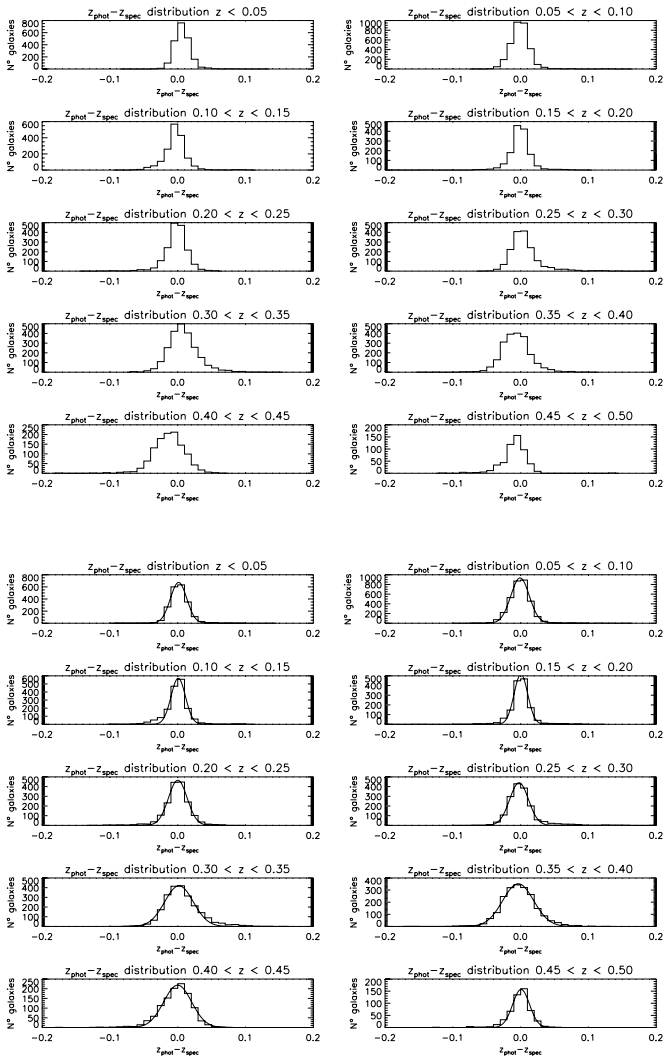


FIG. 9.—Same as Fig. 8, but for the LRG sample.

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 color. 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 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 redshift 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 catalogs, we applied to both data sets an interpolative correction computed separately in the two redshift intervals. We used a $\chi^2$ 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,

$$P_4[0.005, 1.570, -12.577, 78.948, -157.961] \quad (4)$$

$$P_5[12.15, -178.2, 1039.3, -2959.0, 4135.5, -2271.3], \quad (5)$$

and for the LRG sample,

$$P_4[0.011, 0.885, -1.820, 21.350, -53.159] \quad (6)$$

$$P_5[13.1, -192.5, 1123.3, -3207.2, 4504.5, -2491.6]. \quad (7)$$

Thus, the correction to be applied is

$$z_{\rm phot}^{\rm corr} = z_{\rm phot} - \left(z_{\rm phot}^{\rm calc} - z_{\rm spec}\right), \quad (8)$$

where $z_{\rm phot}^{\rm calc} = P_4(z_{\rm spec})$ for near objects and $z_{\rm phot}^{\rm calc} = P_5(z_{\rm spec})$ 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. [8]) to correct NNs' $z_{\rm phot}$ estimates for systematic trends, and we are obliged to use an approximation. In other words, we replace the unknown $z_{\rm spec}$ with $z_{\rm phot}$ in the equation. (8), obtaining the relation

$$\tilde{z}_{\rm phot}^{\rm corr} = z_{\rm phot} - \left(\tilde{z}_{\rm phot}^{\rm calc} - z_{\rm phot}\right), \quad (9)$$

where $\tilde{z}_{\rm phot}^{\rm calc} = P_4(z_{\rm phot})$ or $P_5(z_{\rm phot})$, depending on the redshift range.

This is equivalent to assuming that the same NN's $z_{\rm phot}$ distribution represents, with good approximation, the underlying and

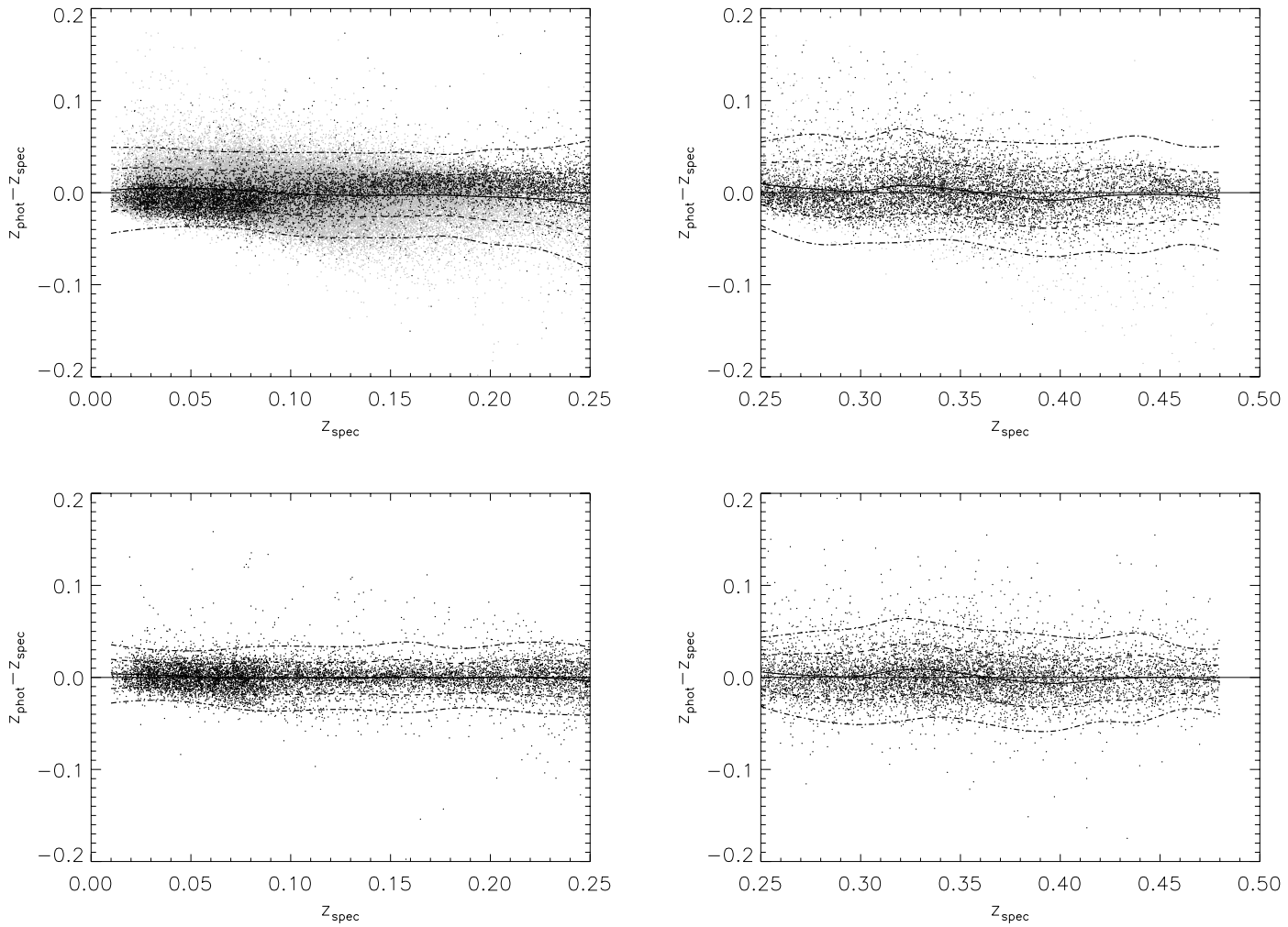


FIG. 10.—Distribution of the residuals vs. spectroscopic redshift after the correction for systematic trends. *Top panels*: GG nearby and distant samples. *Bottom panels*: LRG nearby and distant samples. The central line marks the average value of the residuals. The 1 and 2 $\sigma$ confidence levels are also shown. [*See the electronic edition of the Journal for a color version of this figure*.]

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 Figures 6 and 7.

## 7. DISCUSSION OF THE SYSTEMATICS AND OF THE ERRORS

As noticed by several authors (see, e.g., 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 (Padmanabhan et al. 2005), since 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 modeled. 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 & Jain 2004; Huterer et al. 2006; Ishak 2005; Ma et al. 2006), and baryon oscillations (Zhan 2006; Zhan & 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. 2001, 2002), or more complex error distributions, as can be seen in Figure 8 within the GG sample.

In order to evaluate the robustness of the $\sigma_r$, 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 Figure 10 separately for the near and distant objects. We have also studied the dependence of such residuals on the $r$-band luminosity of the galaxies in the two different magnitude ranges (cf. § 5; $r < 17.7$ and >17.7) and in the near and intermediate redshift bins, as shown in Figures 11 and 12 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 by remembering 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.

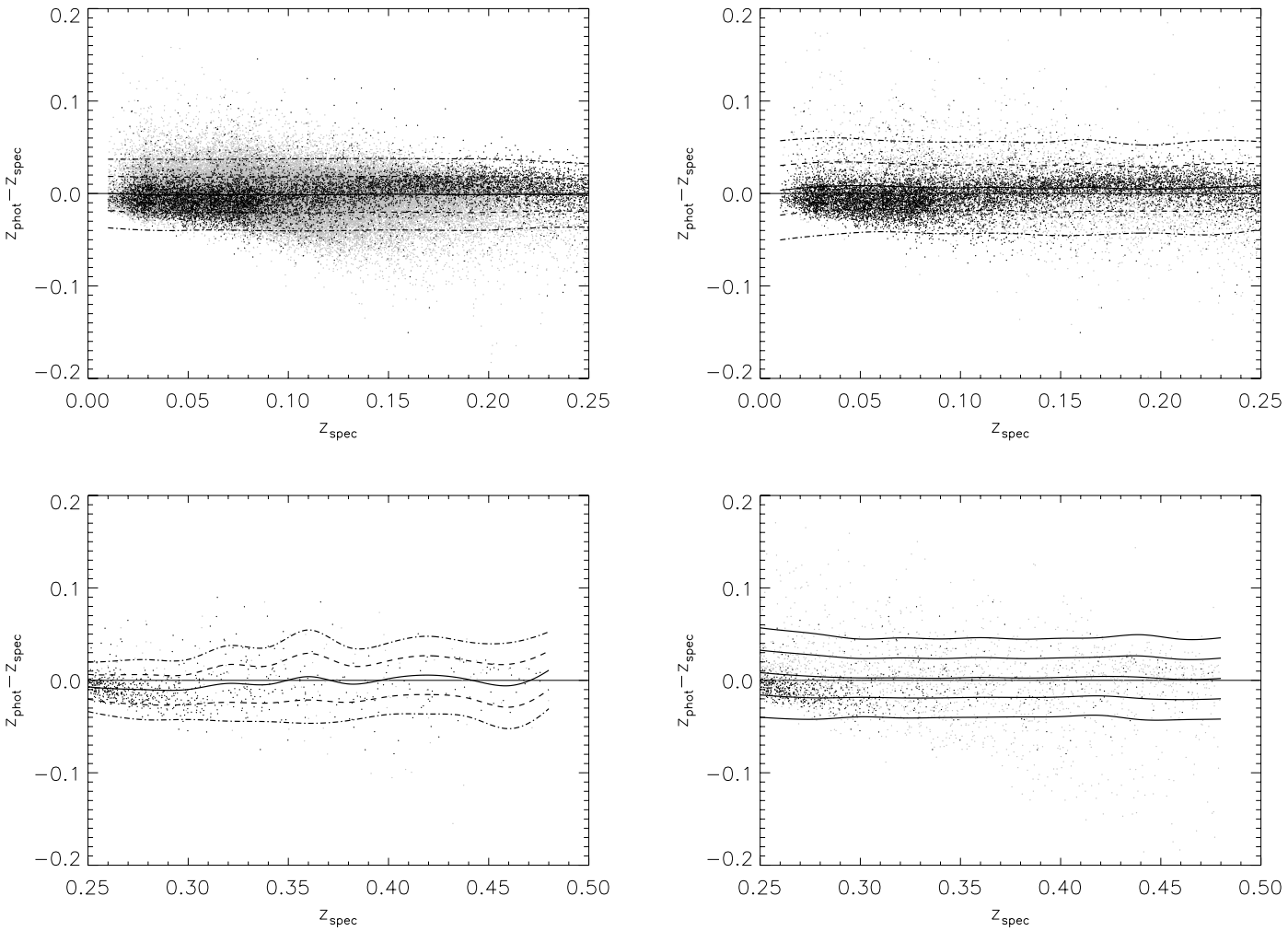


FIG. 11.—Distribution of residuals for the GG sample divided in magnitude bins. *Top left*, Nearby sample, $r < 17.7$; *top right*, nearby sample, $r > 17.7$; *bottom left*, distant sample, $r < 17.7$; bottom right, $r > 17.7$. [*See the electronic edition of the Journal for a color version of this figure*.]

In Figure 13 we show the same plot as in Figure 6 but without isocontours and plotting as black 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 0.020. Note, however, that the LRG objects show a clear residual systematic trend. This behavior 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 robust variances instead, given by $\sigma_3 = 0.321$ and 0.021. Also, in this case the observed behavior can be easily explained as being due to the heavy bias toward the LRGs which form ~88.5% of the sample. It must be stressed that while the remaining 11.4% of the objects still constitute a fairly large sample, 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 was already found by several authors (Padmanabhan et al. 2005), the derivation of photometric redshifts requires not only an accurate evaluation of the errors, but also the identification of a homogeneous sample of objects.

Objects not matching the 3 $\sigma$ 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 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 & Srivastava (2006). In the former various methods were tested against the EDR data. With reference to their Table 3, and using the "iterated" $\sigma$ 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 $\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 (~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 & 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 in contrast with our method, their methods greatly benefit by the use of

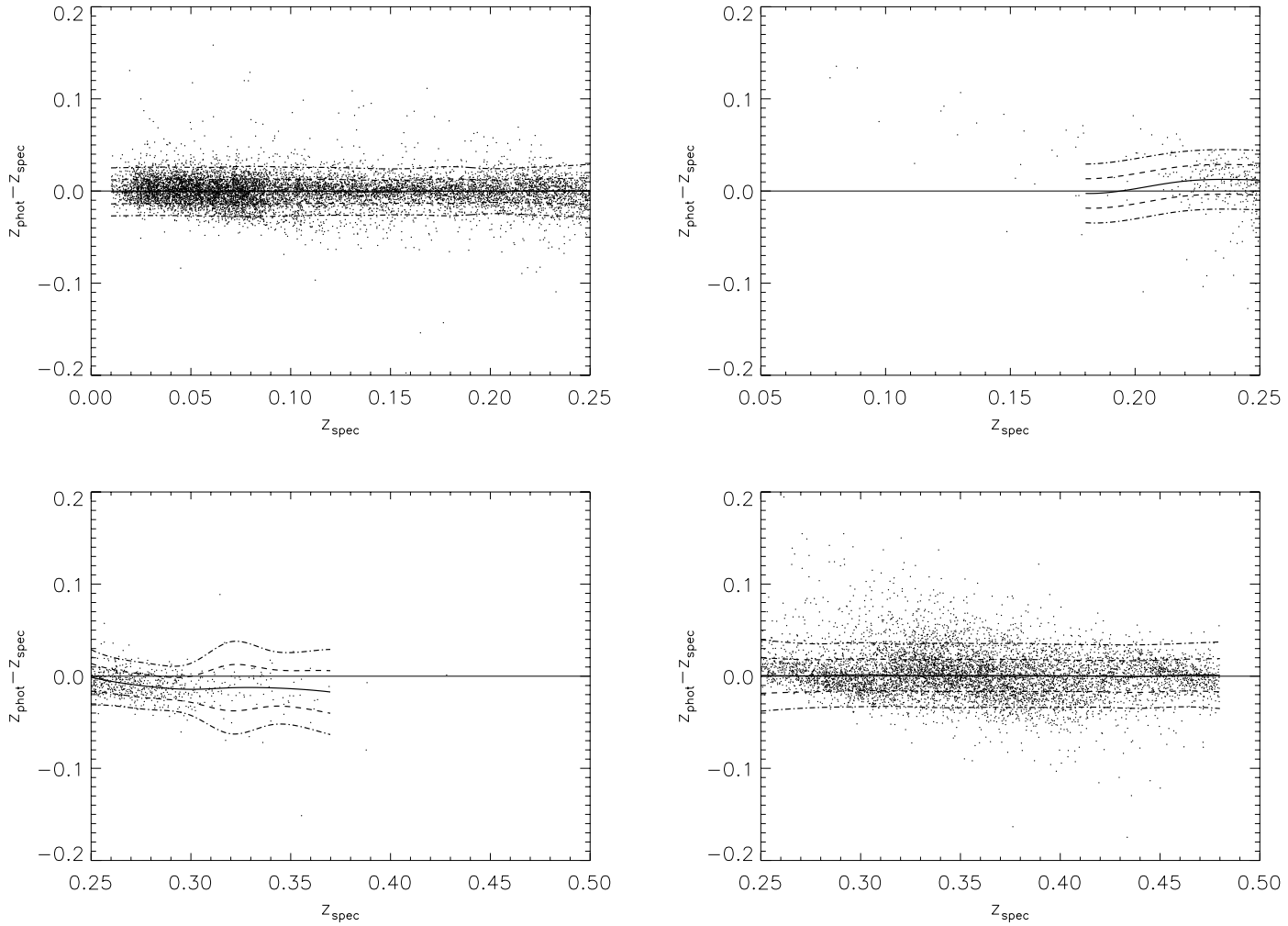


FIG. 12.—Distribution of residuals for the LRG sample divided in magnitude bins. Top left, Nearby sample, $r < 17.7$; top right, nearby sample, $r > 17.7$; bottom left, distant sample, $r < 17.7$; bottom right, $r > 17.7$. [*See the electronic edition of the Journal for a color version of this figure*.]

additional parameters such as the Petrosian radii, the concentration index, and the shape parameter.

Two points are worth stressing. First of all, their selection criteria for the construction 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 an "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 & 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, which had a large number of training data and few input features, the NN result is very stable, and therefore other combining procedures, such as AdaBoost (Freund & Schapire 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.

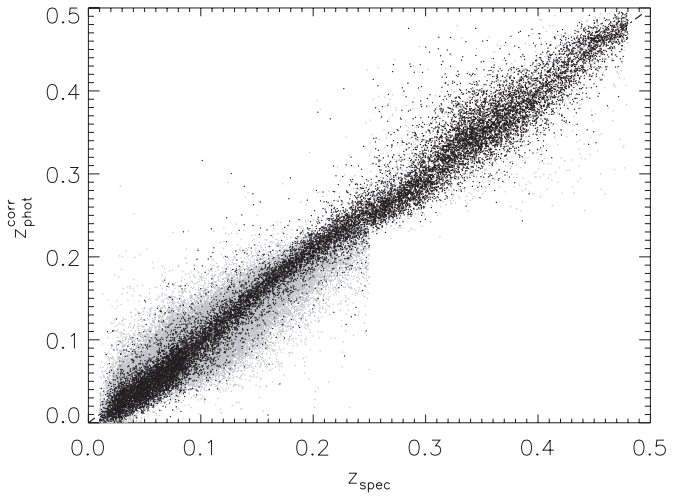


FIG. 13.—Plot of the same data shown in the lower panel of Fig. 6, with the LRG and GG objects marked as black and gray dots, respectively.

An additional machine learning approach, namely support vector machines, was used by Wadadekar (2005). In Table 4 we briefly summarize the main results of the above-quoted papers.

### 7.1. *Contamination by Distant Galaxies*

The fact that our NNs are trained on a sample of galaxies with observed redshift $z_{spec} < 0.5$ introduces some contamination from objects which, even though they are 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 covering, in a uniform way, all significant

TABLE 4
COMPARISONS OF VARIOUS METHODS FOR THE PHOTOMETRIC REDSHIFT ESTIMATION APPLIED TO THE SDSS DATA

| Method (1) | Data (2) | $\Delta z$ (3) | $\sigma$ (4) | Range (5) | Reference (6) |
|---|---|---|---|---|---|
| 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 | ... | Csabai et al. (2003) |
| Class X | DR2 | ... | 0.0340 | ... | Suchkov et al. (2005) |
| Gaussian Process | DR3 | ... | 0.0230 | ... | Way & Srivastava (2006)[a] |
| Ensemble | DR3 | ... | 0.0205 | ... | Way & Srivastava (2006)[a] |
| ANNz | EDR | ... | 0.0229 | ... | Collister & Lahav (2004) |
| SVM | DR2 | ... | 0.027 | ... | Wadadekar (2005) |
| SVM | DR2 | ... | 0.024 | ... | Wadadekar (2005)[a] |
| MLP ff | DR1 | 0.016 | 0.022 | $<0.4$ | Vanzella et al. (2004) |
| Template fitting and hybrid | DR1-LRG | $<0.01$ | $\sim0.035$ | $<0.55$ | Padmanabhan et al. (2005) |
| MLP | DR5-GG | $-0.0036$ | 0.0197 | 0.01, 0.25 | This work before interpolation |
| | DR5-GG | $-0.0036$ | 0.0245 | 0.25, 0.48 | This work before interpolation |
| | DR5-GG | ... | 0.0197 | 0.01, 0.25 | This work after interpolation |
| | DR5-GG | ... | 0.0238 | 0.25, 0.48 | This work after interpolation |
| | DR5-LRG | $-0.0029$ | 0.0154 | 0.01, 0.25 | This work before interpolation |
| | DR5-LRG | $-0.0029$ | 0.0189 | 0.25, 0.48 | This work before interpolation |
| | DR5-LRG | $-0.0029$ | 0.0160 | 0.01, 0.25 | This work after interpolation |
| | DR5-LRG | $-0.0029$ | 0.0183 | 0.25, 0.48 | This work after interpolation |

NOTES.—Col. (1): Method (for the acronyms, see text). Col. (2): Data set (EDR=Early Data Release; DR1 through DR5 the various SDSS data releases). Col. (3): Systematic offset. Col. (4): Standard deviation. Col. (5): Redshift range over which the average error is estimated. Col. (6): References.

[a] Additional morphological and photometric parameters.

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 levels by including statistically significant and complete samples of spectroscopic redshifts from other, deeper surveys. The feasibility of using a third NN to classify (and eventually dispose of) objects having $z > 0.5$ is under study. At the moment, however, since we are interested in validating the method and in producing catalogs to be used for statistical applications, we shall estimate the number and distribution in magnitude of such contaminants on statistical grounds using only the $r$-band luminosity function derived from SDSS data by Blanton et al. (2003). This function, in fact, allows us to derive for any given absolute magnitude the number of objects which, even though they are 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 Figure 14 which corresponds to a total number of contaminants of $\sim 3.74 \times 10^6$. It has to be noted, however, that for magnitudes brighter than 20.5, the fraction of contaminants is less than 0.04 and drops below 0.01 for $r < 20$.

## 8. THE CATALOGS

The catalogs containing the photometric redshift parameters together with the parameters used for their derivation can be downloaded at http://people.na.infn.it/~astroneural/SDSSredshifts.htm. This data, for consistency with the SDSS survey, has been subdivided into several files, each corresponding to a different SDSS "stripe" of the observed sky. A "stripe" is defined by a line of constant survey latitude $\eta$, bounded on the north and south by the edges of the two strips (scans along a constant $\eta$ 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.[8] The data for both GG and LRG samples have been extracted using the queries described in § 4. The catalogs 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 detail (SDSS database names of the parameters are in brackets): unique SDSS identifier (objID),

[8] For more details see http://www.sdss.org.

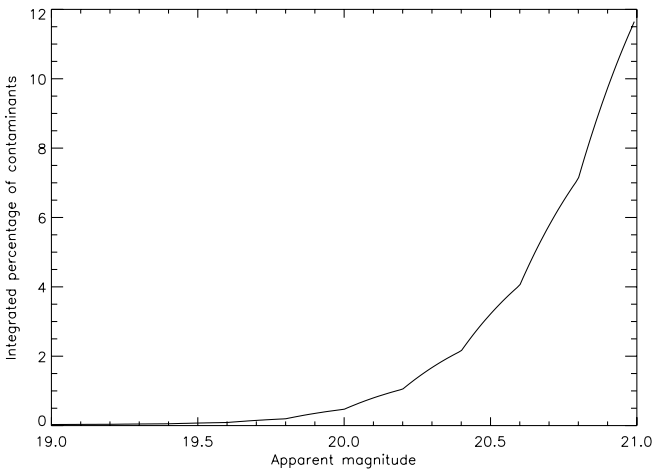


FIG. 14.—Estimated distribution of contaminants as a function of the apparent $r$ magnitude. The $y$-axis gives the expected fraction of objects at $z > 0.5$ that are erroneously evaluated by our procedure.

right ascension J2000.0 (R.A.), declination J2000.0 (decl.), dereddened magnitudes (`dered_u`, `dered_g`, `dered_r`, `dered_i`, `dered_z`), and the estimated value of photometric redshift before correction (`zphot`) and after correction (`zphot_corr`).

## 9. CONCLUSIONS

In the previous sections, we discussed a two-step 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 retrained 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 hand, the method suffers from 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 us 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 0.0238 for the nearby and distant sample, respectively) and $\sigma_3 = 0.0164$ for the LRG sample ($\sigma_3 = 0.0160$ and 0.0183). This accuracy was reached using a two-step approach that permits the building of 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, i.e., 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 the absence of a priori information and therefore to rely only on the photometric data themselves. This task, as it has been shown for instance by Suchkov et al. (2005) and Bazell & Miller (2004) is a nontrivial 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 buildup a of a suitable base of knowledge. A possible way to solve this problem will be discussed in Paper III.

This work was supported by the European Virtual Observatory Technological (VO-Tech) Infrastructure project, the MIUR-PRIN program, and by the SCOPE Consortium. The authors also thank the Regione Campania for partial financial Support. We thank Gennaro Miele and Giampiero Mangano for useful discussions.

## REFERENCES

Adelman-McCarthy, J. K., et al. 2006, ApJS, 162, 38
Baum, W. A. 1962, in IAU Symp. 15, Problems of Extra-Galactic Research, ed. G. C. McVittie (New York: Macmillan Press), 390
Bazell, D., & Miller, D. J. 2005, ApJ, 618, 723
Benítez, N. 2000, ApJ, 536, 571
Bernstein, G., & Jain, B. 2004, ApJ, 600, 17
Bishop, C. M. 1995, Neural Networks for Pattern Recognition (New York: Oxford Univ. Press)
Blake, C., & Bridle, S. 2005, MNRAS, 363, 1329
Blanton, M. R., et al. 2003, ApJ, 592, 819
Bolzonella, M., Miralles, J.-M., & Pelló, R. 2000, A&A, 363, 476
Bolzonella, M., Pelló, R., & Maccagni, D. 2002, A&A, 395, 443
Brodwin, M., et al. 2006, ApJ, 651, 791
Brunner, R. J., Connolly, A. J., & Szalay, A. S. 1999, ApJ, 516, 563
Bruzual A., G., & Charlot, S. 1993, ApJ, 405, 538
Budavári, T., et al. 2003, ApJ, 595, 59
Butchins, S. A. 1981, A&A, 97, 407
Coleman, G. D., Wu., C.-C., & Weedman, D. W. 1980, ApJS, 43, 393
Collister, A. A., & Lahav, O. 2004, PASP, 116, 345
Connolly, A. J., Csabai, I., Szalay, A. S., Koo, D. C., Kron, R. G., & Munn, J. A. 1995, AJ, 110, 2655
Csabai, I., et al. 2003, AJ, 125, 580
Dietterich, T. G. 2002, in The Handbook of Brain Theory and Neural Networks, ed. M. A. Arbib (2nd ed.; Cambridge: MIT Press), 405
Donalek, C. 2007, Ph.D. thesis, University of Napoli (Federico)
Duda, R. O., Hart, P. E., 1973, Pattern Classification and Scene Analysis (New York: Wiley)
Edmondson, E. M., Miller, L., & Wolf, C. 2006, MNRAS, 371, 1693
Eisenstein, D. J., et al. 2001, AJ, 122, 2267
———. 2003, ApJ, 585, 694
Fernández-Soto, A., Lanzetta, K. M., Chen, H.-W., Levine, B., & Yahata, N. 2002, MNRAS, 330, 889
Fernández-Soto, A., Lanzetta, K. M., Chen, H.-W., Pascarelle, S. M., & Yahata, N. 2001, ApJS, 135, 41
Firth, A. E., Lahav, O., & Somerville, R. S. 2003, MNRAS, 339, 1195
Freund, Y., & Schapire, R. E. 1996, in Proc. 13th Int. Conf. Machine Learning, ed. L. Saitta (San Francisco: Morgan Kaufmann), 148
Gladders, M. D., & Yee, H. K. C. 2000, AJ, 120, 2148
Hamilton, D. 1985, ApJ, 297, 371
Haykin, S. 1999, Neural Networks: A Comprehensive Foundation (2nd ed.; Upper Saddle River: Prentice Hall)
Huterer, D., Kim, A., Krauss, L. M., & Broderick, T. 2004, ApJ, 615, 595
Huterer, D., Takada, M., Bernstein, G., & Jain, B. 2006, MNRAS, 366, 101
Ishak, M. 2005, MNRAS, 363, 469
Koo, D. C. 1999, in ASP Conf. Ser. 191, Photometric Redshifts and the Detection of High Redshift Galaxies, ed. R. Weymann, et al. (San Francisco: ASP), 3
Ma, Z., Hu, W., & Huterer, D. 2006, ApJ, 636, 21
Massarotti, M., Iovino, A., & Buzzoni, A. 2001a, A&A, 368, 74
Massarotti, M., Iovino, A., Buzzoni, A., & Valls-Gabaud, D. 2001b, A&A, 380, 425
Padmanabhan, N., et al. 2005, MNRAS, 359, 237
Schneider, D. P., Gunn, J. E., & Hoessel, J. G. 1983, ApJ, 264, 337
Schneider, M., Knox, L., Zhan, H., & Connolly, A. 2006, ApJ, 651, 14
Suchkov, A. A., Hanish, R. J., & Margon, B. 2005, AJ, 130, 2439
Tagliaferri, R., Longo, G., Andreon, S., Capozziello, S., Donalek, C., & Giordano, G. 2003a, Lecture Notes in Computer Science, 2859, 226
Tagliaferri, R., et al. 2002, preprint (astro-ph/0203445)
———. 2003b, in Neural Network Analysis of Complex Scientific Data: Astronomy and Geosciences, ed. R. Tagliaferri (Amsterdam: Pergamon), 297
Tegmark, M., et al. 2006, Phys. Rev. D, 74, 123507
Vanzella, E., et al. 2004, A&A, 423, 761
Wadadekar, Y. 2005, PASP, 117, 79
Way, M. J., & Srivastava, A. N. 2006, ApJ, 647, 102
Willis, J. P., Hewett, P. C., & Warren, S. J. 2001, MNRAS, 325, 1002
Winter, C., Jeffery, C. S., & Drilling, J. S. 2004, Ap&SS, 291, 375
Yasuda, N., et al. 2001, AJ, 122, 1104
Zhan, H. 2006, J. Cosmol. Astropart. Phys., 08, 008
Zhan, H., & Knox, L. 2006, ApJ, 644, 663