SLAC-TN-12-012

An Improved Technique for Increasing the Accuracy of Photometrically Determined Redshifts for "Blended" Galaxies

Ashley Marie Parker

Marietta College, Marietta, Ohio

Office of Science, Science Undergraduate Laboratory Internship Program

SLAC National Accelerator Laboratory

Menlo Park, California

August 13, 2011

Prepared in partial fulfillment of the requirements of the Office of Science, Department of Energy's Science Undergraduate Laboratory Internship, under the direction of Dr. Deborah Bard, Kavli Institute for Particle Astrophysics and Cosmology at SLAC.

Participant: ______________________________

Research Advisor: ______________________________

Work supported in part by US Department of Energy contract DE-AC02-76SF00515.

TABLE OF CONTENTS

Abstract

An improved technique for increasing the accuracy of photometrically determined redshifts for “blended” galaxies. ASHLEY M. PARKER (Marietta College, Marietta, OH 45750) DEBORAH J. BARD (Kavli Institute for Particle and Astrophysics, Menlo Park, CA 94025)

The redshift of a galaxy can be determined by one of two methods; photometric or spectroscopic. Photometric is a term for any redshift determination made using the magnitudes of light in different filters. Spectroscopic redshifts are determined by measuring the absorption spectra of the object then determining the difference in wavelength between the “standard” absorption lines and the measured ones, making it the most accurate of the two methods.

The data for this research was collected from SDSS DR8 and then separated into blended and non-blended galaxy sets; the definition of “blended” is discussed in the Introduction section. The current SDSS photometric redshift determination method does not discriminate between blended and non-blended data when it determines the photometric redshift of a given galaxy. The focus of this research was to utilize machine learning techniques to determine if a considerably more accurate photometric redshift determination method could be found, for the case of the blended and non-blended data being treated separately. The results show a reduction of 0.00496 in the RMS error of photometric redshift determinations for blended galaxies and a more significant reduction of 0.00827 for non-blended galaxies, illustrated in Table 2.

Introduction

The goal of this project is to utilize Sloan Digital Sky Survey's (SDSS) data along with machine learning techniques to ultimately increase the reliability of photometric redshift analysis for blended and non-blended galaxies. Currently in SDSS database, self adjusting algorithms are used to determine photometric redshifts for all objects as a set, although none of these algorithms are optimized for galaxies [1]. This summer research aims determine if more accurate photometric redshift measurements will result from looking at only galaxy data and separating blended from non-blended.

A "blended" object is defined by the SDSS database as a light source (e.g. galaxy, star, etc.) for which intensity analysis shows multiple intensity peaks in the single light source, meaning there are multiple objects present [1]. Figure 1 shows a simple illustration of the deblending process, taken from a paper on the SDSS deblending algorithm [2]. Within the database the "frames pipeline" analyzes the data to determine if a light-emitting object is blended. If so, a de-blending algorithm is used to separate the multiple objects into "child" objects whose spectra add to become the "parent" image [1]. After de-blending these "child" light sources are all treated as separate objects, and are categorized as blended data. Objects that are flagged as blended are given a unique parentID number greater than zero, otherwise parentID is set to zero for the non-blended [1].

This parentID numbering is used in SDSS's newest data release, DR8, which covers approximately one third of the sky and includes all spectroscopic measurements that will be taken with this imaging camera [1]. This research project will use data acquired from SDSS which includes approximately 900,000 galaxies for which both photometric and spectroscopic data has been recorded. The reason both photometric and spectroscopic methods will be analyzed is that an accurate redshift measurement, assuming that the spectroscopic is the “true” redshift, must exist to test results from the new machine learning techniques.

A photometric redshift is measured using photometry, a method of looking at the light from an object through 5 standard filters (u, g, r, i, z) and using the overall magnitudes per filter to determine the redshift. The average wavelengths for the SDSS filters are shown in Table 1 [1]. Photometry is much less time consuming than the alternate method of spectroscopic redshift determination. In order to spectroscopically measure redshift there must be significantly more light collected for the object, so the full spectrum can be seen rather than just the intensities per filter. This makes spectroscopic far more accurate, however due to the large amount of telescope time it requires, the photometric is the most commonly used method. Telescope time is even more precious when it is used for a large scale sky survey, this is the reasoning behind efforts to increase the accuracy of the photometric method.

This is, to my knowledge, an original research project which will yield a photometric redshift determination method, specifically for blended or non-blended galaxies, which could be implemented in the next generation of sky survey databases, namely LSST, Large Synaptic Survey Telescope. This increase in accuracy of photometric redshift measurements will impact many scientific measurements, which rely on redshift, such as the study of large scale structure of the universe and gravitational lensing.

The goal was to find a method which yielded results for blended galaxies which were more accurate than SDSS's photometric redshift values. The goal for the non-blended galaxies was to determine a method equivalent to SDSS's method, although we did not expect to be capable of doing a significantly more accurate determination on this data set.

In the future I will investigate the scientific application of the, more accurate, photometric redshifts. Ultimately this work is hoped to be useful for the future LSST database which will not contain spectroscopic redshift measurements for all objects, and will therefore need to make use of the photometric method.

Methods

The initial step of the project was to learn SQL, Structured Query Language, which is used to write queries that acquire data from SDSS. The multitude of data available on SDSS's database makes it ideal for “training” a machine learning program such that the example data should show nearly every variation in galaxy type. The data used for the “training” consisted of approximately 100,000 non-blended galaxies and 800,000 that were blended. The difference in sample size was unintentional; it is a product of the fact that most of SDSS’s galaxy data is blended.

This project made use of CasJobs DR8, a program which utilizes SQL queries to acquire large amounts of data, from the most recent data release. The queries allowed for request of specific useful quantities such as: spectroscopic redshift measurement, two separate photometric redshift measurements using “random forest” and “robust fit” methods, magnitudes in the bands u, g, r, i, z, parentID and uncertainty measurements for all relevant quantities. An example of one SQL query for non-blended galaxy data, which was used for this research, is shown in Figure 3.

Data was requested for all objects which are of the type “galaxy” and downloaded for use in the ROOT data analysis framework. Data was requested for both blended and non-blended objects separately, so that a determination could be made if the photometric redshift measurements for blended galaxies are less accurate than measurements of non-blended objects. There was an investigation into which of the predetermined photometric redshift determination methods is most accurate, which was determined to be the “robust fit” technique.

For this research a machine learning technique, specifically TMVA, was used in order to find a more accurate method for determination of photometric redshift for blended galaxies. The Toolkit for Multivariate Analysis (TMVA) is a 'ROOT-integrated environment' which allows multivariate regression techniques to be used to analyze large data sets [3]. A TMVA regression script template was edited to include all necessary information about the studied galaxies. The goal of the script is to start with only the galaxy magnitudes in the 5 different bandwidth filters (u, g, r, i, z) and from that determine a redshift which is in good agreement with the spectroscopic value, which is considered the “correct” redshift. Within the script various methods attempt to determine redshifts, afterwards a given method's results were compared with the spectroscopic redshift determined by SDSS to see which method gave the best approximation.

The TMVA regression script was “trained” separately for blended and non-blended galaxies. All available TMVA methods were tested on the data, which include: PDERS, PDERSkNN, KNN, LD, FDA_GA, FDA_MC, FDA_MT, FDA_GAMT, MLP, SVM, BDT, and BDTG. All method titles are acronyms which describe the underlying mathematics of the method, there are far too many methods to describe them all in sufficient detail in this paper, for specific information regarding the individual methods see reference [3], the TMVA users guide.

The data analysis began by determining the accuracy of the current SDSS photometric method by comparing it to the spectroscopic data, which was needed to show the improvement of the machine learning methods used. In this study the accuracy of a given method was determined by an RMS error, given by:

$$\sigma_z \equiv \sqrt{\overline{(\delta z)^2} - (\overline{\delta z})^2} \qquad \text{Equation 1.}$$

Where $\delta z = z_{method} - z_{spectroscopic}$, with z being the determined redshift and the overhead bars in the equation representing when the average is taken [4].

Results

On the far right of Figures 4 and 5 the graph shows the difference in redshifts as determined by spectroscopic and SDSS photometric methods, for non-blended and blended data respectively. From these figures, 4 and 5, it is clear that for both data sets there are large discrepancies between the photometric and spectroscopic redshift determinations. It is also clear from the shapes of the curves that the blended and non-blended data are distinctly different in their shape and reaction to the SDSS photometric method, thus showing the need for this research. The RMS errors for non-blended and blended data for the SDSS photometric method are shown in Table 2, 0.0724 and 0.04587, respectively.

In order to determine the most accurate of the TMVA methods, many histograms were produced showing the photometrically determined redshifts to be compared with the spectroscopic. These are depicted in Figures 7 and 8, which show the redshifts for the 7 most accurate TMVA methods and the SDSS photometric method. The best method should have a 2 peak profile similar to the graphs on the far left of figures 4 and 5, which show the spectroscopic or “true” redshift. Notice that the values end promptly at zero and do not go negative. Since the current “standard” cosmological theory infers that the universe is expanding, all galaxies should be moving away from the Milky-Way, making all redshift values of galaxies positive, therefore a negative value is considered non-physical.

Figures 9 and 10 show the “best” (lowest RMS error) TMVA method redshift values as a function of the spectroscopic redshift and compare those to the SDSS photometric data as a function of spectroscopic redshift, for blended and non-blended data respectively. When comparing figures 9 and 10, it becomes obvious that the non-blended data seems very tight where the blended data shows a large spread, which can partially be contributed to the blended data set being much larger but is also believed to be attributed to shortcomings of the SDSS deblending algorithm.

Discussion and Conclusion

Table 2 shows that for blended galaxy data the KNN method gave the most accurate results which are slightly more accurate than the SDSS determination. Table 2 also shows that for non-blended galaxies the MLP method yields the most accurate results for redshift, significantly more accurate than the SDSS determination. This was not an expected result; this research began with the hypothesis that a better photometric method could be determined for blended galaxy data and a similar, but not significantly more accurate, method would be found for non-blended data. This hypothesis was exactly the opposite of what was determined by the research: a significantly more accurate photometric redshift determination method was found for non-blended galaxy data while only a slight improvement was made on the blended galaxy data.

It is clear, from figures 7 and 8, that the BDT, LD and FDA_GAMT methods, first row right, second row left and third row right, respectively in both figures 7 and 8, are not the best methods due to the fact that they clearly include many non-physical redshift values. The BDTG method, shown on first row right in both figures 7 and 8, is also clearly not the best method due to the large amount of noise in the data which leads to the strange shape of the curve. The PDERSkNN method, shown in figures 7 and 8, third row left, yielded an unusual pile-up of redshift determinations at 0 which is not representative of the actual data. The data from these figures was input into Equation 1, along with the spectroscopic data from figures 4 and 5, in order to determine the RMS errors for all methods, which are displayed in Table 2.

In conclusion, a significant improvement over the SDSS photometric redshift determination method was made for non-blended galaxies. This will have far reaching effects on how photometric redshifts are determined for future large-scale sky-surveys, such as LSST. This work should be taken into account when the de-blending algorithms for the LSST database are being developed since the blended data clearly behaves differently from the non-blended due to the effects of the de-blending process. These improved photometric redshift determination methods should also be applied to existing data so that a more accurate representation of the universe can be seen.

Acknowledgments

I would like to express my gratitude to my mentor, Dr. Deborah Bard, for her assistance with this project. Dr. Bard was invaluable in my research effort; providing her guidance as well as many useful references along with assistance in troubleshooting and de-bugging my many ROOT macros. I would also like to thank Dr. Steven Rock, director of the SULI program at SLAC, for his efforts in making this year's SULI program a success. This research was funded by the U.S. Department of Energy, and the Kavli Institute for Particle Astrophysics and Cosmology.

References

[1] http://sdss.org, July 2011.

[2] http://www.astro.princeton.edu/~rhl/photomisc/deblender.ps.gz, August 2011.

[3] “TMVA Users Guide”, http://tmva.sourceforge.net/, August 2011.

[4] http://arxiv.org/PS_cache/astro-ph/pdf/0605/0605303v2.pdf, August 2011.

Figures

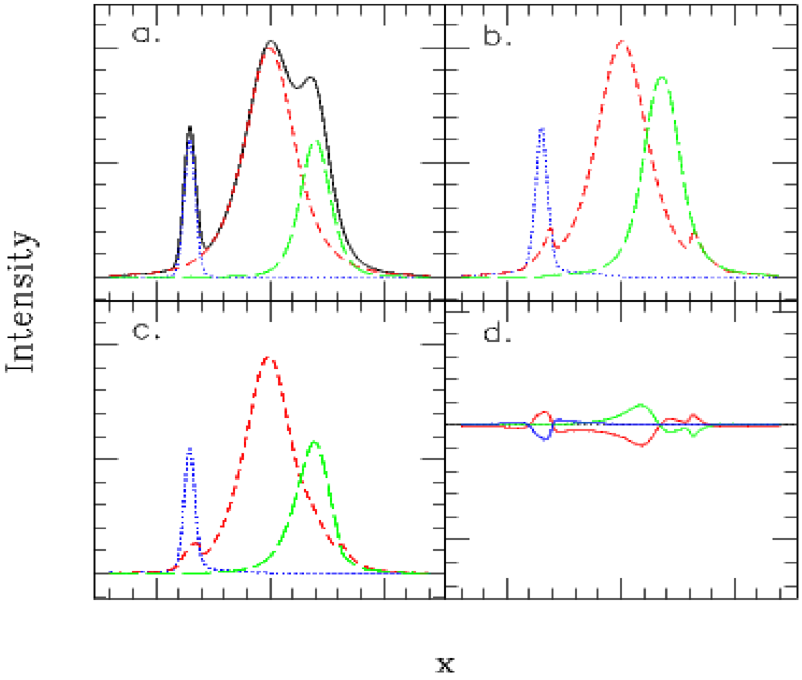
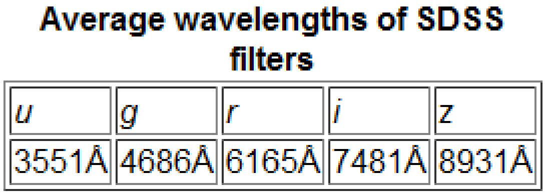


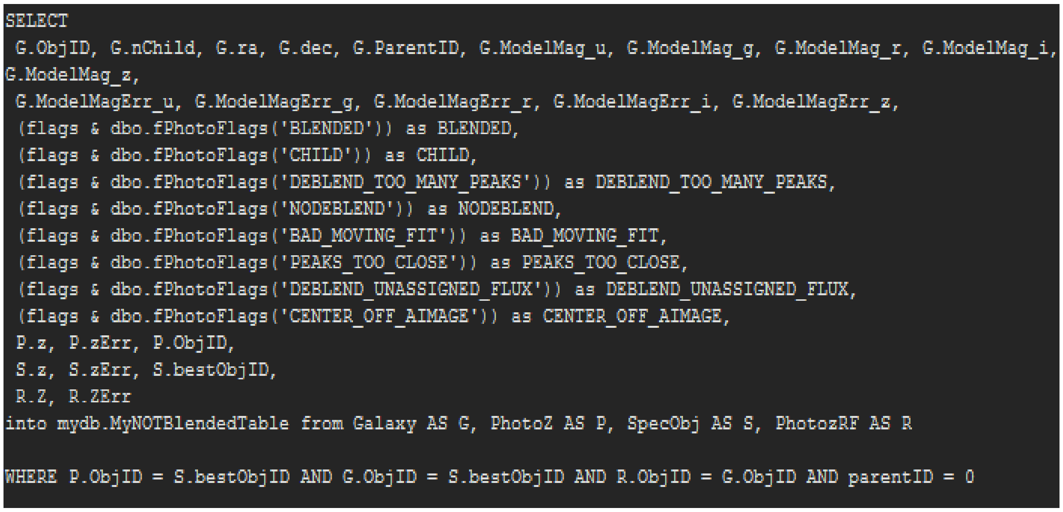
Figure 1: a) here is a “blended” parent object, solid black line, consisting of 3 child objects shown as colored dotted lines. b and c) shows the corresponding deblended children d) shows the difference between the sum of the children and the original parent, illustrating the imperfection of the method.



Average wavelengths of SDSS filters

u	*g*	*r*	*i*	z
3551Å	4686Å	6165Å	7481Å	8931Å

Table 1: The average wavelength, in angstrom, of the 5 model magnitude filters for SDSS DR8 data [1].



```
SELECT
 G.ObjID, G.nChild, G.ra, G.dec, G.ParentID, G.ModelMag_u, G.ModelMag_g, G.ModelMag_r, G.ModelMag_i,
G.ModelMag_z,
 G.ModelMagErr_u, G.ModelMagErr_g, G.ModelMagErr_r, G.ModelMagErr_i, G.ModelMagErr_z,
 (flags & dbo.fPhotoFlags('BLENDED')) as BLENDED,
 (flags & dbo.fPhotoFlags('CHILD')) as CHILD,
 (flags & dbo.fPhotoFlags('DEBLEND_TOO_MANY_PEAKS')) as DEBLEND_TOO_MANY_PEAKS,
 (flags & dbo.fPhotoFlags('NODEBLEND')) as NODEBLEND,
 (flags & dbo.fPhotoFlags('BAD_MOVING_FIT')) as BAD_MOVING_FIT,
 (flags & dbo.fPhotoFlags('PEAKS_TOO_CLOSE')) as PEAKS_TOO_CLOSE,
 (flags & dbo.fPhotoFlags('DEBLEND_UNASSIGNED_FLUX')) as DEBLEND_UNASSIGNED_FLUX,
 (flags & dbo.fPhotoFlags('CENTER_OFF_AIMAGE')) as CENTER_OFF_AIMAGE,
 P.z, P.zErr, P.ObjID,
 S.z, S.zErr, S.bestObjID,
 R.Z, R.ZErr
into mydb.MyNOTBlendedTable from Galaxy AS G, PhotoZ AS P, SpecObj AS S, PhotozRF AS R

WHERE P.ObjID = S.bestObjID AND G.ObjID = S.bestObjID AND R.ObjID = G.ObjID AND parentID = 0
```

Figure 3: This SQL query is requesting data on non-blended galaxies where both spectroscopic and photometric data is available.

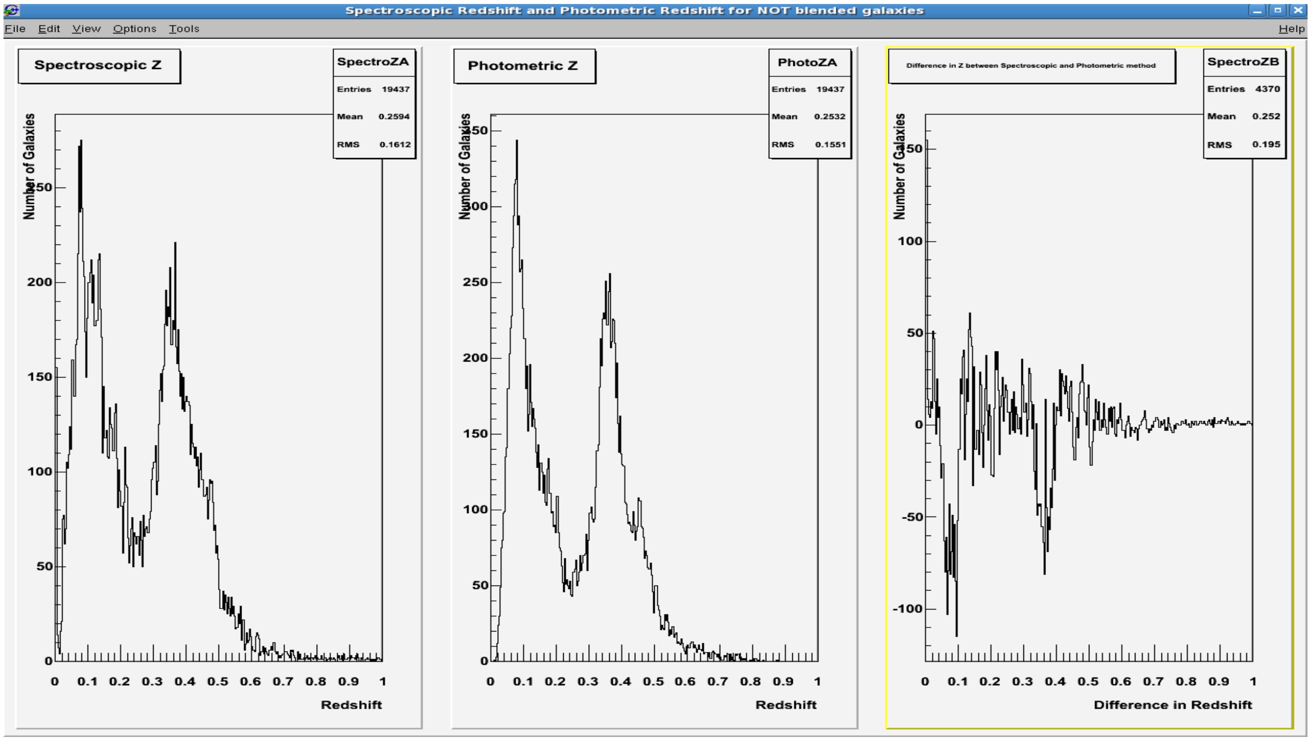


Figure 4: Shows non-blended data: on the far left is the spectroscopic redshift determination, middle shows the SDSS photometric method and on far right is the difference between the two.

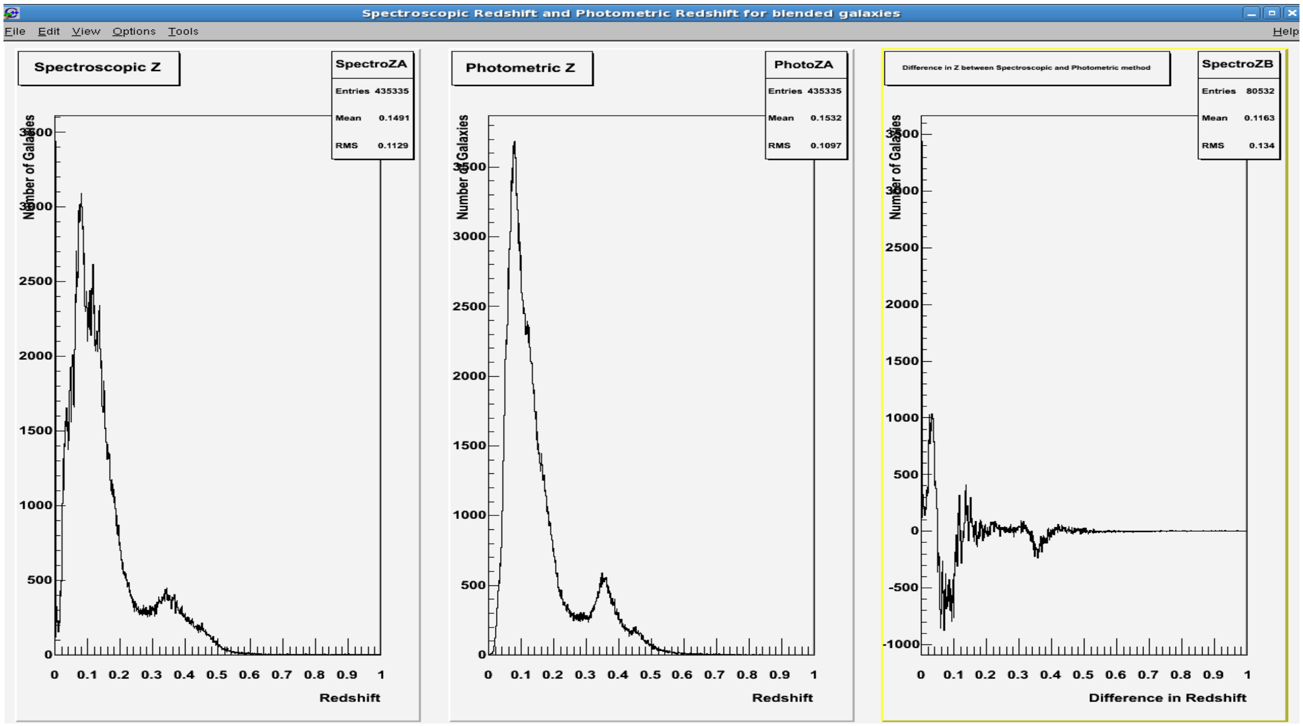
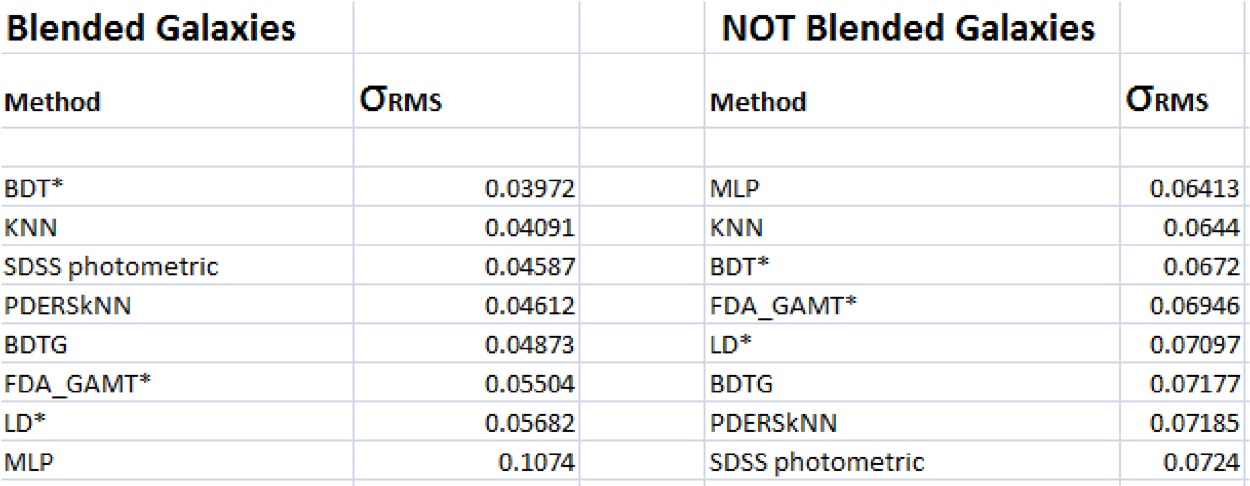


Figure 5: Shows blended data: on the far left is the spectroscopic redshift determination, middle shows the SDSS photometric method and on far right is the difference between the two.



Blended Galaxies			NOT Blended Galaxies	
Method	σ_{RMS}		Method	σ_{RMS}
BDT*	0.03972		MLP	0.06413
KNN	0.04091		KNN	0.0644
SDSS photometric	0.04587		BDT*	0.0672
PDERSkNN	0.04612		FDA_GAMT*	0.06946
BDTG	0.04873		LD*	0.07097
FDA_GAMT*	0.05504		BDTG	0.07177
LD*	0.05682		PDERSkNN	0.07185
MLP	0.1074		SDSS photometric	0.0724

Table 2: Above are the RMS errors, in accending order, computed using Equation 1, of the various TMVA machine learning techniques as compared to the SDSS photometric method. The * next to some methods is to indicate the fact that although the RMS may be low, because it was calculated on the redshift interval of 0 to 1, it is not the best method due to the large number of non-physical redshift determinations.

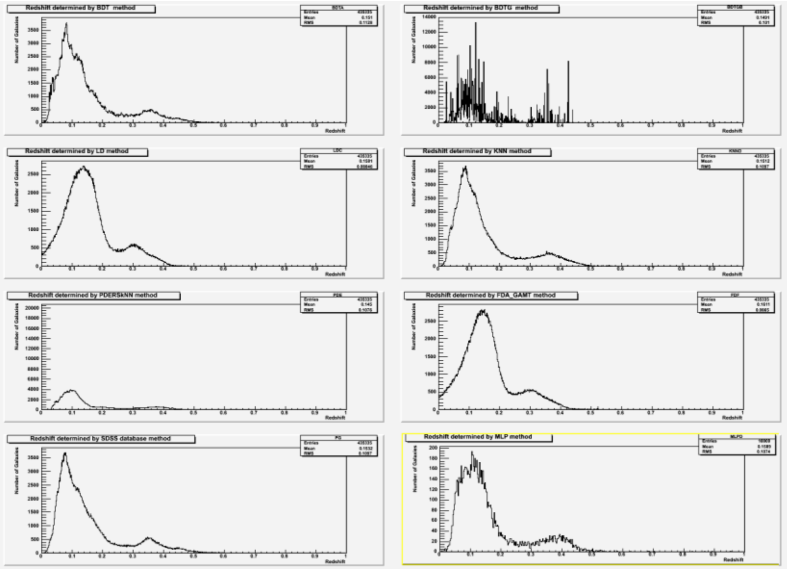


Figure 7: The above figure shows the redshift results of various methods for blended data. Method titles from left to right beginning with top; BDT, BDTG, LD, KNN, PDERSkNN, FDA_GAMT, SDSS photometric, MLP.

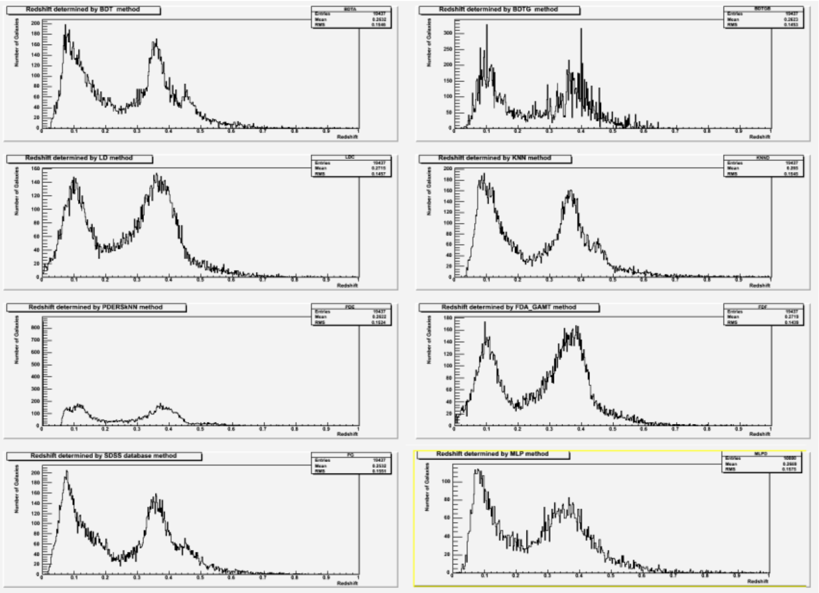


Figure 8: The above figure shows the redshift results of various methods for non-blended data. The method titles from left to right beginning with top; BDT, BDTG, LD, KNN, PDERSkNN, FDA_GAMT, SDSS photometric, MLP.

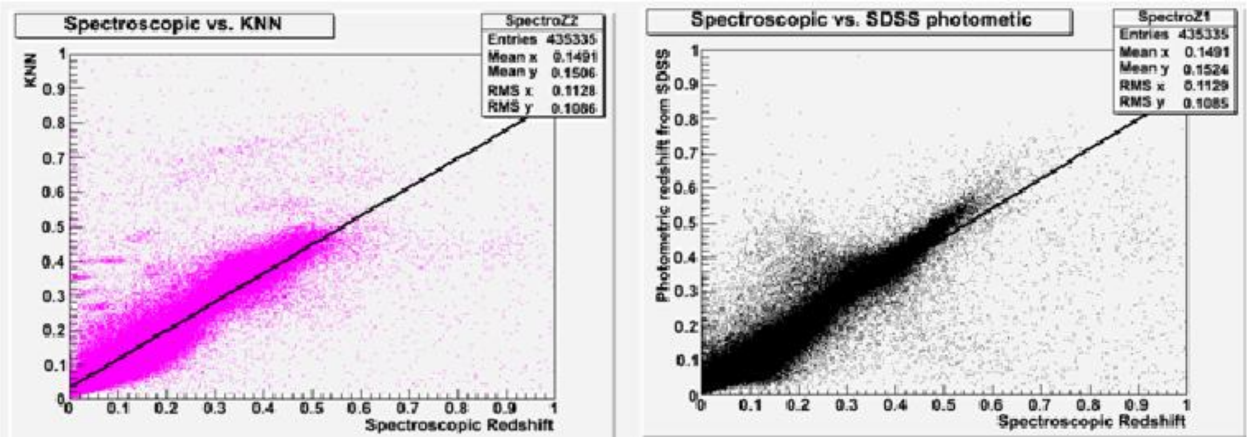


Figure 9: (Left) The “best” redshift determining TMVA method for blended data, KNN, as a function of the spectroscopic redshift.(Right) The SDSS photometric data as a function of spectroscopic redshift.

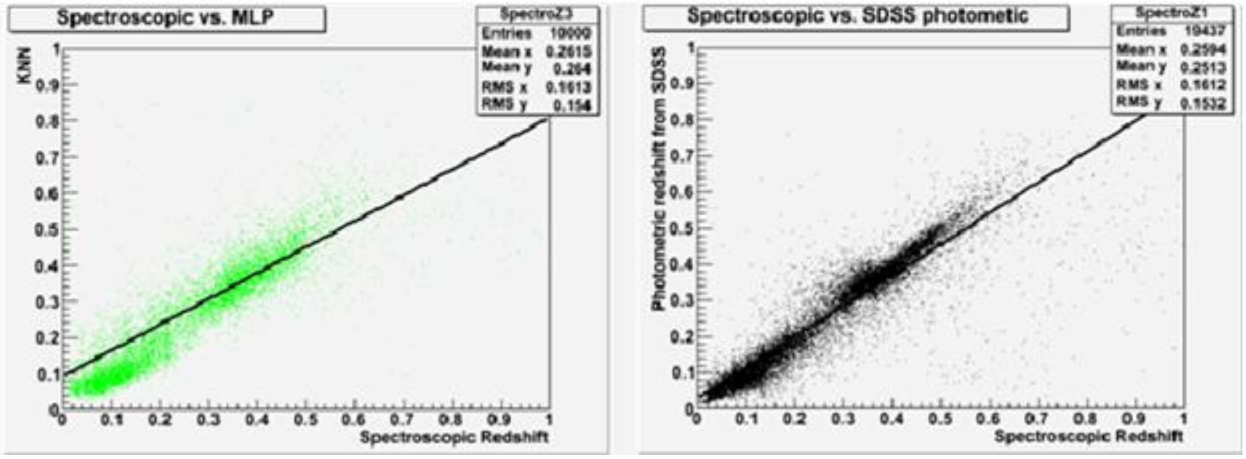


Figure 10: (Left) The “best” redshift determining TMVA method for non-blended data, MLP, as a function of the spectroscopic redshift.(Right) The SDSS photometric data as a function of spectroscopic redshift.