THE APPLICATION OF PHOTOMETRIC REDSHIFTS TO THE SDSS EARLY DATA RELEASE

István Csabai,[1,2] Tamás Budavári,[2] Andrew J. Connolly,[3] Alexander S. Szalay,[1,2] Zsuzsanna Győry,[1]
Narciso Benítez,[2] Jim Annis,[4] Jon Brinkmann,[5] Daniel Eisenstein,[6] Masataka Fukugita,[7] Jim Gunn,[8]
Stephen Kent,[4] Robert Lupton,[8] Robert C. Nichol,[9] and Chris Stoughton[4]

ABSTRACT

The Early Data Release (EDR) from the Sloan Digital Sky Survey provides one of the largest multicolor photometric catalogs currently available to the astronomical community. In this paper we present the first application of photometric redshifts to the ~6 million extended sources in these data (with 1.8 million sources having $r' < 21$). Utilizing a range of photometric redshift techniques, from empirical to template and hybrid techniques, we investigate the statistical and systematic uncertainties present in the redshift estimates for the EDR data. For $r' < 21$, we find that the redshift estimates provide realistic redshift histograms with an rms uncertainty in the photometric redshift relation of 0.035 at $r' < 18$ and rising to 0.1 at $r' < 21$. We conclude by describing how these photometric redshifts and derived quantities, such as spectral type, rest-frame colors, and absolute magnitudes, are stored in the SDSS database. We provide sample queries for searching on photometric redshifts and list the current caveats and issues that should be understood before using these photometric redshifts in statistical analyses of the SDSS galaxies.

Key words: galaxies: distances and redshifts — galaxies: photometry — methods: statistical

1. INTRODUCTION

From their inception (Koo 1985; Connolly et al. 1995; Gwyn & Hartwick 1996; Sawicki, Lin, & Yee 1997; Hogg et al. 1998; Wang, Bahcall, & Turner 1998; Fernández-Soto, Lanzetta, & Yahill 1999; Benítez 2000; Csabai et al. 2000; Budavári et al. 2000) photometric redshifts have been seen as an efficient and effective means of studying the statistical properties of galaxies and their evolution. They are essentially a mechanism for inverting a set of observable parameters (e.g., colors) into estimates of the physical properties of galaxies (e.g., redshift, type, and luminosity). To date, photometric redshifts have typically been employed on small multicolor photometric surveys such as the Hubble Deep Field (HDF, Williams et al. 1996). While these applications have demonstrated the power of the estimated redshifts in studying galaxy evolution, they have an underlying limitation. The cosmological volumes probed by the narrow pencil-beam surveys are small, and consequently it is not clear whether these data provide a representative sample of the universe. With the development of large wide-field survey cameras, this volume limitation can be overcome and large statistically complete studies of the properties of galaxies can be undertaken.

One of the largest ongoing multicolor photometric surveys currently underway is the Sloan Digital Sky Survey (SDSS, York et al. 2000). This imaging and spectroscopic survey provides an ideal base from which to apply photometric redshifts to large samples of galaxies. In the Early Data Release (EDR, Stoughton et al. 2002) there are over 6 million galaxies, an order-of-magnitude increase in sample size when compared with existing public multicolor surveys. From these galaxies there are approximately 35,000 galaxies with published spectroscopic redshifts from which to determine the statistical and systematic uncertainties in the SDSS photometric redshift relation.

In this paper we describe the first application of photometric redshifts to the SDSS data. We provide a background to the redshift estimation techniques but do not go into the technical details of the individual methods. We focus on providing the astronomical community with details of how to use the photometric redshifts in the SDSS EDR database and emphasize the caveats and limitations present in the current photometric redshift catalog (due to photometric errors and uncertainties in the SDSS zero points). We plan to have a more detailed analysis of systematic errors on the soon-to-appear Data Release 1, where most of these problems will be eliminated. Sample queries for the EDR database are provided in § 6.1, together with details of value added parameters that can be derived from the photometric redshifts, such as rest-frame colors, *k*-corrections, and absolute magnitudes.

2. THE EARLY SDSS DATA RELEASE

In this section we provide a brief description of the Early Data Release (Stoughton et al. 2002) of the SDSS and introduce the subsets of the data that will be used throughout this paper. The EDR has five-band photometry (Fukugita et al. 1996; Gunn et al. 1998; Smith et al. 2002; Hogg et al. 2001;

[1] Department of Physics, Eötvös Loránd University, Budapest, Pf. 32, H-1518 Budapest, Hungary.

[2] Department of Physics and Astronomy, Johns Hopkins University, 3701 San Martin Drive, Baltimore, MD 21218-2868.

[3] Department of Physics and Astronomy, University of Pittsburgh, 3941 O'Hara Street, Pittsburgh, PA 15260.

[4] Fermi National Accelerator Laboratory, P.O. Box 500, Batavia, IL 60510.

[5] Apache Point Observatory, P.O. Box 59, Sunspot, NM 88349.

[6] Steward Observatory, University of Arizona, 933 North Cherry Avenue, Tucson, AZ 85721.

[7] Institute for Cosmic Ray Research, University of Tokyo 3-2-1, Midori, Tanashi, Tokyo 188-8502, Japan.

[8] Princeton University Observatory, Peyton Hall, Princeton, NJ 08544-1001.

[9] Department of Physics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15232.

Pier et al. 2002) for over 6 million galaxies, out of which 1.8 million galaxies have $r' < 21$. The five filters of the u', g', r', i', z' system have effective wavelengths of 3540, 4750, 6222, 7632, and 9049 Å, respectively, and the goal of the survey is to achieve a level of photometric uniformity and accuracy such that the systemwide rms errors in the final SDSS photometric catalog will be less than 0.02 mag in r', 0.02 mag in $r'-i'$ and $g'-r'$, and 0.03 mag in $u'-g'$ and $i'-z'$, for objects bluer than an M0 dwarf. All analyses in this paper are based on the dereddened model magnitudes in the EDR data set. A relatively small subset of these galaxies ($>30,000$) have measured redshifts. The objects for spectroscopic observation were selected using the SDSS target-selection algorithm, which is discussed in detail in Stoughton et al. (2002) and Strauss et al. (2002). This selection algorithm results in two subsets of the SDSS data, a main galaxy sample and a luminous red galaxy (LRG) sample (Eisenstein et al. 2001). The main galaxy sample contains 27,797 galaxies, with a mean redshift of $z = 0.116$ and a photometric limit of $r' = 18$. The LRG sample was selected from galaxies with colors similar to that of an elliptical galaxy and contains 6698 galaxies, with a mean redshift of $z = 0.227$ (though extending out to $z > 0.5$). The redshift histograms of these two subsets of the data are given in Figure 1, and they demonstrate that the main sample should provide a good training/test set out to $z = 0.2$, and the LRG data set out to $z = 0.5$.

In order to test the accuracy of the photometric redshifts derived from the SDSS, we supplement the SDSS redshifts with a subset of galaxies selected from published redshift catalogs. At low redshift and for bright magnitudes, the 2dF Galaxy Redshift Survey (Colless et al. 2001) contains 5642 galaxies for which we have matching SDSS photometry. These galaxies have a limiting magnitude of approximately $r' = 18.5$ and a mean redshift of $z = 0.112$. The redshift range sampled by these galaxies is therefore well matched to

that of the SDSS redshift catalog, with a limiting redshift of approximately $z = 0.2$. At higher redshifts and for fainter magnitudes the Canada Network for Observational Cosmology (CNOC2) survey (Yee, Ellingson, & Carlberg 1996) has a magnitude limit of approximately $r' < 21.0$, with a mean redshift of $z = 0.274$ and upper redshift limits of approximately $z = 0.7$. The photometric depth of the 2697 galaxies in the CNOC2 sample provides not just a test of the accuracy of the photometric redshifts but also a measure of how the redshift uncertainties scale with magnitude limit. We designate these "blind" test samples as "2dF" for the low-redshift samples and "CNOC2" for the CNOC2 data.

In the following sections we will use the main EDR and the EDR LRG samples as training sets and all of the above data sets as test sets.

3. STANDARD PHOTOMETRIC REDSHIFT TECHNIQUES

A wide range of techniques has been employed in the literature to estimate redshifts of galaxies with broadband photometric colors. Approaches have ranged from the purely empirical relations to comparisons of the colors of galaxies with the colors predicted from galaxy spectral energy distributions. Each approach has its own set of advantages and disadvantages. Empirical approaches, where the color-redshift relations are derived directly from the data themselves, are relatively free from possible systematic effects in the photometric calibration. As such, they provide a simple measure of the statistical uncertainties with the data and can demonstrate the accuracy to which we should be able to estimate redshifts once we can control the systematic errors. Their underlying disadvantage is that we can typically only apply these relations to galaxies with colors that lie in the range of colors and redshifts found in the training set. Template-based techniques are free from the limitation of a training set and can be applied over a wide range of redshifts and intrinsic colors. They rely, however, on having a set of galaxy templates that accurately map the true distribution of galaxy spectral energy distributions (and their evolution with redshift) and on the assumption that the photometric calibration of the data is free from systematics.

In this section we consider both empirical and template-based approaches to photometric redshift estimation for SDSS data. We demonstrate the redshift accuracy that it should be possible to achieve from the EDR sample and describe the current limitations of using standard galaxy spectral energy distributions.

3.1. *Empirical Redshift Estimation Methods*

We consider here the standard empirical redshift estimation techniques that have been used in the literature (Connolly et al. 1995; Wang et al. 1998; Brunner, Connolly, & Szalay 1999) and develop a new technique based on a hierarchical indexing structures (kd-trees, Moore, Schneider, & Deng 1997). One of the first successful empirical methods is based on fitting a functional form for the relation between the spectroscopic redshift of a galaxy and its colors or magnitudes (Connolly et al. 1995). This function is typically a second- or third-order polynomial. Figure 2 shows the photometric versus the spectroscopic redshifts, using the EDR main galaxy and LRG spectroscopic samples. As the size of the training set is large ($>30,000$) when compared with the
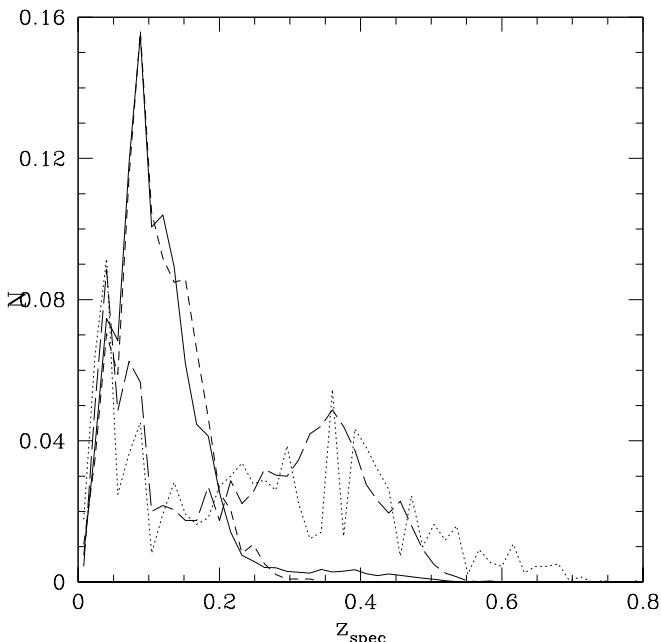

Fig. 1.—Spectroscopic redshift histogram for the SDSS main EDR (*solid line*), the EDR LRG (*long dashed line*), the 2dF (*short dashed line*), and the CNOC2 sets.
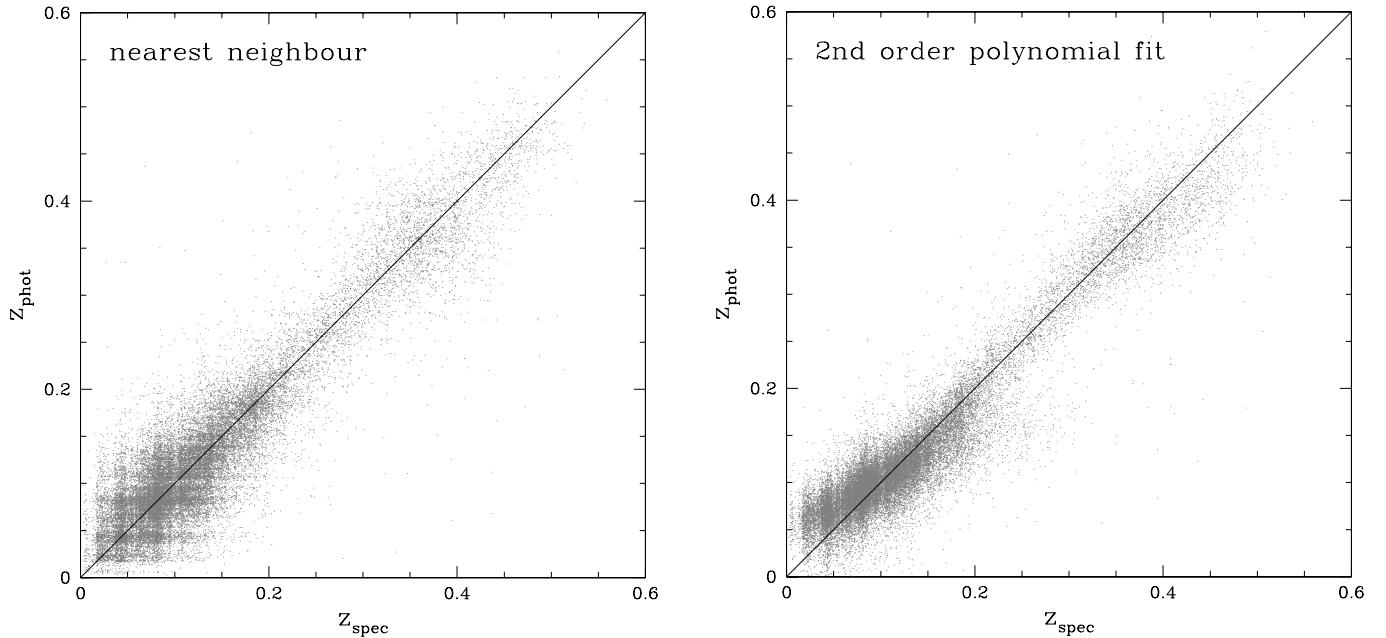
FIG. 2.—Photometric redshift estimations with the simple empirical methods

number of the fitted parameters (21), we can expect that this fit will work for other objects with the same dispersion, as seen in Figure 2 (as long as the data are selected over the same sample color and redshift range as the training set). The dispersion in this photometric redshift relation is $\sigma_z = 0.027$ (see Table 1 for comparison with other values). One possible uncertainty in this technique comes from the fact that the fitting function is just an approximation of the (possibly) more complex relation between the colors and the redshift of a galaxy. We would therefore expect the fitting function to accurately follow the redshift-color relation over a narrow range of redshift. A technique to avoid this is to use separate functions in different redshift (Brunner et al. 1999) or color ranges.

A second (and possibly the simplest) empirical estimator is the nearest-neighbor method. For a test galaxy, this finds the galaxy in the training set with the smallest distance in the color (or magnitude) space (weighted by the errors). The redshift of this closest match is then assigned to the test galaxy. In the ideal case the training set contains a sufficient

TABLE 1
ERRORS ON PHOTOMETRIC REDSHIFTS

Estimation Method	rms	log	Iterated	Nonoutliers (%)
Polynomial	0.0318	0.0277	0.0273	98.0
Nearest neighbor	0.0365	0.0321	0.0327	98.5
Kd-tree	0.0254	0.0224	0.0226	98.4
CWW	0.0666	0.0598	0.0621	99.1
Bruzual-Charlot	0.0552	0.0501	0.0509	99.2
Bayesian	0.0476	0.0415	0.0422	98.4
CWW LRG	0.0473	0.0332	0.0306	97.1
Repaired LRG	0.0476	0.0319	0.0289	96.5
Interpolated	0.0451	0.0359	0.0352	97.7
2dF	0.0528	0.0455	0.0433	97.1
CNOC2	0.1358	0.0989	0.0842	93.0
CNOC2 $17.8 < r < 19.5$	0.0801	0.0614	0.0614	97.1

NOTES.—We list three different estimated rms values in the table. The first is the usual standard deviation σ_{rms} computed for all galaxies as defined by $\sigma_{rms}^2 = \langle \Delta z^2 \rangle$, where $\Delta z = z_{spec} - z_{phot}$. The standard deviation is very sensitive to outliers; it is a common trick to assign less weight to them by defining another quantity that measures the scatter in a more reliable way: $\sigma_{log}^2 = \langle A^2 \log(1 + \Delta z^2/A^2) \rangle$, where A is a large number compared with Δz. We use $A^2 = 20 \times \Delta z_{med}^2$, where Δz_{med} is the median. Without outliers, σ_{rms} and σ_{log} were basically same, because $\epsilon \approx \log(1 + \epsilon)$ for small ϵ values, but large outliers only affect the standard deviation dramatically. Another way of suppressing the effect of outliers is excluding them. The last rms column (σ_z; we use this values in the text) lists the standard deviation for galaxies that are within the 3 σ limits of the distribution, which often has a value similar to σ_{log}. The very last column of the table shows the fraction of galaxies included in the 3 σ limit.
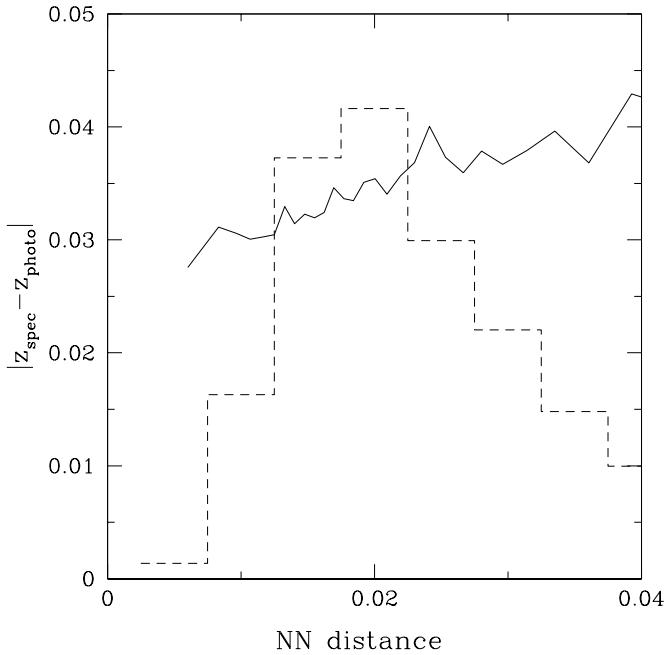
FIG. 3.—Dependence of redshift average estimation error on the color space distance from the nearest reference object (*solid line*). As expected, smaller distances result smaller error. The dashed line is for the histogram of number of objects with a given nearest neighbor distance. One can see that for most of the objects the nearest neighbor is not close enough to get the best estimation.

number of galaxies that for each unknown object there is a close neighbor. In Figure 3 we show that redshift estimation error increases with the distance from the nearest neighbor in color space. The larger the data set, the more accurate this method becomes, as long as all galaxy types are represented in the training set. From the technical viewpoint, larger training sets mean that the search time increases, so one has

to use an efficient multidimensional search technique (e.g., kd-trees) instead of a standard linear search. The comparison between the estimated and spectroscopic redshifts for the nearest-neighbor technique is given in Figure 2. The dispersion about this relation is $\sigma_z = 0.033$.

A natural limitation of the nearest-neighbor technique is that a large number of training galaxies alone is not enough; they must cover the range of the colors of the unknown objects in a more or less uniform way. Unfortunately, this is usually not the case. To resolve this problem, one can search for more than one nearest neighbor and apply an interpolation or a fitting function. This also helps resolve a second problem, namely, that, because of the finite number of objects in the training set, the photometric redshifts will have discrete values, making them problematic to use in some statistical studies. We have created a hybrid version of these two empirical methods: we partitioned the color space into cells containing the same number of objects from the training set, using a kd-tree tree (a binary search tree, Bluntly 1979). In each cell we fitted a second-order polynomial. The results, together with a demonstration of a two-dimensional version of the kd-tree partitioning of the EDR training set, are given in Figure 4. The dispersion about this relation is $\sigma_z = 0.023$.

For each of these approaches the resulting dispersion in the photometric redshift relation is found to be approximately 0.03 (see Table 1), with the hybrid method being marginally more accurate. As these empirical approaches do not rely on the absolute photometric calibration of the data (other than that the calibration should be stable across the data sets), they are somewhat insensitive to systematic errors in the data. If the SDSS redshifts (or external redshift samples) sampled the full redshift range of the data to the limit of the survey, these empirical techniques would provide an ideal mechanism for deriving redshift estimates for the SDSS. As the redshift range of the spectroscopic samples are fairly limited, the application of these techniques to
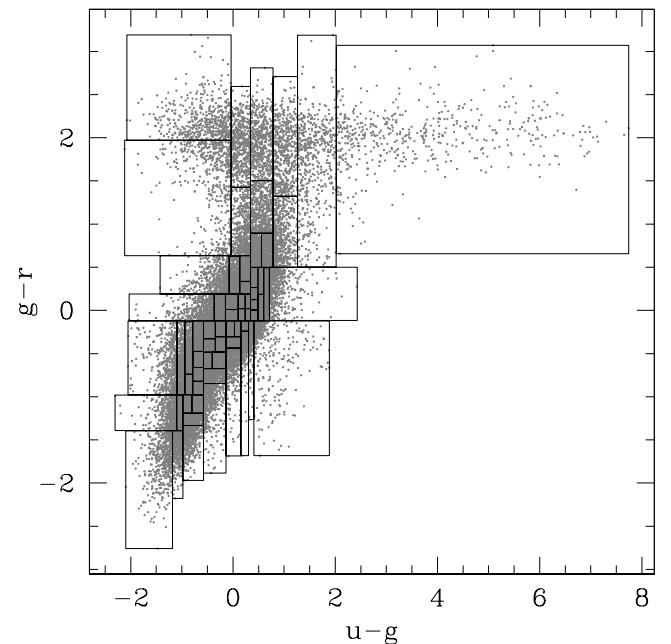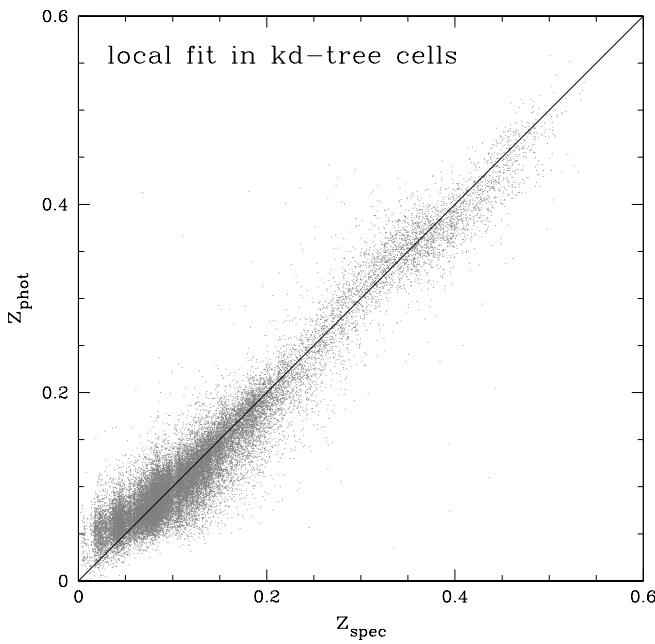

FIG. 4.—*Right*: Plot of two-dimensional demonstration of the color space partitioning. In each of these cells we applied the polynomial-fitting technique to estimate redshifts. *Left*: Results.

the full data set is nontrivial. We can, however, use these results to demonstrate that the accuracy we should be able to derive from the template-based techniques (once any systematics in the data are accounted for) should be $\sigma_z \sim 0.03$ at $r' < 18$.

3.2. Template-Based Redshift Estimation Methods

As noted previously, the advantages of using templates to estimate redshifts of galaxies (Koo 1985; Gwyn & Hartwick 1996; Sawicki, Lin, & Yee 1997; Connolly et al. 1999; Fernández-Soto et al. 1999; Benítez 2000; Bozonella, Miralles, & Pelló 2000; Budavári et al. 1999; Budavári et al. 2000; Csabai et al. 2000) are numerous. This approach simply compares the expected colors of a galaxy (derived from template spectral energy distributions) with those observed for an individual galaxy. The standard scenario for template fitting is to take a small number of spectral templates T (e.g., E, Sbc, Scd, and Irr galaxies) and choose the best fit by optimizing the likelihood of the fit as a function of redshift, type, and luminosity $p(z, T, L)$. Variations on this approach have been developed in the last few decades, including ones that use a continuous distribution of spectral templates, enabling the error function in redshift and type to be well defined.

A representative set of spectrophotometrically calibrated spectral templates is not easy to obtain. One problem with measured spectra is that to calibrate them spectrophotometrically over the full spectral range is nontrivial. A second problem is that, because of the redshift of a galaxy, we need spectra over a wavelength range that is wider than the range of our optical filters (3000–12000 Å). Such spectra cannot currently be measured by a single spectrograph. Third, even if we could measure calibrated spectra over the required range, spectrographs, especially modern multifiber ones, usually sample only the central region of the galaxy, while photometric measurements integrate over the full spatial extent of a galaxy. The alternative to empirical templates is to use the outputs of spectral synthesis models. The accuracy of spectral models is improving (Bruzual & Charlot 1993), but they are not yet as accurate as direct measurements of galaxy spectra. Modern surveys will improve on this situation (e.g., the SDSS will measure spectrophotometrically calibrated spectra for a million objects in the 3800–9200 Å range at a resolution $R = \lambda/\Delta\lambda$ of about 1800), but to date there does not exist an optimal set of galaxy spectral templates.

The most frequently used set of spectral energy distributions (SEDs) used in photometric redshift analyses are those from Coleman, Wu, & Weedman (1980, hereafter CWW; see also Bolzonella et al. 2000). In Figures 5 and 6 we demonstrate the results of the template-fitting technique using the CWW templates and a set of SEDs from the spectral synthesis models of Bruzual & Charlot (1993). The dispersion about this relation is 0.062 and 0.051 for the CWW and BC templates, respectively. While this is only a factor of 2 worse than that achieved by the empirical methods, there appear to be systematic deviations in these photometric redshift relations. The CWW templates produce a photometric redshift relation where the majority of galaxies have a systematically lower redshift than that given by the spectroscopic data (by approximately 0.03 in redshift), and there exists a broad tail of galaxies for which the photometric redshifts are systematically overestimated. For the BC templates the galaxy redshifts tend to be systematically
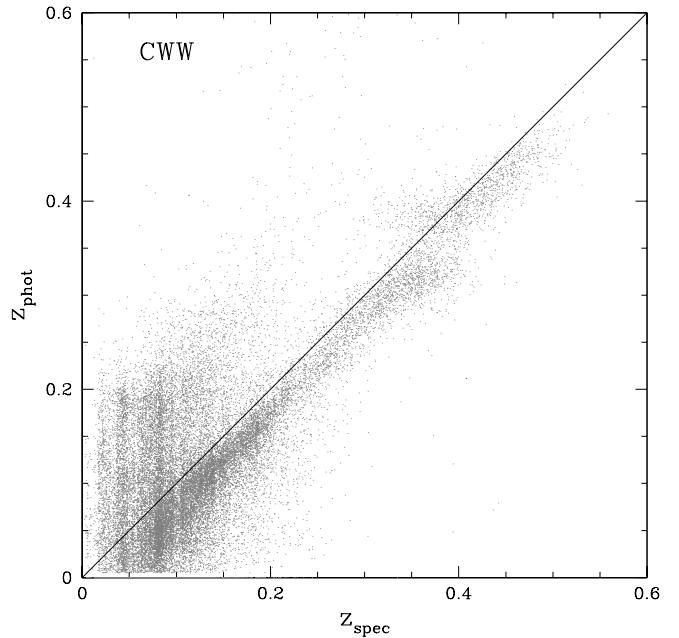

Fig. 5.—Photometric redshift estimation using the CWW spectral energy distributions. The rms dispersion about this relations is 0.062 in redshift. For the majority of the SDSS galaxies, the CWW templates perform reasonably well (with the core of the photometric redshift relation having a tight correlation with spectroscopic redshift, though about 0.03 below the one-to-one relation). It is clear, however, that there remains a population of galaxies for which the CWW templates do not well match the galaxy colors, leading to an overestimate of the redshift of the galaxy.

underestimated (with this effect becoming more pronounced as a function of redshift out to redshifts $z = 0.3$).

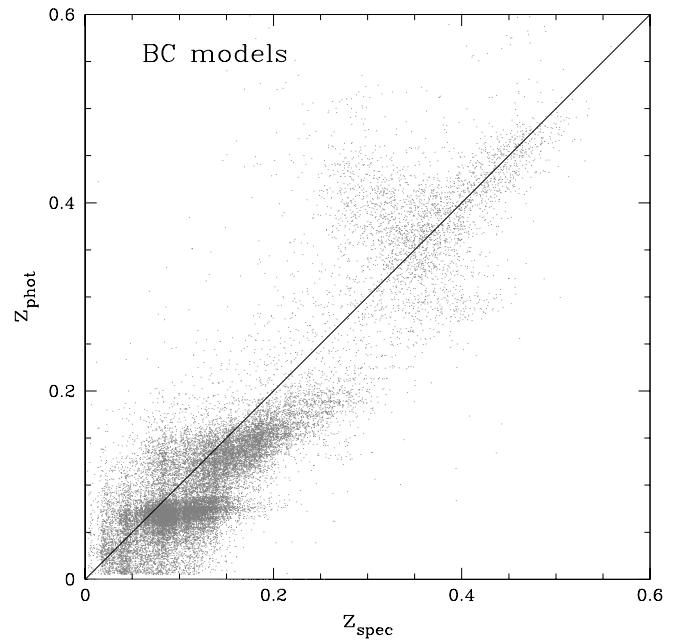An improvement over standard template methods, which rely uniquely on the galaxy colors, is the introduction of

Fig. 6.—Photometric redshift estimation using the Bruzual & Charlot spectral energy distributions. The rms dispersion about this relations is 0.051 in redshift. While this dispersion is within a factor of 2 of that derived from adapting the templates, it is also clear that there remain systematic offsets in the photometric redshift relation with the Bruzual & Charlot templates underpredicting (on average) the true spectroscopic redshift.
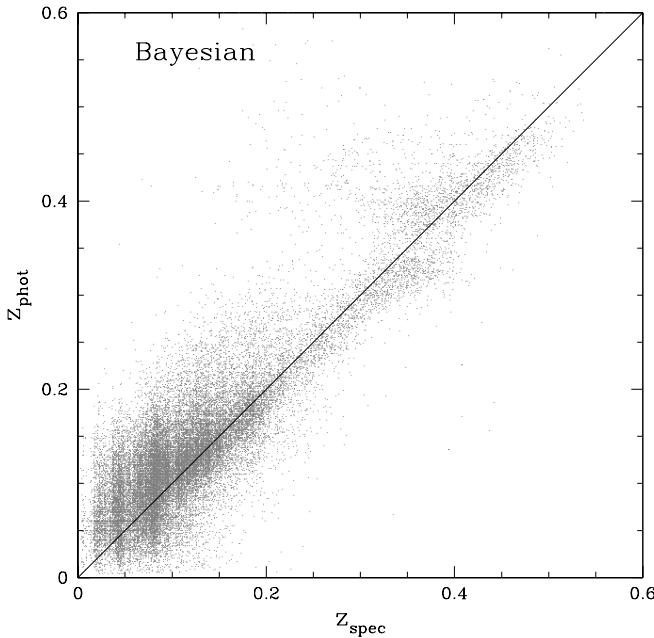
Fig. 7.—Photometric redshift estimation, using the Bayesian method. The rms dispersion about this relations is 0.042 in redshift.

magnitude priors in a Bayesian framework (Benítez 2000). The redshift distribution of the main EDR sample is well fitted by the relationship $p(z) \propto z^2 \exp{[-(z/z_m)^{1.5}]}$ for $i \lesssim 18$, and a continuous prior can be constructed by measuring z_m in five different magnitude bins and interpolating. Since the EDR spectroscopic sample redshift distribution is "contaminated" by LRGs at faint magnitudes and turns bimodal, we have assumed a flat redshift/magnitude prior for $i \gtrsim 18$. Using this magnitude prior we run Bayesian estimation with two further refinements: (1) setting the minimal photometric error in each band to 0.03, which mimics the intrinsic fluctuations in the colors of galaxies described by a same template and produces more realistic redshift likelihoods, and (2) using linear interpolation between the main CWW types to improve the color resolution. Using this setup, the dispersion for the CWW templates *without using any prior* decreases from 0.06 to 0.05, with an offset of 0.0156; introducing the prior described above further decreases the dispersion to $\sigma_z = 0.0415$ (see Fig. 7) for the whole sample, but an offset of 0.0144 still remains.

It is clear from these tests that, while the template-fitting methods should be directly applicable to the SDSS EDR data, there remain significant systematics in either the templates or the photometric calibrations (or both) that will add artifacts into any photometric redshift relation. We must therefore recalibrate the template spectra to minimize these systematic effects.

4. HYBRID PHOTOMETRIC REDSHIFT TECHNIQUES

Recently, new hybrid techniques have been developed to calibrate template SEDs (Csabai et al. 2000; Budavári et al. 1999; Budavári et al. 2000; Budavári et al. 2001a), using a training set of photometric data with spectroscopic redshifts. These combine the advantages of the empirical methods and SED fitting by iteratively improving the agreement between the photometric measurements and the spectral templates. The basic approach is to divide a set of galaxies

into a small number of spectral classes (using the standard template based photometric redshifts) and then adjust the template SEDs to match the mean colors of the galaxies in these spectral classes. By repeating this classification and repair procedure, the template spectra converge toward the observed colors. In this paper we will not review the details of these techniques, but direct the reader to Csabai et al. (2000) and Budavári et al. (1999; 2000, 2001a) for a full description of the algorithms. As we shall show in the following sections, the application of these techniques yields more reliable photometric redshifts for the SDSS EDR catalog than the standard template fitting.

4.1. A Single Template: The Luminous Red Galaxy Sample

In addition to providing a training set for redshift estimation in the SDSS data, the LRG sample is extremely useful in identifying systematic uncertainties in the SDSS photometric system. The LRG galaxies have a strong continuum feature, namely, the break at around 4000 Å. As a result of the depth of this feature, photometric redshifts are easily estimated for these galaxies. In addition, because of the high luminosity of these galaxies, they can be observed spectroscopically over a larger redshift range than the main galaxy sample. Systematics in the photometric data can therefore be identified as this spectral feature passes through the filters as a function of redshift. In fact, we can simply use a single SED for the LRG sample to test how we must optimize the template spectra to accurately represent the observed colors.

For the 6698 LRG galaxies we start with an initial template spectrum selected from the CWW elliptical spectrum and apply the training techniques of Budavári et al. (2000). In Figure 8 we show the original CWW elliptical spectrum, together with our reconstructed template. From these spectra we can see that, in order to represent the colors of the LRGs, we need a template spectrum that is redder than the standard CWW elliptical. To demonstrate how well these respective spectral templates cover the photometric observations, in Figure 9 we have plotted the colors of the EDR LRG galaxies together with the traces of the original and repaired spectral templates. The color-redshift relation for the repaired spectrum clearly traces the locus of the LRG galaxy sample more accurately than the original CWW SED. The most obvious improvement in the comparative colors is found in the $u'-g'$ and $i'-z'$ colors.

Although the repair procedure does not optimize directly for photometric redshifts, the improvement in the match between the observed and predicted colors should lead to an improved photometric redshift relation for the LRG
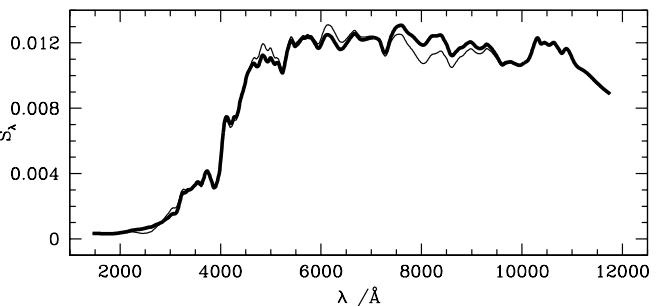

Fig. 8.—Repaired (*thick line*) spectral template is redder than the original elliptical galaxy template (*thin line*).
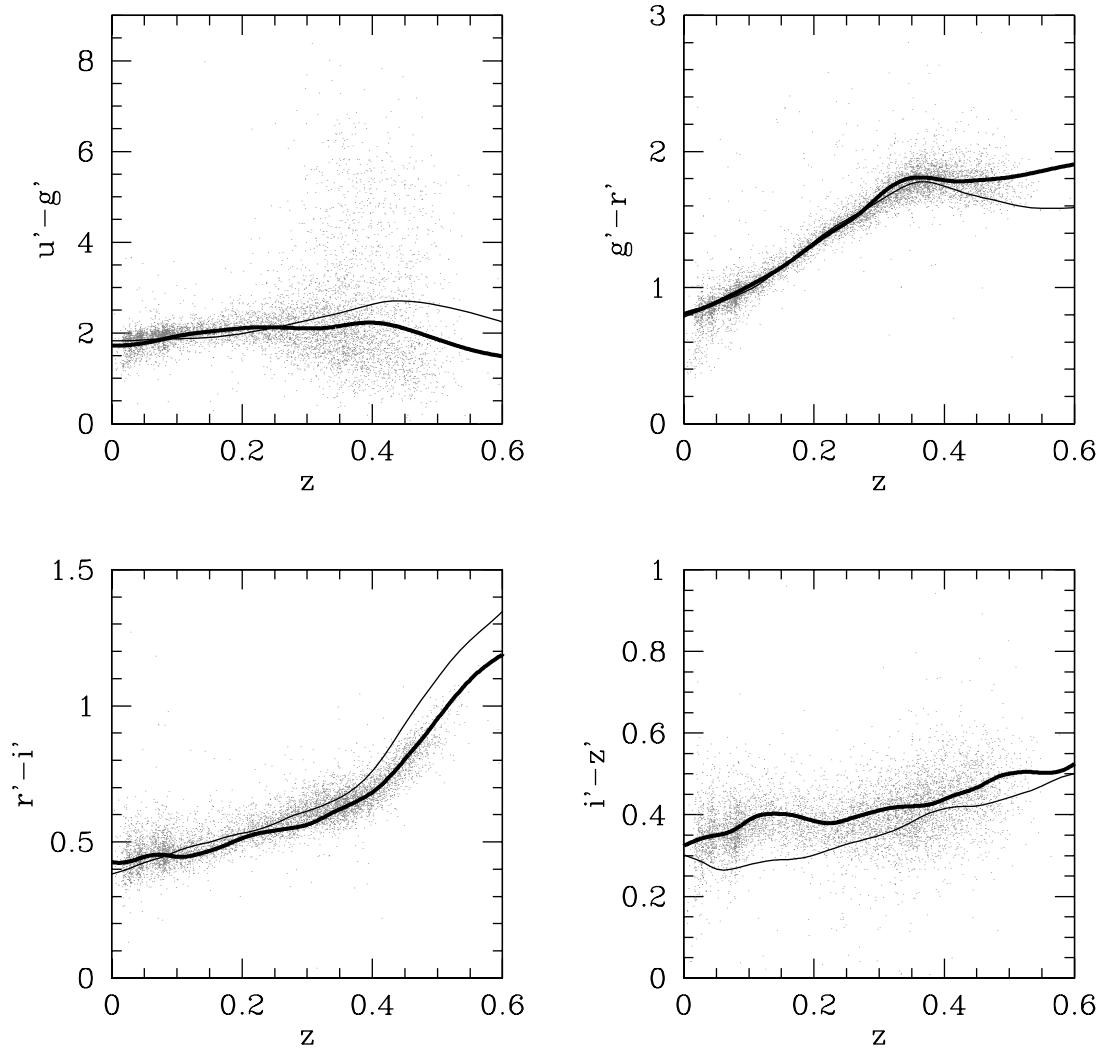
Fig. 9.—Four SDSS colors of ∼6000 red galaxies vs. the redshift. The color trace of the repaired spectral template (*thick line*) follows the data better than the trace of the original CWW E0 template (*thin line*).

sample. Figure 10 compares the performance of the photometric redshift estimators utilizing the two original and repaired template SEDs. The repair procedure decreases the overall scatter in the redshift relation from $\sigma_z = 0.031$ to $\sigma_z = 0.029$. The main improvement is, however, that the systematic underestimation of the redshift at redshifts $z > 0.2$ is reduced. There remains a feature in the redshift relation at $z \approx 0.4$, an increase by a factor of 2 in the dispersion. This arises as a result of a degeneracy in the $u'-g'$ versus $g'-r'$ colors in red galaxies at a redshift of $z \sim 0.4$ (the color-color tracks loop on top of each other). The degeneracy is a result of the Balmer break shifting between the g' and r' filters, making it difficult to estimate the exact redshift (Budavári et al. 2001b). This problem cannot be removed by using better template spectra.

4.2. The Distribution of Galaxy Types: The Main Galaxy Sample

The entire sample of the SDSS galaxies (including the LRGs) poses a more difficult question because of the spectral composition of the data. Spectral variations cannot be neglected, and, in fact, one would like to get a continuous parameterization of the spectral manifold. To accomplish this, we adopt a variant of the ASQ algorithm (Budavári et

al. 2001a). First we reconstruct a small number of discrete SEDs, using the techniques described previously, and then we use an interpolation scheme to provide a continuous distribution of spectral types that evenly sample between the discrete spectra.

The training set consists of all galaxies with spectroscopic redshifts and the five-band SDSS photometry. The large number of galaxies is very promising, but the spectral resolution of the reconstructed templates also depends on the redshift baseline of the input galaxy training set. This redshift range is significantly smaller than, for example, those derived from the Hubble Deep Field (Hogg et al. 1998; Budavári et al. 2000). Ideally, one would like to have a training set that uniformly samples the color space to ensure that no extra weight is assigned to any particular type of galaxy. The limited color range of the galaxies with spectroscopic redshifts will, therefore, ultimately limit the accuracy of our final redshift relations.

The iterative ASQ method was applied to the initial set of four CWW spectra. The spectral templates are found to converge rapidly, in a few iterations. After 10 iterations the repaired templates yield the photometric redshifts shown in the top panels of Figure 11. The left panel shows all galaxies assigned to the reddest template, and the galaxies assigned
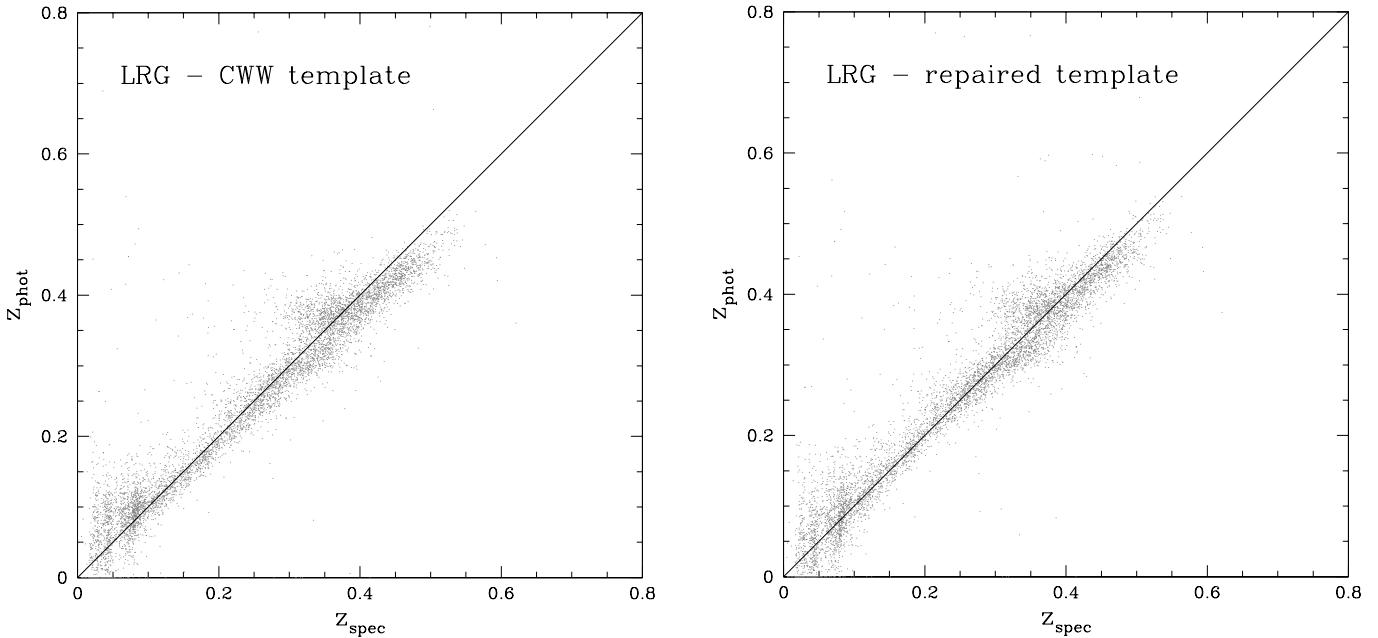
FIG. 10.—Photometric vs. spectroscopic redshifts for the EDR LRG set. *Left*: Original CWW spectral templates. *Right*: Repaired templates. One can see that the redshift prediction improves, especially for higher redshifts.

to the remaining three templates are given on the right panel. The rms in the red and blue sample are $\sigma_z = 0.028$ and 0.05, respectively. This plot should be compared with the redshift relations derived from the standard CWW templates, as shown in Figure 5. The training of these templates both removes the systematics in the data and reduces the dispersion about the photometric redshift relation.

The large estimation error for the late-type galaxies is partly caused by the small number of discrete templates used in the redshift estimation. We can improve on our estimates if we derive an interpolation scheme that provides a finer sampling of the distribution of late-type spectral templates. Figure 12 illustrates the one-dimensional continuous spectral manifold derived from the discrete SEDs by plotting equally spaced (in type) interpolated spectra using a simple spline interpolation. Based on the following tests, this simple interpolation scheme provides sufficient accuracy for mapping the color distribution of late-type galaxies.

The first test of the interpolation scheme was a simple sanity check of the type histogram. If the interpolated spectra are not physical, we expect to see humps at the basis templates (i.e., the colors of the majority of galaxies will be better matched to the original templates than the interpolated templates). For this test, we used the known redshift of each galaxy in the training set and only fitted the spectral type (and apparent luminosity). In Figure 13 we show this interpolated type histogram. The smooth transition between interpolated types shows no evidence for any discreteness in assigning a spectral template to an individual galaxy. The second test of the interpolation was to determine whether the interpolated templates would evolve if we applied the ASQ training algorithm. Fixing the four basis-trained SEDs, we introduced three interpolated classes at the center of the intervals between these spectral types. We find no significant change in the spectral properties of these interpolated spectra as a function of iteration of the training algorithm.

The redshift estimates based on the continuous one-dimensional type parameter are shown in the bottom panels

of Figure 11 for both the early- and late-type subsamples (*left and right, respectively*). Compared with the top panels of the discrete version (discussed previously), the new estimates seem to be superior for the intrinsically blue subset and slightly worse for the early types.

For early-type galaxies, it would be better to use the original discrete template set to avoid the systematic overestimation around $z = 0.2$ and 0.3. Since we want to have a simple estimation for the spectral type, we would like to avoid using a separate (discrete) template set for early-type galaxies, so we use the above scheme, keeping in mind the systematic errors and working on a better interpolated template set. Note that SDSS will measure spectroscopic redshift for most of the luminous early-type galaxies, so the number of objects for which this problem arises is somewhat smaller than in our test sample. However, for the less luminous early-type galaxies, the above problem still exists.

In terms of rms values of the scatter this translates into an increase from $\sigma_z = 0.028$ to 0.029 for the red galaxies and a decrease from $\sigma_z = 0.05$ to 0.04 for the blue ones. To quote an rms for the entire training set would not be too meaningful, because it depends on the ratio of the number of early- and late-type galaxies. For the main SDSS galaxy sample the scatter is $\sigma_z = 0.035$. We will use the above template-fitting method with repaired interpolated templates to create the EDR photometric redshift catalog.

5. COMPARISONS WITH INDEPENDENT REDSHIFT SAMPLES

5.1. *The 2dF and CNOC2 Redshift Samples*

In the preceding sections we have used data from the same subsets for training and testing. We now perform a blind test, using the independent data sets. Details of the 2dF and CNOC2 data sets are given in § 2. Figure 14*a* compares the spectroscopic and photometric redshifts for the 2dF spectroscopic sample. The dispersion in the photometric redshift relation for these data is $\sigma_z = 0.043$. This
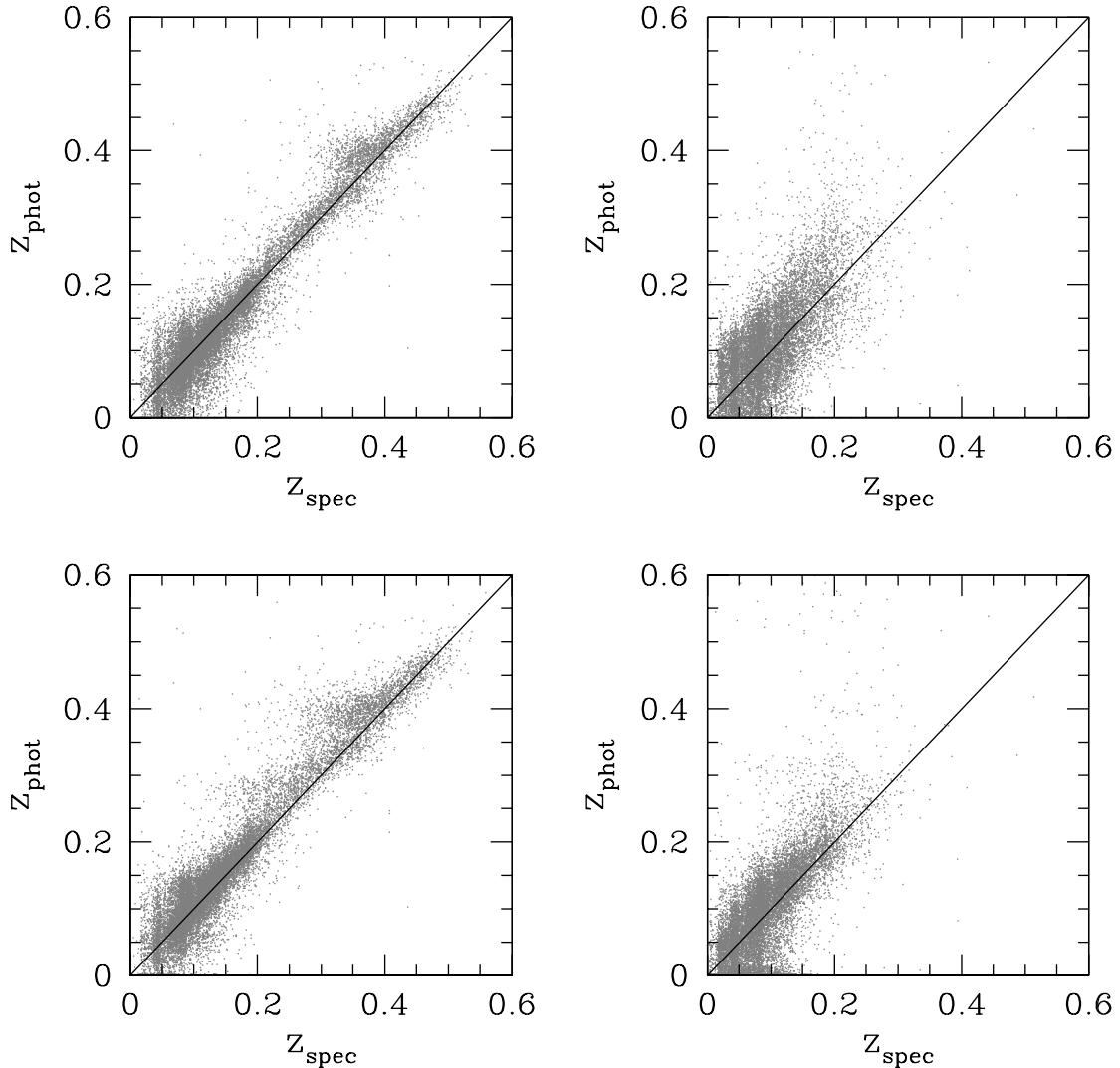
Fig. 11.—Photometric redshifts of intrinsically red (*left*) and blue (*right*) galaxies. The figures show the spectroscopic vs. photometric redshifts for the four discrete templates (*top*) and for the continuous type (*bottom*) estimators. The type interpolation makes things slightly worse for the red early-type galaxies because of the type-redshift degeneracy, but the estimates for the late-type galaxies get significantly better.

compares with the dispersion in the relation for the full SDSS sample of $\sigma_z = 0.035$. The increase in the dispersion arises from two effects. The r'-band magnitudes of the 2dF data are intrinsically fainter than the SDSS spectroscopic sample (by approximately 0.2 mag), and the 2dF data are
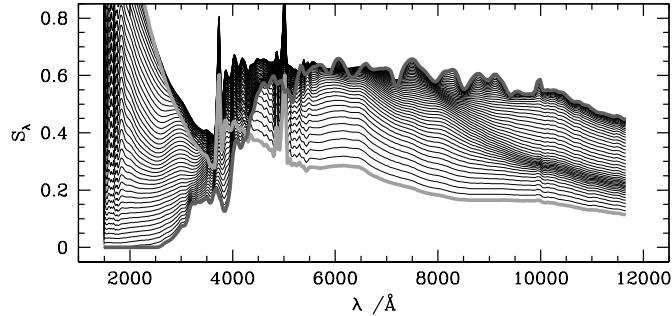

Fig. 12.—Illustration of the one-dimensional–type manifold. A few SEDs are plotted here for a equally spaced type parameter values. The reddest and bluest SEDs are shown with the thick dark and light gray curves, respectively.

selected based on their B_j photographic magnitudes, which will provide an intrinsically bluer galaxy sample than the r'-selected SDSS data. As the dispersion in the redshift relation increases with limiting magnitude and for blue galaxies, the difference in the observed photometric redshift relation is not surprising.

To determine how well the templates extrapolate to higher redshift data, we apply the photometric redshifts to the CNOC2 data set (with a redshift range $0 < z < 0.7$ and a magnitude limit of $r' < 21.0$). As we can see in Figure 14, the dispersion in the relation increases for the fainter magnitude sample as a result of the increase in photometric error. The average estimation error for the whole set is $\sigma_z = 0.084$. If we consider only those galaxies with $17.8 < r' < 19.5$, the uncertainty in the redshift estimates decreases to $\sigma_z = 0.061$. In Figure 15 we show the absolute deviation between the photometric and spectroscopic redshifts for the CNOC2 galaxy sample as a function of r'. The cumulative rms of these data (as a function of r') is shown by the solid line. For $r' < 21$, the rms uncertainty about this relation is 0.1 in redshift.
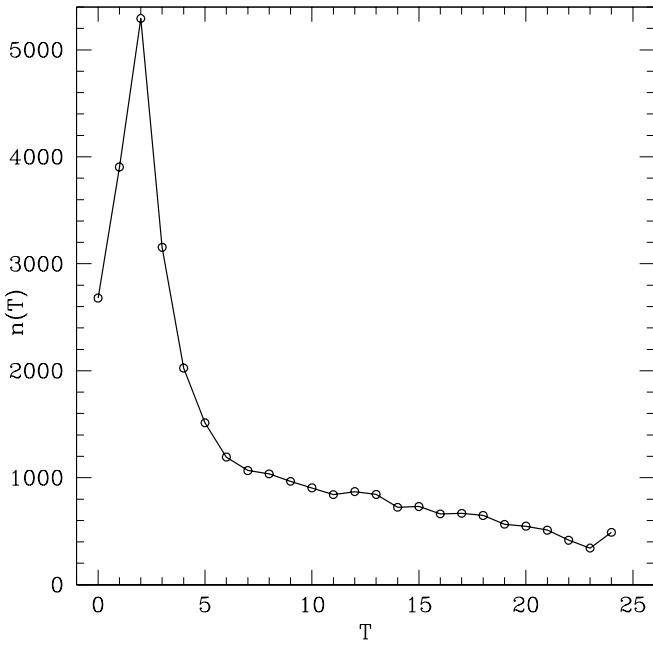
FIG. 13.—Distribution of interpolated spectral templates that fit the observed colors in the EDR main galaxy sample. The smooth distribution shows that no particular spectral template is preferred (i.e., the galaxies do not fall into a small number of spectral types). This implies that the spline used to interpolated between the trained spectral energy distributions accurately maps the distribution of galaxy colors.

6. THE EARLY DATA RELEASE PHOTOMETRIC REDSHIFT CATALOG

6.1. *Selecting Galaxies from the EDR Database*

The goal of our analysis has been to obtain photometric redshifts for all SDSS galaxies in the Early Data Release. We have, therefore, created the first EDR photo-z catalog (version 1.0), which has now been included in the publicly available EDR database.[10]

We used the template-fitting method with repaired interpolated templates (see § 4) to estimate photometric redshifts in the above public catalog. Though the empirical methods (see § 3.1) give a smaller estimation error, we have chosen to use the template-fitting method, since it estimates not just redshift but spectral type and rest-frame magnitude as well. Also we hope that, with the accumulation of more precisely calibrated data in further SDSS releases, the disadvantage of this method will lessen.

The photometric redshift table (see Table 2 for the list of parameters) in the database has more than 6 million entries, one for every galaxy in the EDR. Each entry contains the unique object ID (`objID`, for quick cross-matching), the most likely redshift (`z`) and type (`t`). The uncertainties of redshift and type are calculated from the 68% confidence regions of the fit, assuming Gaussian errors. Note that the true error distribution for higher redshift object is not known, and probably not Gaussian. The elements of the covariance matrix are stored in the database and represented by `c_zz`, `c_tt`, `c_tz`. The errors in columns `zErr` and `tErr` are simply taken from the diagonal elements of the covariance matrix. The χ^2 value of the fit (`chiSq`) measures the absolute " goodness " of the fit. The catalog contains a preliminary quality flag (`quality`), which scales between zero and five, where the larger the number the more confident the photometric redshift. This flag is assigned to objects in the process of fitting the confidence region and seems to correlate with the rms of the photometric and spectroscopic redshifts. In the current version this correlation is quite weak, and we would like to improve the calculation of this flag in the next version.

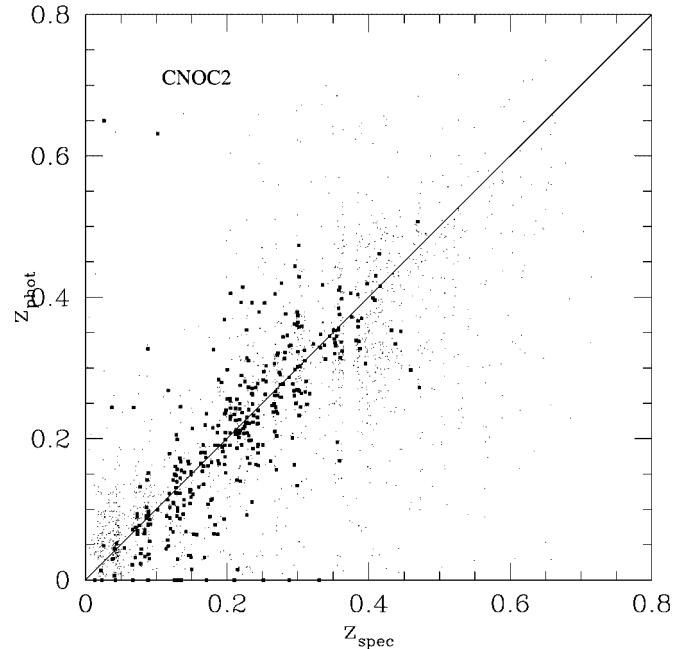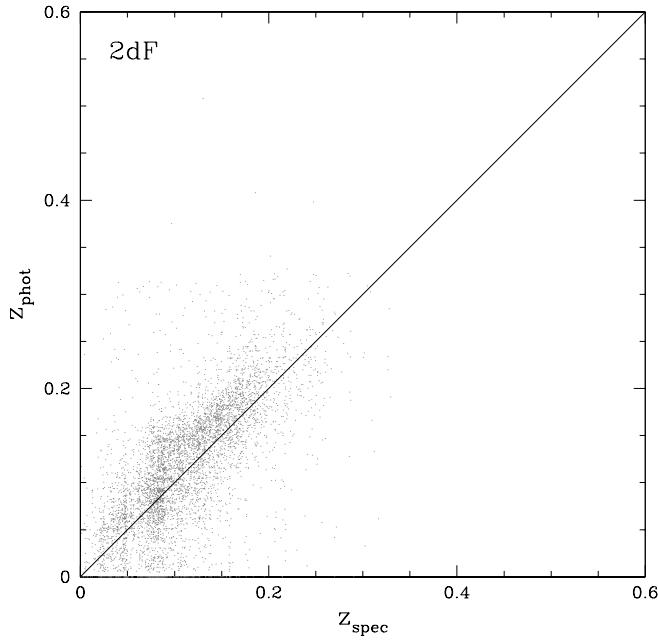[10] Available at http://skyserver.sdss.org.

FIG. 14.—Checking the extrapolation capabilities of the photometric redshift estimator: the predicted vs. the spectroscopic redshift. *Left*: 2dF set. *Right*: CNOC2 set; since most of these objects are too faint, we show with larger symbols the objects with reasonable SDSS photometry ($17.8 < r < 19.5$). Note the different redshift range.
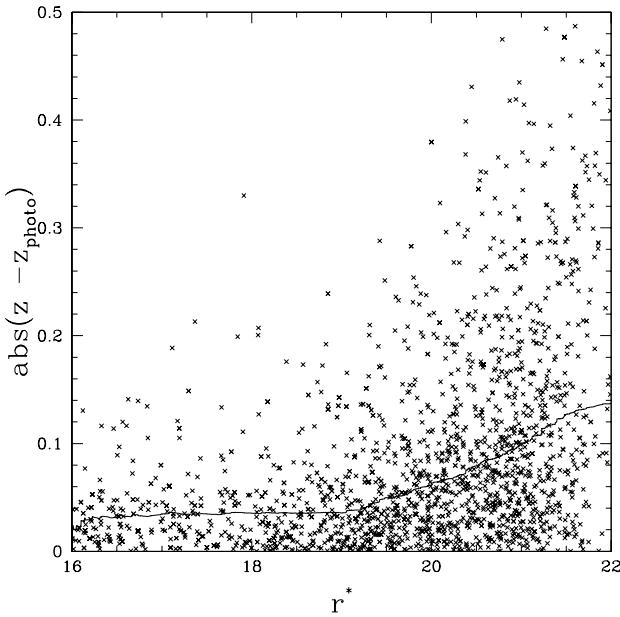
Fig. 15.—Cumulative rms of the SDSS photometric redshift as a function of limiting magnitude. The points represent the absolute deviation between the spectroscopic and photometric redshifts for the CNOC2 sample of galaxies. The solid line is the cumulative rms of the sample as a function of the r' magnitude. At a limiting magnitude of $r' < 21$ the rms error on the photometric redshift rises to 0.1.

In addition to the redshift estimates, physical parameters derived from the estimated redshift are also stored in the database. These include the distance modulus (dmod) for the standard ΛCDM cosmology ($\Omega_M = 0.3$, $\Omega_\Lambda = 0.7$, h^{-1} units), rest-frame colors (rest_ug, rest_gr, rest_ri, rest_iz), and k-corrections (kcorr_u, kcorr_g, kcorr_r, kcorr_i, kcorr_z) derived directly from the templates and the rest-frame absolute magnitudes (absMag_u, absMag_g, absMag_r, absMag_i, absMag_z) as computed from the distance modulus and k-correction,

$$M = m - \mathrm{DM}(z) - k(t, z) \ .$$

Access to these parameters is straightforward through the Structured Query Language (SQL). A sample query to extract the object identification and photometric redshift of five galaxies in the redshift range of $0.2 < z < 0.3$ would look like this:

```
select top 5 objId, z from PhotoZ where z > 0.2 and
z < 0.3
```

All parameters stored in the SDSS database (including the derived parameters) can be searched upon.

6.2. Caveats and Limitations of the Current Photometric Redshifts

While, as the comparisons between the photometric and spectroscopic redshift show, the current implementation of

TABLE 2
PHOTOMETRIC REDSHIFT PARAMETERS

Name	Type	Length	Units	Description
pId	int	4	...	Unique ID for photoz version
rank	int	4	...	Rank of the photoz determination; default is 0
version	varchar	6	...	Version of photoz code
class	int	4	...	Number describing the object type (galaxy = 1, QSO = tbd, ...)
objID	bigint	8	...	Unique ID pointing to PhotoObj table
chiSq	real	4	...	The χ^2 value for the fit
z	real	4	...	Photometric redshift
zErr	real	4	...	Marginalized error of the photometric redshift
t	real	4	...	Photometric SED type between 0 and 1
tErr	real	4	...	Marginalized error of the photometric type
c_tt	real	4	...	tt-element of covariance matrix
c_tz	real	4	...	tz-element of covariance matrix
c_zz	real	4	...	zz-element of covariance matrix
fitRadius	int	4	pixels	Radius of area used for covariance fit
fitThreshold	real	4	...	Probability threshold for fitting, peak normalized to 1
quality	int	4	...	Integer describing the quality (best:5, lowest 0)
dmod	real	4	mag	Distance modulus for $\Omega_M = 0.3$, $\Omega_\Lambda = 0.7$ cosmology
rest_ug	real	4	mag	Rest-frame $u-g$ color
rest_gr	real	4	mag	Rest-frame $g-r$ color
rest_ri	real	4	mag	Rest-frame $r-i$ color
rest_iz	real	4	mag	Rest-frame $i-z$ color
kcorr_u	real	4	mag	k-correction
kcorr_g	real	4	mag	k-correction
kcorr_r	real	4	mag	k-correction
kcorr_i	real	4	mag	k-correction
kcorr_z	real	4	mag	k-correction
absMag_u	real	4	mag	Rest-frame u' absolute' magnitude
absMag_g	real	4	mag	Rest-frame g' absolute magnitude
absMag_r	real	4	mag	Rest-frame r' absolute magnitude
absMag_i	real	4	mag	Rest-frame i' absolute magnitude
absMag_z	real	4	mag	Rest-frame z' absolute magnitude

NOTE.—Parameters contained in the photoz table of the SDSS Science Archive (http://skyserver.sdss.org). See text for more details.

SDSS photometric redshifts provides an accurate estimate of the redshifts, there are a number of limitations and caveats pertaining to the EDR data. We describe here the results of a series of tests of the quality of the SDSS photometry and how these issues affect the accuracy and possible uses of the photometric redshifts in the EDR catalog. We advise any potential user of the current photometric redshift implementation to be aware of these caveats prior to undertaking any statistical analysis.

Even though the photometric calibration of the SDSS survey has been shown to be accurate to a few percent for the SDSS standard stars, galaxy colors appear to have a slight offset from SED-based estimated values (Eisenstein et al. 2001). As part of this analysis of the SDSS EDR data we compare measured colors not only with the spectrophotometrically calibrated SEDs (e.g., CWW), but we have also carried out experiments where small offsets were applied before refining the template spectra. In this way we can identify systematic photometric offsets from the mean deviation of the colors from the SEDs. The g'-band offset we found is in the same sense as that given in Eisenstein et al. (2001) but with a smaller amplitude of $\Delta g' \sim 0.05$. All galaxies in the SDSS catalog had this g' offset applied prior to calculation of the photometric redshifts.

Our SED reconstruction algorithm ideally requires a training set with reasonably uniform redshift distribution over a large baseline. The SDSS spectroscopic survey delivers excellent quality data for this kind of analyses. However, the main galaxy sample has a median redshift of approximately 0.1, which does not enable the use photometric data from different bands to constrain the SEDs at all wavelengths. In principle, if there exist photometric zero-point uncertainties in the data, the reconstruction could introduce artificial continuum spectral features in the templates, which would make the extrapolation to higher redshifts impossible (in a similar sense to the limitations of the empirical techniques). The repaired spectral energy distributions show no obvious trace of such features.

Finally, we consider how the increasing photometric uncertainty at fainter magnitudes affects the redshift histograms. In Figure 16 we show the redshift distributions in different r' magnitude bins 16–17, 17–18, 18–19, 19–20, and 20–21. The histograms built in different magnitude bins peak around values consistent with published redshifts surveys, and they move toward higher values as a function of the magnitude. Beyond a magnitude limit of $r' > 21$ artifacts are seen in the redshift histograms that are due to the large photometric errors. We therefore advise caution when using the current EDR photometric redshift catalog for galaxies with $r' > 21$. Also, one should take into consideration the fact that, for some objects, the photometric redshift would be negative because the estimation is based on photometric data with errors. But the algorithm allows only positive redshift values, so all negative redshifts pile up at $z = 0$.

7. CONCLUSIONS

We present the first application of photometric redshifts to the SDSS EDR data. From a comparison of the photometric and spectroscopic redshifts we find that the rms error in the redshift relation is 0.035 for $r' < 18$, rising to 0.1 at $r' < 21$. For magnitude intervals $r' < 21$ the photometric redshift relation and redshift histogram are well matched to existing redshift surveys (with comparable median redshifts and dis-
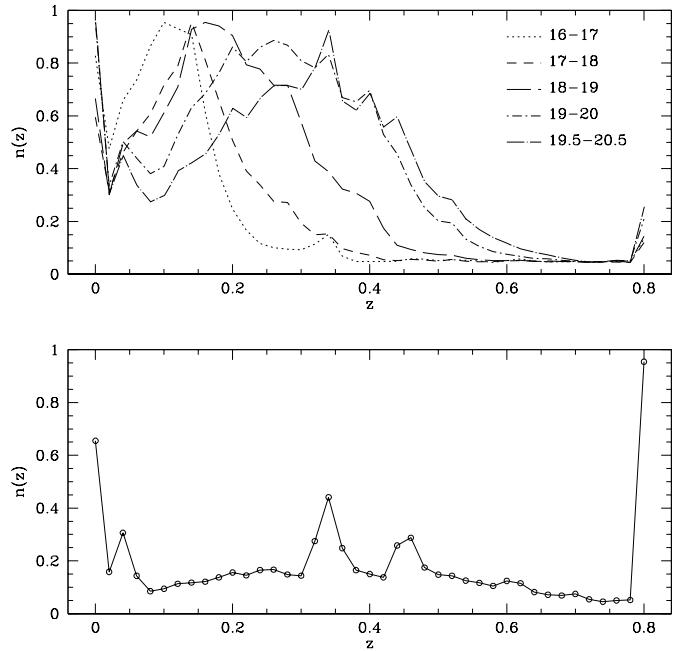

Fig. 16.—Redshift distributions plotted to show the trend with the apparent r'-band magnitude. As expected, the histograms in the top figure are shifted to the right as we go with r' magnitude bins from $16 < r' < 17$ to $19.5 < r' < 20.5$. The histogram in the bottom figure is built using all galaxies in the EDR catalog and has artifacts. We therefore advise caution when using the current EDR photometric redshift catalog for galaxies with $r' > 21$.

persions). Implementing these redshift estimates in the SDSS EDR database, together with derived quantities such as the absolute magnitudes, k-corrections, and rest-frame colors, we provide a simple interface to one of the largest publicly accessible catalogs of photometric redshifts available to the astronomical community. We conclude by providing a description of the limitations and caveats present in the current photometric redshift implementation. We caution all users to be aware of these limitations before applying the EDR photometric redshifts in any statistical analyses.

CSABAI ET AL.

REFERENCES

Benítez, N. 2000, ApJ, 536, 571
Bentley, J. L. 1979, Commun. ACM, 19, 509
Bolzonella, M., Miralles, J.-M., Pelló, R. 2000, A&A, 363, 476
Brunner, R. J., Connolly, A. J., & Szalay, A. S. 1999, ApJ, 516, 563
Bruzual, A. G., & Charlot, S. 1993, ApJ, 405, 538
Budavári, T., Szalay, A. S., Connolly, A. J., Csabai, I., & Dickinson, M. E. 1999, in ASP Conf. Ser. 191, Photometric Redshifts and High Redshift Galaxies, ed. R. J. Weymann, L. J. Storrie-Lombardi, M. Sawicki, & R. Brunner (San Francisco: ASP), 19
———. 2000, AJ, 120, 1588
Budavári, T., et al. 2001, AJ, 122, 1163
———. 2001, AJ, 121, 3266
Coleman, G. D., Wu, C.-C., & Weedman, D. W. 1980, ApJS, 43, 393
Colless, M. M., et al. 2001, MNRAS, 328, 1039
Connolly, A. J., Budavári, T., Szalay, A. S., Csabai, I., & Brunner, R. J. 1999, in ASP Conf. Ser. 191, Photometric Redshifts and High Redshift Galaxies, ed. R. J. Weymann, L. J. Storrie-Lombardi, M. Sawicki, & R. Brunner (San Francisco: ASP), 13
Connolly, A. J., Csabai, I., Szalay, A. S., Koo, D. C., Kron, R. G., & Munn, J. A. 1995, AJ, 110, 2655
Csabai, I., Connolly, A. J., Szalay, A. S., & Budavári, T. 2000, AJ, 119, 69
Eisenstein, D. J., et al. 2001, AJ, 122, 2267

Fernández-Soto, A., Lanzetta, K. M., & Yahil, A. 1999, ApJ, 513, 34
Fukugita, M., Ichikawa, T., Gunn, J. E., Doi, M., Shimasaku, K., & Schneider, D. P. 1996, AJ, 111, 1748
Gunn, J. E., et al. 1998, AJ, 116, 3040
Gwyn, S. D. J., & Hartwick, F. D. A. 1996, ApJ, 468, L77
Hogg, D. W., et al. 1998, AJ, 115, 1418
Hogg, D. W., Schlegel, D. J., Finkbeiner, D. P., & Gunn, J. E. 2001, AJ, 122, 2129
Koo, D. C. 1985, AJ, 90, 418
Moore, A. W., Schneider, J., & Deng, K. 1997, Proc. Int. Machine Learning Conference, ed. D. Fisher (San Francisco: Morgan Kaufmann)
Pier, J. R., Munn, J. A., Hindsley, R. B., Hennessy, G. S., Kent, S. M., Lupton, R. H., & Ivezic, Z. 2002, AJ, submitted
Sawicki, M. J., Lin, H., & Yee, H. K. C. 1997, AJ, 113, 1
Smith, J. A., et al. 2002, AJ, 123, 2121
Stoughton C., et al. 2002, AJ, 123, 485
Strauss, M. A., et al. 2002, AJ, 124, 1810
Yee, H. K. C., Ellingson, E., & Carlberg, R. G. 1996, ApJS, 102, 269
York, D. G., et al. 2000, AJ, 120, 1579
Wang, Y., Bahcall, N., & Turner, E. L. 1998, AJ, 116, 2081
Williams, R. E., et al. 1996, AJ, 112, 1335