SKY MINING
- APPLICATION TO PHOTOMORPHIC REDSHIFT ESTIMATION

by

Pragyansmita Nayak
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
In Partial fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Computational Sciences and Informatics

Committee:

_____	Dr. Kirk Borne, Dissertation Director
_____	Dr. James E. Gentle, Committee Member
_____	Dr. Janusz Wojtusiak, Committee Member
_____	Dr. Ruixin Yang, Committee Member
_____	Dr. Jie Zhang, Committee Member
_____	Dr. Maria Dworzecka, Director, School of Physics, Astronomy, and Computational Sciences
_____	Dr. Donna M. Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science
_____	Dr. Peggy Agouris, Dean, College of Science
Date: _____	Spring Semester 2015 George Mason University Fairfax, VA

Sky Mining - Application to Photomorphic Redshift Estimation

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Pragyansmita Nayak
Master of Science (thesis incomplete)
IIT Madras, India, 2003
Bachelor of Engineering (Honors)
BITS Pilani, India, 1999

Director: Dr. Kirk Borne, Professor
School of Physics, Astronomy, and Computational Sciences

Spring Semester 2015
George Mason University
Fairfax, VA

Dedication

I dedicate this dissertation to:

My loving and ever supportive parents, Pradeep Kumar and Anuja Paul
My best friend and amazing husband, Rakesh Nayak
And all of my wonderful family, friends and teachers,
Without whom none of this would have been possible.

Acknowledgments

First and foremost, I would like to thank my esteemed advisor, Dr. Kirk Borne, whose input, guidance, understanding and patience, added to my graduate studies experience. Without his impeccable help and valuable time, always shared generously, the completion of this work would have been difficult. His encouraging words and upbeat attitude is one of the prime reasons that I could accomplish the set goals for this work. I will forever cherish the long hours of discussions related to astronomy and data science that has lead me to have a much better and deeper appreciation of both the domains. I look forward to grow further the learning spirit that he has instilled in me as part of my long-term career goals - "my universe is indeed expanding".

I would like to thank the other honorable members of my committee, Dr. James E. Gentle, Dr. Janusz Wojtusiak, Dr. Ruixin Yang, and Dr. Jie Zhang, for their time and assistance that they so generously provided me at all the levels of the research. Their thought-provoking questions helped form different and interesting perspectives of the original research question. They were paramount to keeping me adequately challenged and focused in answering the right questions.

I would like to express my gratitude to my employer CGI Federal Inc. and my colleagues over the years for all the encouragement and well wishes. I would like to thank my managers and team-mates for letting me focus on the classwork, research and defense when I needed to and for always being understanding of my schedule. Their playfully prodding question "When can we call you Dr.?" can finally be answered in the affirmative.

I would also like to thank my parents for the support they have provided me through my entire life. Without their constant push to value education above anything else and to strive for perfection with due diligence and good work ethics, none of this would have been possible. No problem of mine is ever insignificant for them and knowing that I will not get judged for anything is precious.

I cannot thank enough my husband and best friend, Rakesh, for our open debates on every topic under the sun, exchange of ideas, and venting of frustration during my graduate program, which helped enrich the overall experience (hoping he too feels the same!). Our common love for food served as a good distraction in moments of despair when I seriously needed to clear my head and refocus. I appreciate his always being there as my pillar of strength and for always making me smile.

I would like to convey my heartfelt thanks to my family, friends and teachers for enriching my life with so many positive experiences. That is one of the huge factors that pushed me to pursue my dreams knowing very well the safety net that they will provide me in my time of need. Their camaraderie, insight and motivation is invaluable.

Table of Contents

List of Tables

List of Figures

Abstract

SKY MINING - APPLICATION TO PHOTOMORPHIC REDSHIFT ESTIMATION

Pragyansmita Nayak, PhD

George Mason University, 2015

Dissertation Director: Dr. Kirk Borne

The field of astronomy has evolved from the ancient craft of observing the sky. In it's present form, astronomers explore the cosmos not just by observing through the tiny visible window used by our eyes, but also by exploiting the electromagnetic spectrum from radio waves to gamma rays. The domain is undoubtedly at the forefront of data-driven science. The data growth rate is expected to be around 50%-100% per year. This data explosion is attributed largely to the large-scale wide and deep surveys of the different regions of the sky at multiple wavelengths (both ground and space-based surveys).

This dissertation describes the application of machine learning methods to the estimation of galaxy redshifts leveraging such a survey data. Galaxy is a large system of stars held together by mutual gravitation and isolated from similar systems by vast regions of space. Our view of the universe is closely tied to our understanding of galaxy formation. Thus, a better understanding of the relative location of the multitudes of galaxies is crucial. The position of each galaxy can be characterized using three coordinates. Right Ascension (ra) and Declination (dec) are the two coordinates that locate the galaxy in two dimensions on the plane of the sky. It is relatively straightforward to measure them. In contrast, fixing the third coordinate that is the galaxy's distance from the observer along the line of sight (redshift z) is considerably more challenging.

Spectroscopic redshift method gives us accurate and precise measurements of z. However, it is extremely time-intensive and unusable for faint objects. Additionally, the rate at which objects are being identified via photometric surveys far exceeds the rate at which the spectroscopic redshift measurements can keep pace in determining their distance. As the surveys go deeper into the sky, the proportion of faint objects being identified also continues to increase. In order to tackle both these drawbacks increasing in severity every day, alternative method Photometric redshift has been studied in the past. It uses the brightness of the object viewed through various standard filters, each of which lets through a relatively broad spectrum of colors. However, these methods are bound by the degeneracy problem (objects with different color profiles have the same redshift) which leads to low predictive accuracy.

As part of our study, we are looking beyond color attributes to identify other measured attributes as degeneracy resolvers as well as generate estimators that are highly accurate; termed as Photomorphic redshift estimators. The present study investigates the photometric information of the objects such as color and magnitude (= observed flux) and morphology attributes such as shape, size, orientation and concentration in the different wavelengths. The specific type of magnitude used in this study are the PSF, Fiber and Petrosian magnitude. The morphology attributes are the ratio of Fiber to Petrosian magnitude, concentration index and Petrosian radius. All these attributes are in the five bands ugriz of the Sloan Digital Sky Survey (SDSS).

Machine learning techniques based on Naïve Bayes (NB), Bayesian Network (BN) and Generalized Linear Model (GLM) are researched to better understand their applicability, advantages and resulting predictive performance in terms of efficiency and accuracy. Note: The SDSS Data Release (DR) 10 data was used in the executed experiments (total of 700,777 galaxies with forty-five attributes associated with each galaxy).

The significant findings of the present work are as follows:

1. Magnitude and morphology attributes have been found to be successful degeneracy resolvers.

2. Magnitude and morphology attributes have been found to be better redshift estimators than color attributes alone.

3. Naïve Bayes, Bayesian Network and GLM have been found to be viable redshift estimation methods. Attribute selection is an important factor in computational performance.

4. In addition to the redshift estimate, the likelihood distribution of the estimate is even more useful, and my Bayesian Network models provide that information. This is particularly useful in ensemble methods as well as the kernel for mass distribution in the universe.

5. The generated Bayesian Network models can be applied to any of the variables, not just limited to redshift. Example applications include quality analysis and missing value imputation. Different types of Bayesian Network learning algorithms - constraint-based, score-based and hybrid - were investigated in detail.

Chapter 1: Introduction

1.1 Data Avalanche

This is an age of data-avalanche where we are data-rich but information-poor. Large-scale experiments and simulations are producing enormous data volumes involving text, numbers, images, etc. leading to extremely large databases (XLDB). Heterogeneity, inconsistency, incompleteness, timeliness, privacy, visualization and collaboration are some of the key data challenges of the present day [1]. This is resulting in a faster convergence of many fields of academic research in both applied mathematics and computer science, including statistics, databases, artificial intelligence, and machine learning. [2]

Data brokers compile profile information about individuals from a wide variety of online and offline sources that includes email, personal websites, social media posts, census records, retailer systems, Department of Motor Vehicles records and real estate records [3]. For the past decade, e-commerce sites have altered prices based on the web browsing patterns and personal attributes such as location and past purchase history. Factors such as how the user arrived at the e-commerce site and the time of the day of the transaction are of significance as well in profiling potential customers. [4]

ATLAS experiment at Europe's Large Hadron Collider (LHC) is famous for discovering the elusive Higgs boson. It has already generated 140 Petabytes of data, distributed between 100 different computing centers, with most of it concentrated in 10 large computing centers like CERN and U.S. Department of Energy's Brookhaven National Laboratory. [5] In the biological sciences, there is now a well-established tradition of depositing scientific data into a public repository. Furthermore, as technology advances, particularly with the advent of Next Generation Sequencing (NGS), the size and number of experimental datasets available is increasing exponentially. [1]

Apart from the commerical and research sector, the public sector is also cognizant of the imminent value of data. In 2009, the United States (U.S.) Federal Government launched Data.gov as a step toward government transparency and accountability. Data.gov is a warehouse of 124,227 datasets (as of March 2015) covering transportation, economy, health care, education, and human services. These datasets are the source for these type of applications: 409 government APIs, 349 citizen-developed applications, and 140 mobile-oriented applications.

In 2010, the President's Council of Advisors on Science and Technology (PCAST) spelled out a Big Data strategy in its report "Designing a Digital Future: Federally Funded Research and Development in Networking and Information Technology". PCAST is the primary mechanism of the the federal government used to coordinate its unclassified networking and information technology research investments. In 2012, the Big Data Research and Development Initiative was announced which is a $200 million investment involving multiple federal departments and agencies. [6] The other federal government sources of data openly available for access [7] include: USASpending.gov, PaymentAccuracy.gov, Performance.gov, Data.gov and Recovery.gov.

The traditional "legs" (or "pillars") of the scientific method were theory and experimentation. U.S. Presidential Information Technology Advisory Committee issued a report, "Computational Science: Ensuring America's Competitiveness," stating:

> "Together with theory and experimentation, computational science now constitutes the 'third pillar' of scientific inquiry, enabling researchers to build and test models of complex phenomena."

However, this leg has been recently augmented by yet a "fourth paradigm" (or "leg") that refers to the usage of advanced computing capabilities to manipulate and explore massive datasets. A scientific theory is an explanatory framework for a body of natural phenomena. A theory can be thought of as a model of reality at a certain level of abstraction. For a theory to be useful, it should explain existing observations as well as generate predictions for new

observations. What has changed is the scale of computation and the nature of the theories. While previous scientific theories were typically framed as mathematical models, today's theories are often framed as computational models. So science is still carried out as an ongoing interplay between theory and experimentation. The complexity of both, however, has increased to such a degree that they cannot be carried out without computation. [8]

The astronomy domain is also affected by the above and is considered to be at the forefront of data-driven science. This quote rightly elucidates the impact of data-related science on the domain,

> "Astronomy faces a data avalanche. Breakthroughs in telescope, detector, and computer technology allow astronomical instruments to produce terabytes of images and catalogs. These technological developments will fundamentally change the way astronomy is done. These changes will have dramatic effects on the sociology of astronomy itself." [9]

There was 1 PB (Petabyte) of public data electronically accessible as of 2011. This number is growing at approximately 50%-100% per year. For instance, the datasets released by Spitzer Space Telescope and Wide-field Infrared Survey Explorer (WISE) exceeded the total volume of data from approximately thirty-five missions and projects archived prior to their release. [10] This has encouraged the creation of a new branch of astronomy termed Astroinformatics. It includes a set of naturally-related specialties including data organization, data description, astronomical classification taxonomies, astronomical concept ontologies, data mining, machine learning, visualization, and astrostatistics. [11] [12] [13]

[14] is the decadal survey of astronomy and astrophysics is charged to survey the field of space-and ground-based astronomy and astrophysics and to recommend priorities for the most important scientific and technical activities of the decade 2010-2020. The science objectives chosen as a priority by the survey committee for the decade 2012-2021 are searching for the first stars, galaxies, and black holes; seeking nearby habitable planets; and advancing understanding of the fundamental physics of the universe. New optical and infrared survey telescopes on the ground and in space will employ a variety of novel techniques to

investigate the nature of dark energy. These same telescopes will determine the architectures of thousands of planetary systems, observe the explosive demise of stars, and open a new window on the time-variable universe. In the category termed On-the-ground, the Large Synoptic Survey Telescope (LSST) is listed as large-scale in priority order. LSST is a wide-field optical survey telescope that is aimed to address broad questions that range from indicating the nature of dark energy to determining whether there are objects that may collide with Earth. The final source table generated from this effort will consist of approximately 20 trillion rows with over 200 columns of scientific information per source.

Numerous dramatic discoveries have been accomplished through the application of modern technology and human ingenuity to the ancient craft of observing the sky. The universe is explored today not only by observing through the tiny visible window used by our eyes, but also by exploiting the electromagnetic spectrum from radio waves to gamma rays at multiple wavelengths. A great mystery now confronts us: when and how did the first galaxies form out of cold clumps of hydrogen gas and start to shinewhen was our cosmic dawn? Observations and calculations suggest that this phenomenon occurred when the universe was roughly half a billion years old, when light from the first stars was able to ionize the hydrogen gas in the universe from atoms into electrons and protonsa period known as the epoch of reionization. These events lie largely in the realm of theory today, and existing telescopes can barely probe this mysterious era. Over the next decade, we expect this to change. Our view of the universe is closely tied to our understanding of galaxy formation. Thus, the position and properties of the galaxies are crucial information to aid this understanding.

> "Over the next decade it will be a high priority to extend such precision mapping over cosmic time: to have, in effect, a 13-billion-year-long movie that traces the buildup of structure since the universe first became transparent to light. This can be done by using radio telescopes to provide more detailed maps of the cosmic microwave background and to detect the atomic hydrogen gas all the

way back into the dark ages; large spectroscopic surveys in the visible and near-infrared to trace the distribution of galaxies; gravitational lensing to trace the distribution of the dark matter halos; ultraviolet spectroscopic surveys to map out the warm tenuous gas lying in the vast cosmic filaments; and radio Sunyaev-Zeldovich effect and X-ray surveys that reveal the distribution of the hot gas found in groups and clusters of galaxies." [14]

1.2 Photomorphic Redshift and Galaxies

We need to characterise the position of each galaxy using three coordinates. Two of these (commonly RA and DEC) locate the galaxy in two dimensions on the plane of the sky. It is relatively straightforward to achieve a precise measurement of sky position, with accuracies of sub-arcsecond achievable even for ground based observations. In contrast, fixing the third coordinate that is the galaxys distance from the observer along the line of sight is considerably more challenging. [15] [16] This thesis evaluates methods and techniques to estimate the photomorphic redshift of galaxies. The methods are applications of machine learning methods on multi-wavelength data that consists of the color, magnitude (=observed flux) and morphology attributes measured for a given galaxy. This helps overcome the performance and accuracy issues prevelant in the existing methods of spectroscopic redshift and photometric redshift respectively; these are discussed in further detail in Section 1.5 and Section 1.6 respectively.

Galaxy is a large system of stars held together by mutual gravitation and isolated from similar systems by vast regions of space. For instance, Stephan's Quintet as shown in Figure 1.1 is a compact group of galaxies discovered about 130 years ago and located about 280 million light years from Earth. The curved, light blue ridge running down the center of the image shows X-ray data from the Chandra X-ray Observatory. Four of the galaxies (NGC 7317, NGC 7318a, NGC 7318b and NGC 7319) in the group are visible in the optical image (yellow, red, white and blue) from the Canada-France-Hawaii Telescope as shown in

Figure 1.2. It includes a prominent foreground galaxy (NGC 7320) that is not a member of the group and is eight-times closer than the rest of the group. The galaxy NGC 7318b is passing through the core of galaxies at almost 2 million miles per hour, and is thought to be causing the ridge of X-ray emission by generating a shock wave that heats the gas. Stephans Quintet provides a rare opportunity to observe a galaxy group in the process of evolving from an X-ray faint system dominated by spiral galaxies to a more developed system dominated by elliptical galaxies and bright X-ray emission. [17] More importantly, it is an example showing the importance of using redshift as a measure of distance.

Galaxies exhibit a bewildering array of shapes and sizes that are determined largely by the mass of the halo of dark matter surrounding them. A collection of galaxy images can be found at [18]. The different types of galaxies based on their shape are as shown below [19]:

- Elliptical galaxies have very little gas and dust. There are very few young stars in elliptical galaxies because gas and dust are found in the clouds that are the birthplaces of stars. They contain primarily old, red stars known as Population II stars and vary widely in size. Both the largest and the smallest known galaxies are elliptical. Very large elliptical galaxies can reach 300 million light years in diameter. Dwarf ellipticals, which are very common, may contain only 1/100,000th as many stars as the Milky Way.

- Spiral galaxies have two distinct regions. The thin and rapidly rotating disk of the galaxy contains the spiral arms. The disk is a region of star formation and has a great deal of gas and dust. It is dominated by young, blue Population I stars. The nearly spherical and slowly rotating central bulge is devoid of gas and dust and is composed primarily of Population II stars. Type c spiral galaxies have the most gas and dust.

- Lenticular galaxies have a central bulges and disks, but no spiral arms. If the disk is faint, it is easy to mistake a lenticular galaxy for an elliptical galaxy (sometimes called armless spirals). There is a second type of lenticular galaxy called a barred

Figure 1.1: Stephan's Quintet-A Galaxy Collision in Action

lenticular galaxy. Barred lenticular galaxies have bars, much like the barred spirals, and so they are denoted SB0.

- Galaxies that do not fit into either the spiral or elliptical classes are called irregular galaxies. Irregular galaxies, such as M82 (right), have a wide variety of shapes and characteristics. They are frequently the result of collisions between galaxies or gravitational interactions between galaxies.

Note: Quasar (contraction of QUASi-stellAR radio source) is an extremely powerful and distant Active Galactic Nucleus (AGN). They were first identified as being high redshift sources of electromagnetic energy, including radio waves and visible light that were point-like, similar to stars, rather than extended sources similar to galaxies. Quasars are not included in this study.

The most distant object (a gamma-ray burst GRB 090423 identified in early 2010) is at redshift $z = 8.26^{+0.07}_{-0.08} \sim 13.1$ billion years ago - when the universe was 600 million years old.

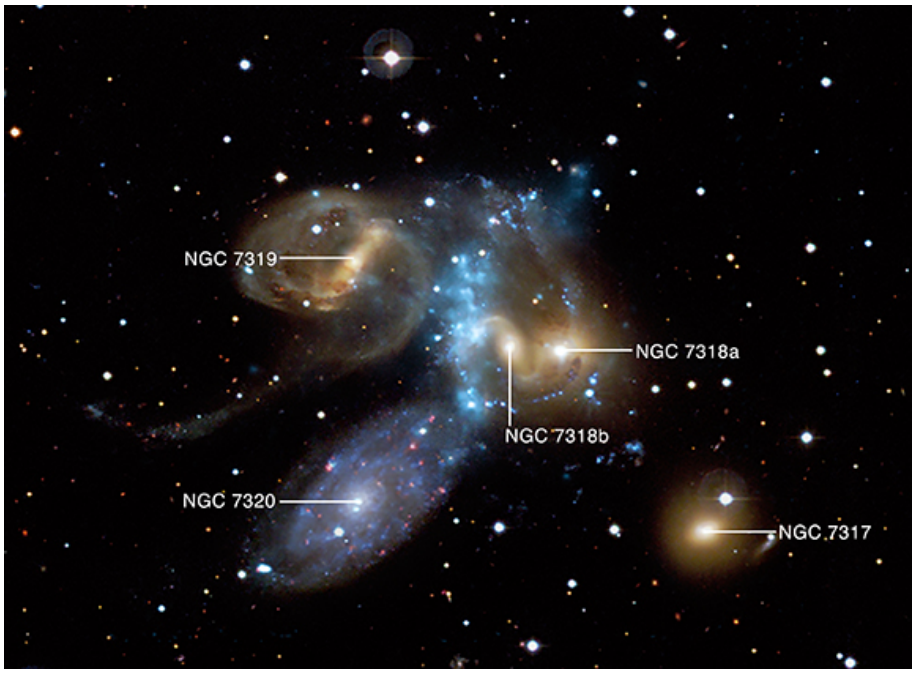


Figure 1.2: Stephan's Quintet - Labeled Galaxies

[20] It was detected by a NASA Explorer program satellite called Swift, and its distance was measured by follow-up observations from telescopes on the ground. The most distant galaxy (z8 GND 5296 identified in 2013) has redshift $z = 7.51$, when the universe was 700 million years old. It was identified using near-infrared (near-IR) spectroscopy of galaxies in the Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey (CANDELS) with the newlycommissioned near-infrared spectrograph MOSFIRE12 on the Keck I 10 meter telescope. [21] Refer Table 1.1 and Table 1.2 for more examples of the most distant and typical nearby objects respectively with their spectroscopically confirmed redshift.

The observable universe contains more than 100 billion galaxies, including our own Milky Way. Although galaxies are made of stars and clouds of gas and dust, ninety percent of the mass of galaxies is the hypothetical dark matter. The existence and properties of dark matter are inferred from its gravitational effects on visible matter, radiation, and the large-scale structure of the universe. A supermassive black hole lies at the center of most or all galaxies. While we have a rather good description of the properties of galaxies in

Table 1.1: Most distant objects

Object Name	**Redshift (z)**	**Object Type**
GRB 090423	8.2	Gamma-Ray Burst
EGS-zs8-1	7.73	Galaxy
z8 GND 5296	7.51	Galaxy
SXDF-NB1006-2	7.215	Galaxy
GN-108036	7.213	Galaxy
BDF-3299	7.109	Galaxy
ULAS J1120+0641	7.085	Quasar
A1703 zD6	7.045	Galaxy

Table 1.2: Typical Redshifts of Nearby Galaxies

Object Name	**Redshift (z)**
UGC 8837	0.00048
NGC 5204	0.00067
UGC 9405	0.00074
Pinwheel Galaxy (M101)	0.00080
NGC 5474	0.00091
NGC 5585	0.00098
NGC 5477	0.00101

the present-day universe, we have far less information about how these properties have changed over the 13.7-billion-year history of the universe. The galaxies we can observe in detail teach us of the complex interplay among the components of normal and dark matter, constrained by the physical laws of the cosmos. A high priority in the coming decade will be to undertake large and detailed surveys of galaxies as they evolve across the wide interval of cosmic timeto have a movie of the lives of galaxies rather than a snapshot.

The surprising discovery in 1998 was that the expansion of the universe is accelerating rather than slowing. This acceleration is attributed to an unknown form of energy called dark energy that accounts for 75 percent of the mass-energy of the universe today. The remainder of the mass-energy is 4.6 percent regular matter and 20 percent dark matter. Recent observations of the microwave background are consistent with the theory that the

universe underwent a burst of inflation when the expansion also accelerated and the scale of the universe that we see today grew from its infinitesimally small beginnings to about the size of a fist. Gravitational waves created at the end of the epoch of inflation can propagate all the way to us and carry information about the behavior of gravity and other forces during the first moments after the big bang. These waves can be detected through the distinctive polarization pattern that they impose on the relic cosmic microwave background radiation. The 13.7-billion-year-old cosmic microwave background is seen in the millimeter band.

1.3 Sky Surveys in Astronomy

The study of the galaxies and the cosmos is enabled by the sky surveys probing the universe in different regions of the sky at different wavelengths, sometimes focusing on specific types of sky objects. The amount of novel data ever collected can be amalgamated to develop a better and complete understanding of the dynamics of the sky bodies. These datasets also need to be quickly analyzed so that interesting phenomena that is fleeting in nature can be identified in a timely manner and investigated in detail before it disappears. Additionally, cross-correlating surveys at different wavelengths is an important aspect of increasing the potential of the individual datasets. The surveys capture vast numbers of images that are accurately calibrated and stored so that they can be easily accessed in future for further research such as motion or unusual behavior on all timescales among other studies.

The technology and software that enable the access and search of these enormous databases are improving all the time and enable astronomers to search the sky systematically for rare and unexpected phenomena. This is a new window to the universe that is opening thanks to the computer revolution and is changing the astronomers way of working. The sky surveys and the ease of access to their captured and measured data has transformed astronomy from a field where taking pictures of the sky was a large part of an astronomer's job to one where the pictures are already in a database, and the astronomer's task is to find interesting objects and phenomena using the database. In other words, post-SDSS, the focus shifts from data collection to data analysis. This is along the lines of the other

fields that are similarly leveraging on the easy availability of data, memory and processing power - web search, content recommendation, computational advertising, healthcare, urban planning, intelligent transportation, environmental modeling, energy saving, computational social sciences, risk analysis [1].

Each probe of cosmology requires an enormous effort to understand both the underlying physics and subtle systematic and observational effects as well as the creation of innovative new statistical techniques to deal with the sheer quantity of data being produced. In engineering terms these projects are often pushing boundaries in terms of space science, optics, detector design, computation and data storage. Cosmological probes are generally complementary, in that each probes a different combination of the cosmological parameters. The strongest constraints on cosmology come from the proper combination of different probes. [22] used CanadaFranceHawaii Telescope Lensing Survey (CFHTLenS) that spans five optical bands for 4.2 million galaxies between redshifts of $0.2 < z < 1.3$ combined with with 7-year Wilkinson Microwave Anisotropy Probe (WMAP7), Baryonic Acoustic Oscillations (BAO): SDSS-III's Baryon Oscillation Spectroscopic Survey (BOSS) and a Hubble Space Telescope distance-ladder prior on the Hubble constant. [23] used Deep Lens Survey (DLS) that is a deep BVRz multi-band imaging survey combined with Wilkinson Microwave Anisotropy Probe 7 year (WMAP7) likelihood data. These combinations break degeneracies between cosmological parameters and allow a level of precision much beyond any individual probe. Most probes are highly correlated as they probe the same underlying physical processes, whether that is the expansion history of the Universe or the perturbations of the large-scale gravitational potential as it evolves with time. [15]

The Web has emerged as a large, distributed data repository of data of different domains and formats - both measured and simualted. Computing techniques have grown from single-server architectures to client-server architectures to a distributed architecture of the present day. Computing and data resources are distributed physically and logically to prevent resource contention and optimal usage. The Virtual Observatory (VO) is a succinct example of an application harnessing the distributed nature of the present-day Internet. Significant

efforts are underway to maximize the utilization of the catalogs and to enable sharing of data and resources for the benefit of improved research and ease of collaboration. Mining these individual or combination of data catalogs (also termed as "multiple data source (MDS) mining problem [24]) significantly improves the decision quality that is more global in nature than ever possible before.

Powerful telescopes that are engineering marvels of the present day are involved in these sky surveys. The telescope can be ground-based or space-based. Telescopes are essentially time machines because light travels across the cosmos at a finite speed. When the telescope captures the most distant objects, it helps us better understand the universe furthest back in time. The space telescopes operated by NASA or as U.S. participation (and operating spectral bands) are shown below in their distinct categories [14]:

- Great Observatories - Chandra (X-ray), Hubble (Infrared, Optical, Ultraviolet), Spitzer (Infrared)
- Mid-size Telescopes - Fermi (Gamma ray), Kepler (Optical),
- Explorers - GALEX (Ultraviolet), RXTE (X-ray), Swift (X-ray), WISE (Infrared)
- Foreign Telescopes with U.S. Participation - Herschel (Infrared), INTEGRAL (Gamma ray), Planck (Radio), Suzaku (X-ray), XMM-Newton (X-ray)

A few examples of ground-based telescopes used specifically for redshift surveys are mentioned next. The first substantial surveys were conducted in the early 1990s, measuring redshifts of several hundred objects at intermediate redshift. The CanadaFrance Redshift Survey (CFRS) measured redshifts of 591 galaxies and was the first to provide a dense, statistical sample of galaxies out to $0.02 < z < 1.2$ [25]. Other similar surveys include the Low-Dispersion Survey Spectrograph (LDSS) survey, the European Southern Observatory (ESO)-Sculptor Survey, the Canadian Network for Observational Cosmology (CNOC) and CNOC2 surveys, and the Hawaii Deep Fields Survey.

Recent surveys investigate using different wavelengths of the electromagnetic spectrum.

Certain examples of the different wavelength usage by the sky surveys include: Two Micron All Sky Survey (2MASS, near-IR), Faint Images of the Radio Sky at Twenty centimeters (FIRST, 20 cm), Runtgen SATellite (ROSAT, X-ray), GALaxy Evolution Explorer (GALEX, UV), Infra-Red Astronomy Satellite (IRAS, mid/far-IR [26]), Green Bank at 6cm (GB6, 6 cm), NRAO VLA Sky Survey (NVSS, 20 cm), and WEsterbork Northern Sky Survey (WENSS, 92 cm). [27] and [28] provide additional detail beyond the scope of the current discussion.

Surveys also focus on specific redshift ranges. Several large-scale digital sky surveys began in late 1990s which included the SDSS, the Two-degree-Field (2dF) [29] and the 2MASS. Pioneering low-redshift surveys such as the 2MASS that covered $z \leq 0.08$ and the SDSS (refer Figure 1.8) and the 2dF (refer Figure 1.5) that covered $z \leq 0.2$ have demonstrated the value of surveying hundreds of thousands of galaxies. These surveys increased by hundreds of times the observational information on the structure, spectral characteristics, and spatial distribution of galaxies in the nearby volume of the universe. These are deep field surveys as well that allow one to study galaxies at the stage of formation and to trace their evolution over billions of years. These surveys have generated vast volumes of disparate type of data termed as catalogs that encapsulate numeric data, spectroscopy data, images, etc. Redshift surveys of more distant galaxies at $z > 1.0$ have progressed relatively slowly as the targeted galaxies are more than 100 times fainter than the targeted galaxies of the low redshift surveys.

The 2MASS [30] infra-red wavelength survey scanned 91% of the sky and measured redshift for 45,000 galaxies with a mean redshift of $z = 0.03$. The northern 2MASS facility began routine operations in 1997 June, and the southern facility in 1998 March. Survey operations were complete for both hemispheres on 2001 February 15. 2MASS has uniformly scanned the entire sky in three near-infrared bands ($J(1.25\mu m), H(1.65\mu m), and Ks(2.17\mu m)$) to detect and characterize point sources brighter than about 1 mJy in each band, with signal-to-noise ratio (SNR) greater than 10, using a pixel size of 2.0". This had achieved an

80,000-fold improvement in sensitivity relative to earlier surveys. 2MASS used two highly-automated 1.3-m telescopes, one at Mt. Hopkins, AZ, and one at CTIO, Chile. Each telescope was equipped with a three-channel camera, each channel consisting of a 256x256 array of HgCdTe detectors, capable of observing the sky simultaneously at J (1.25 microns), H (1.65 microns), and Ks (2.17 microns).

SDSS generated deep, multi-color images covering a little over 35% of the sky using five broad bands (ugriz). This involved over fourteen years of operations (SDSS-I, 2000-2005; SDSS-II, 2005-2008, SDSS III 2008-2014). This generated three-dimensional maps containing more than 930,000 galaxies and more than 120,000 quasars. The final public data release from SDSS-II was completed in October, 2008. SDSS-III, a program of four additional surveys using SDSS facilities, began observations in July 2008, and continued through 2014. The SDSS used a dedicated 2.5-meter telescope at Apache Point Observatory, New Mexico, equipped with two powerful special-purpose instruments. The 120-megapixel camera imaged 1.5 square degrees of sky at a time, about eight times the area of the full moon. A pair of spectrographs fed by optical fibers measured spectra of (and hence distances to) more than 600 galaxies and quasars in a single observation [31]. The SDSS measured many spectra in a single observation: 640 at a time with the SDSS spectrograph (used in SDSS-I, -II comprising Data Releases 1-8 and in the SEGUE surveys) and 1000 with the BOSS spectrograph (starting with Data Release 9). [32] After the spectra are output from the spectroscopic pipeline, derived quantities were computed by applying stellar population models to derive stellar masses, emission-line fluxes and equivalent widths, and gas kinematics and stellar velocity dispersions. [33]

Deep Extragalactic Evolutionary Probe 2 (DEEP2) Galaxy Redshift Survey is the densest and largest high-precision redshift survey of faint galaxies in the redshift range $0.7 < z < 1.4$. It utilized the DEep Imaging Multi-Object Spectrograph (DEIMOS) spectroscopic on the Keck II telescope (the world's largest optical telescope as of 2005). Objects with $z \leq 0.7$ were readily identifiable using BRI photometry and rejected in three of the

four DEEP2 fields. This allowed galaxies with $z > 0.7$ to be targeted $\sim$2.5 times more efficiently than in a purely magnitude-limited sample. Approximately 60% of eligible targets are chosen for spectroscopy, yielding nearly 53,000 spectra and more than 38,000 reliable redshift measurements (as part of Data Release 4). [34] describes a new catalog that supplements the existing DEEP2 Galaxy Redshift Survey photometric and spectroscopic catalogs with ugriz photometry from the Canada-France-Hawaii Legacy Survey (CFHTLS) and the SDSS. This catalog consists of $\sim$27,000 objects with full ugriz photometry as well as robust spectroscopic redshift measurements. This catalog can be used as a testbed for future photo-z studies, including tests of algorithms for upcoming surveys such as LSST.

Deep fields relate to projects devoted to a detailed exploration of relatively small sky areas. Some of the deep field surveys in the last few years include William Herschel Deep Field (WHDF 1994-1997), Hubble Deep Field North (HDF-N) and South (HDF-S, early 1990s), Chandra Deep Field (CDF, 2005), FORS Deep Field (FDF, 1999-2000), Subaru Deep Field (SDF, 1999-), Subaru/XMM-Newton Deep Survey (SXDS, 2003), COMBO-17 (Classifying Objects by Medium-Band Observations in 17 filters - five broadbands of UBVRI, and 12 medium-band covering the spectral range $3500\mathring{A} - 9300\mathring{A}$), Great Observatories Origins Deep Survey (GOODS, 2002). Hubble Ultra Deep Field (HUDF, 2003-2004), among others. [16]

The Gemini Deep Deep Survey (GDDS) [35] is an infrared-selected ultradeep ($K < 20.6mag, I < 24.5mag$) redshift survey targeting galaxies in the redshift desert (specifically, $1.4 \lesssim z \lesssim 2.5$) [36]. The so-called "redshift desert" corresponds to an era when the universe was between one-third and one-half its present age - that is, when the universe was only 3-6 billion years old ($1 < z < 2$). Relatively little is known of the galaxies in this period due to a combination of intrinsically faint galaxy spectra and contamination by atmospheric emissions in the spectral window used to make these observations. The redshift desert is critically important because it is thought to correspond to the period of peak galaxy building. A sample of 309 spectra, along with redshifts, identifications of spectral features, and

photometry was obtained. This makes the GDDS the largest and most complete infrared-selected survey probing the redshift desert. The seven-band ($VRIzJHK_s$) photometry is taken from the Las Campanas Infrared Survey. The infrared selection means that the GDDS is observing not only star-forming galaxies, as in most high-redshift galaxy surveys, but also quiescent evolved galaxies. The median redshift of the whole GDDS sample is $z = 1.1$. It is designed to bridge the gap between landmark surveys of highly complete samples at $z < 1$ and UV-selected surveys at higher redshift.

Dark Energy Survey (DES) and LSST will extend wide field imaging well beyond the depth of SDSS. [37] The DES and LSST photometry (together with VISTA JHK photometry) yields not only fluxes, colors, and photometric redshifts but also galaxy image shapes and surface brightnesses. All of this information can be exploited to select a sample of galaxies that satisfy the joint requirements of large redshift range, adequate volume sampling, and control over any bias introduced due to sample selection or redshift failures. In practice, we expect to use galaxy flux, color (and photo-z), and surface-brightness to optimize the redshift distribution and galaxy types of the survey. The DES started in August of 2013 and will continue operations for the next five years. [37] [38]

Stanford Linear Accelerator Center (SLAC) National Accelerator Laboratory at Palo Alto, CA is currently building one of the world's largest databases. Scheduled to go live in 2020, the LSST will feature a 8.4-meter optical telescope with 3.2-gigapixel camera sited in Chile capturing ultra-high-resolution images of the sky every 15 seconds, every night, for at least 10 years. LSST will image the entire available sky every 3 nights. Ultimately, the system will store more than 200 Petabytes of image data, but that is barely a fraction of the data that will actually pass through the camera. The system will extract critical data from the images in real time, then simply discard the source images. The final source table consisting of object parameters extracted from images is expected to be about 40 Petabytes. To get a perspective on the amount of data analysis involved, [39] mentions:

> "The new telescope will take high-resolution images that cover an area 7 times the width of the full moon. To do that they build a phone camera the size of a

VW Beetle, containing 3.2 billion pixels. To view one image at full resolution would require 1500 HD TV screens."

The LSST will provide detailed data of the astronomical objects in the sky, more than any of the surveys of the past or ongoing surveys. The LSST will provide the three-dimensional maps of the mass distribution in the Universe, in addition to the traditional images of luminous stars and galaxies. These mass maps can be used to better understand the nature of the newly discovered and utterly mysterious Dark Energy that is driving the accelerating expansion of the Universe. SDSS and 2MASS are two of the potential catalogs from over 15,000 available catalogs whose data or combination of data can be used to design and evaluate techniques and methodologies that can be applied to test methods meant to be applied to LSST data (when it becomes operational).

Virtual observatories have enabled online access to the data from most of the above mentioned surveys. The datasets are publicly accessible via the Internet and a common interface. The International Virtual Observatory Alliance (IVOA) was formed in June 2002 with a mission to "facilitate the international coordination and collaboration necessary for the development and deployment of the tools, systems and organizational structures necessary to enable the international utilization of astronomical archives as an integrated and inter-operating virtual observatory." The IVOA now comprises 16 VO projects from Armenia, Australia, Canada, China, Europe, France, Germany, Hungary, India, Italy, Japan, Korea, Russia, Spain, the United Kingdom, and the United States [40]. US Virtual Astronomical Observatory (VAO) (known earlier as the National Virtual Observatory (NVO)) is the US-based virtual observatory project [41]. VO provides a virtual sky - a single window to the different survey data, combinations of survey data, related publications and other relevant information through uniform data services, compute services and registry services. It enables astronomers and the general public to preview and collaborate with convenience. VO addresses the access need, however, efficient processing of these data inorder to maximize their potential still remains an open computational problem. The properties of quasars and galaxies [42], the mass functions of galaxies, the properties of AGN galaxies detected by

SDSS and ROSAT, the colors of elliptical galaxies and the optical properties of SDSS galaxies are some of the studies that has used data from a single survey or from a combination of multiple surveys.

1.4 Surveys and Catalogs

Databases are enablers of scientific discovery. The databases generated by the sky surveys mentioned in Section 1.3 are known as catalogs. Since the surveys target a specific cause, each catalog has it's own specific purpose - interim or final - in the exhaustive data capture, measurement and collection process. There are approximately 15,000 catalogs of different sizes available via VO and Centre de Donnes astronomiques de Strasbourg (CDS) [43]. The catalogs were previously maintained by the Astronomical Data Center of NASA. CDS maintains the collection of catalogs. Table 1.3 lists the distribution of catalogs based on the type of catalog.

Table 1.3: List of catalogs by Type

Catalogue Type	**Number Available**
Astrometric Data	264
Photometric Data	262
Spectroscopic Data	224
Cross-Identifications	27
Combined data	114
Miscellaneous	105
Non-stellar Objects	213
Radio and Far-IR data	84
High-Energy data	30

VO has enabled this new astronomy that involves the multiple catalogs holding data of millions of sky objects covering data of different wavelengths and object characteristics. Data from multiple catalogs with correlated data can be combined to get a much larger and more complete data-set. Easily joining different catalogs requires that the catalogs should

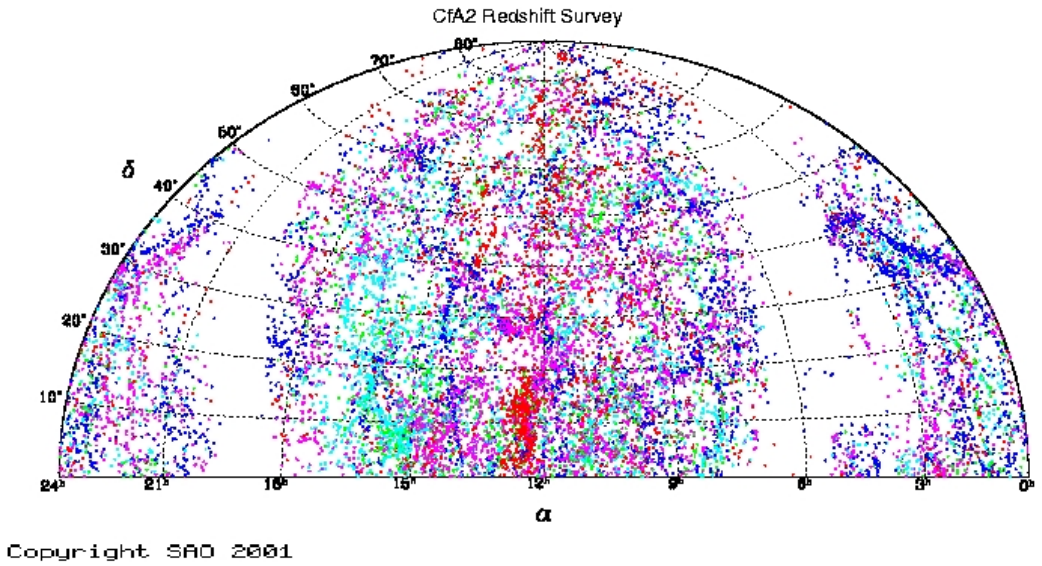


Figure 1.3: CfA2 Northern Sky Redshift Survey Map

be in a standard format so that automated processes can perform the join. The VOTable and FITS format are the two standard format that are largely being used in the astronomy community. VOTable is a standard metadata-rich XML data-interchange format for tabular data across the services of the VO [44]. FITS stands for 'Flexible Image Transport System' and is the standard astronomical data format endorsed by both NASA and the IAU. FITS is much more than an image format and is primarily designed to store scientific data sets consisting of multi-dimensional arrays and two-dimensional tables containing rows and columns of data. Using data across catalogs to solve science problems truly harnesses the potential of the highly distributed and heterogeneous data sources made easily accessible via the uniform single-point interface provided by the VO. Table 1.4 provides a short list of available catalogs in chronological order. It highlights the increasing number of new objects being identified with newer surveys using more powerful telescopes than before their time.

The term Digitized Sky Survey was originally used to refer to a digital version of an all-sky photographic atlas in 1994. For the northern sky, the National Geographic Society - Palomar Observatory Sky Survey provided almost all of the source data. For the southern

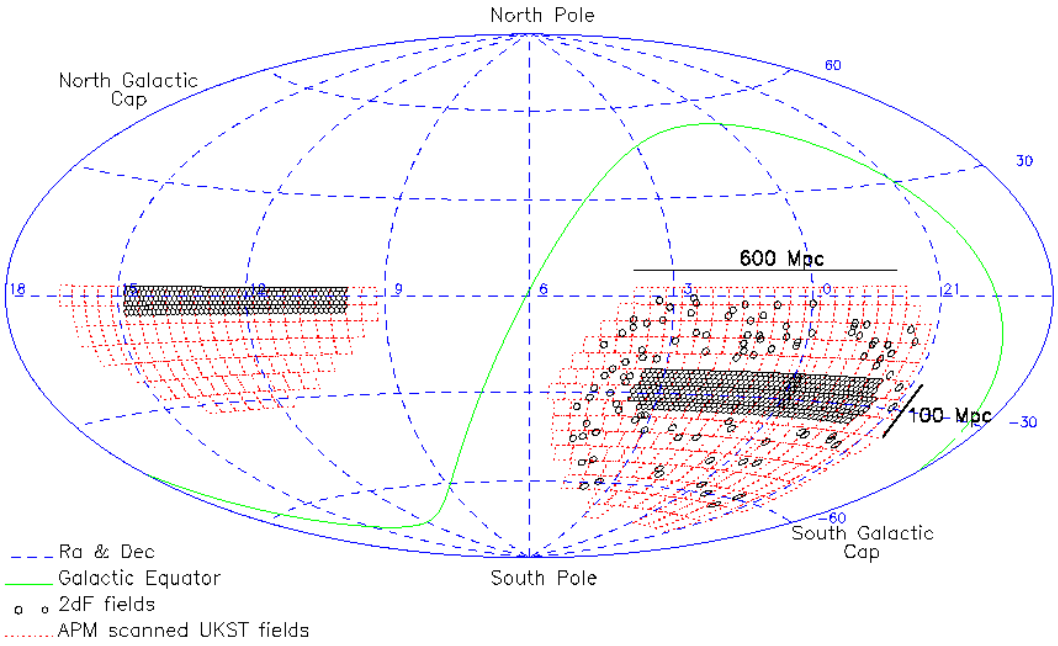


Figure 1.4: 2dF Southern Sky Redshift Survey Map

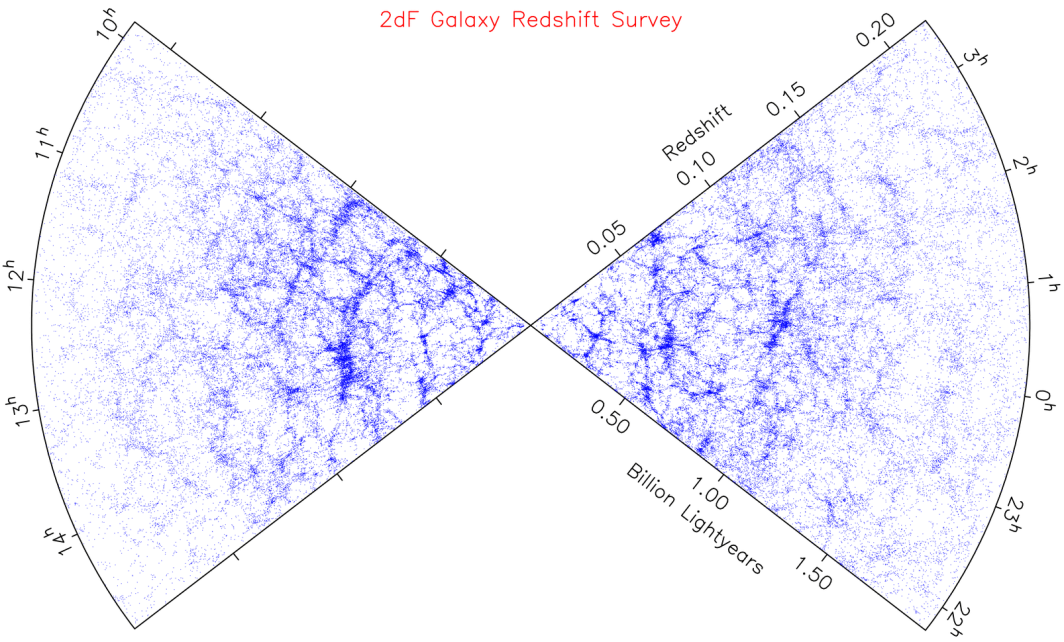


Figure 1.5: 2dF Southern Sky Redshift Survey Map

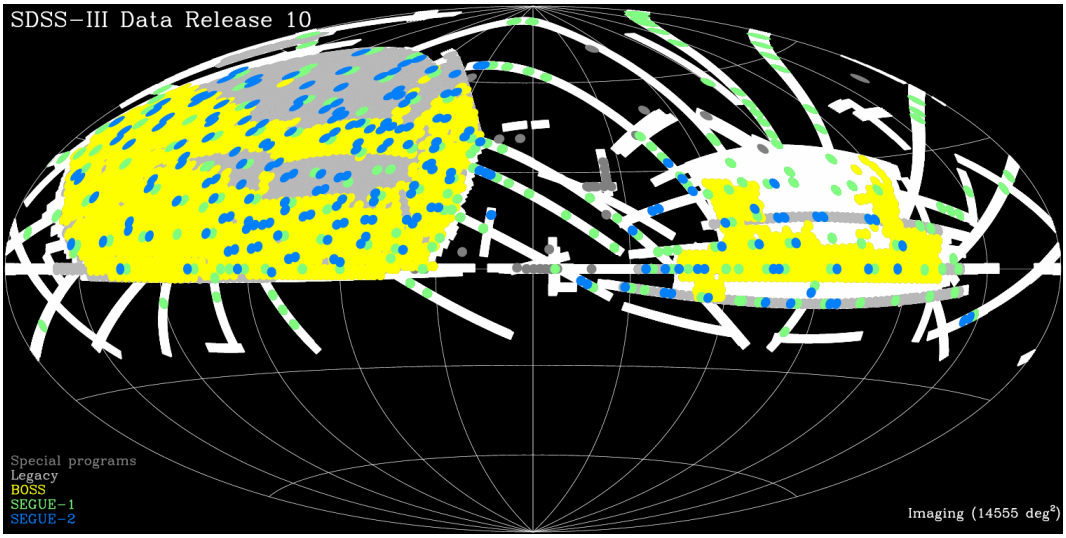


Figure 1.6: SDSS Northern Sky Redshift Survey Map DR10

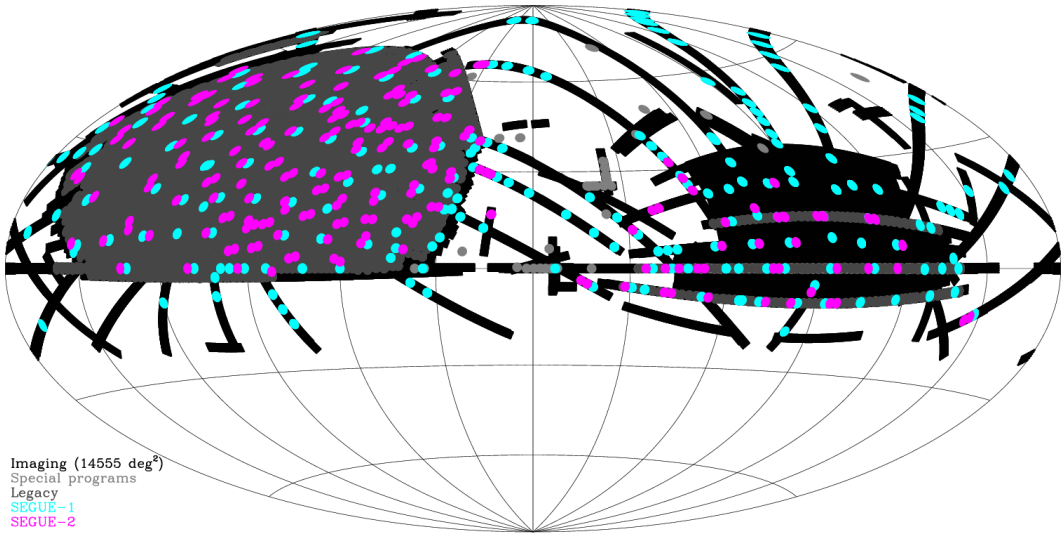


Figure 1.7: SDSS Northern Sky Redshift Survey Map DR12

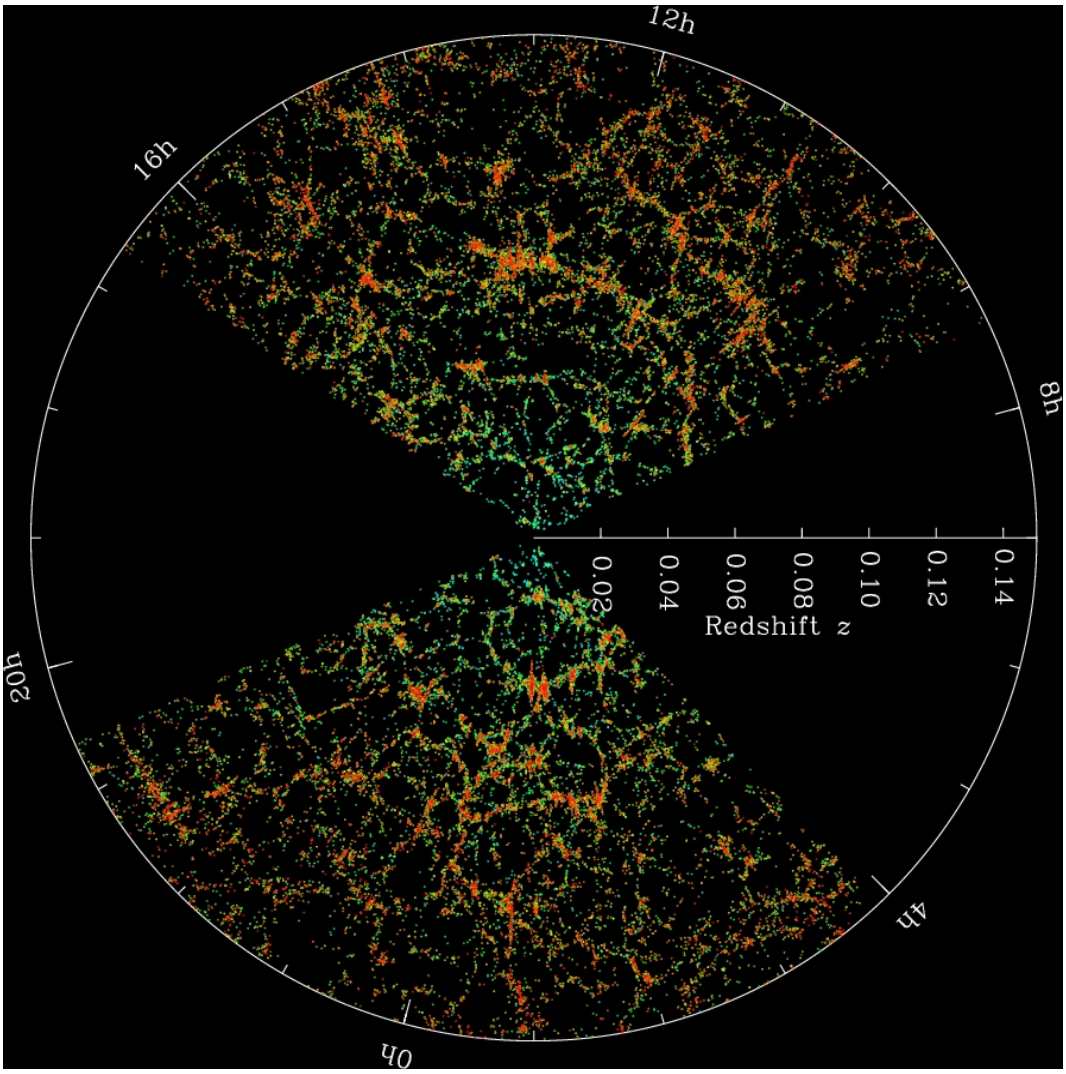


Figure 1.8: SDSS Distribution of Galaxies where color corresponds to galaxy luminosity

Table 1.4: List of catalogs

Catalogue	**Year**	**Objects**
Timocharis and Aristyllus	290BC	∼100
Hipparchus	130BC	1080
Almagest (Ptolemy)	140	1208
Ulugh Begh	1437	∼1000
Tycho Brahe	1598	1000
Historia Coelestis Britannica (Flamsteed)	1725	2935
Catalogue of 9766 Stars (Lacaille)	1847	9766
Bonner Durchmusterung (Argelander)	1862	325,036
Cape Photographic Durchmusterung (Gill et. al.)	1900	454,877
Cordoba Durchmusterung (Thome)	1932	613,959
Center for Astrophysics (CfA) Survey	1977-1982	**First Redshift Survey**
CfA2	1985-1995	18,000 Galaxies [45] *Refer Figure 1.3*
The HST Guide Star Catalogue	1990	$\approx 2 \times 10^7$
ROE/NRL Object Catalogue Southern Sky	1992	$\approx 5 \times 10^8$
Two-degree-Field (2dF)	1997-2002	382,323 incl. 232,155 Galaxies [29],[46] *Refer Figure 1.4*
Sloan Digital Sky Survey (SDSS)	2000-2014	$\approx$ 469 million incl. 2,401,952 Galaxies (DR12) [47] *Refer Figure 1.8; Refer DR10 in Figure 1.6; DR12 in Figure 1.7*
Two Micron All Sky Survey (2MASS)	1997-2001	470 million [30]
The SuperCOSMOS Southern Sky Survey	2001	$\approx 10^9$ [48]
Galaxy Evolution Explorer (GALEX)	2003	More than 10^8
Visible and Infrared Survey Telescope for Astronomy (VISTA)	2010	$\approx 10^{10}$
Panoramic Survey Telescope And Rapid Response System (PanSTARRS)	2010-	Billions of Stars
DEEP2	2002-2008	52,989 [49]
Large Synoptic Survey Telescope (LSST)	2020-	50 billion

sky, the Southern Sky Atlas and its Equatorial Extension (together known as the SERC-J) and the southern Galactic Plane survey (SERC-V), from the UK Schmidt Telescope at Anglo-Australian Observatory, were used. The publication of a digital version of these photographic collections has subsequently become known as the First Generation DSS. [50]

Microsoft World Wide Telescope (WWT) is an integrated amalgam of data and images

from the sky surveys, that is, the 2MASS, the SDSS, the Hubble Space Telescope and the Chandra X-ray Observatory. The images are provided across the multiple wavelengths of the electromagnetic spectrum. The application is coded in Microsofts C# .NET with Microsofts Visual Experience Engine. WWT is a combination of software and Web services that allows users to pan smoothly across the sky while accessing terabytes of images and data from multiple sources. The basic layer of the northern sky in both the WWT and Google Sky is comprised of sky surveys conducted over the years at Palomar Observatory in California, while the southern sky is derived from surveys at the Anglo-Australian Observatory in Australia.

Google Sky also uses multiple information layers that can be selected under its sky database, including constellation figures, the current positions of the planets, and a backyard astronomy layer that labels stars, constellations, and celestial objects. Google Earth was created to project imagery onto the surface of a sphere. For Google Sky, that perspective is reversed by using the same infrastructure to project images of the sky onto the inside of a sphere, creating a realistic representation of the celestial vault. However, due to this latitude and longitude-akin projection, the stars in the original images were significantly distorted between seven and eight degrees of both celestial poles. The stars in those polar regions are obviously not as sharp as the other parts of Google Skys sky (they exhibit a decided radial stretch from the pole outward), but they are properly scaled and their colors are based on real color data. Keyhole Markup Language (KML) which is an XML-based language for displaying geographic data and visualizations for web-based 3D browsers. KML files display celestial objects as well as annotated data files in Google Sky. Users can add their own content by converting it into a KML file. [50]

This multi-wavelength analysis of the astronomy objects enable us to discover significant trends and patterns, including redshift distribution, from the analysis of statistically rich and unbiased image/numeric databases. Larger data-set needs automated and intelligent analysis techniques. It is no longer amenable for manual analysis. Two important characteristics of the algorithms processing these large data-sets (giga-, tera- and peta-scale)

is that they should be distributed and incremental in nature. Distributed algorithms will largely address the vast communication and data-storage needs. Incremental algorithms will reuse and build on past findings. It will avoid re-processing from-scratch every time new data releases are available thereby losing hours and days of processing done with the older data releases. It will also make it easier to add new catalogs to the data analysis thereby facilitating the increase of the wavelength-coverage of the analysis. The algorithms should be able to reuse past findings in a straight-forward manner avoiding any repeat work.

1.5 Spectroscopic Redshift

Einstein's General Theory of Relativity of 1916 was an expansion of his first theory of relativity - Special Theory of 1905. This studies the effect of gravitation on the shape of space and the flow of time. This theory establishes the relationship between matter, space, time and gravity and it governs cosmology's view of the universe. This had brought into question if the universe was expanding or collapsing in contrast to the understanding of those times that the universe is static. Einstein added the cosmological constant to his equations to make his calculations consistent with a static universe. He later termed this as "the greater blunder of my life". A new study confirms that the cosmological constant is the best fit for dark energy, and offers the most precise and accurate estimate yet of its value. The finding comes from a measurement of the universe's geometry that suggests that our universe is flat instead of being spherical or curved. [51]

In the early 20th century, astronomer Vesto Slipher observed in his experiments that the absorption lines in the spectra of most spiral galaxies had longer wavelengths than those observed from stationary objects. Assuming that the redshift was caused by the Doppler shift, he concluded that the red-shifted galaxies were all moving away from us. In 1920's, Hubble observed that the galaxies were receding from us at a velocity proportional to their distance. This led to the Hubble's law - the more distant the galaxy, the greater its redshift and therefore, the higher the velocity. The expansion is believed today to be a result of a "Big Bang" which occurred nearly 14 billion years ago. The Hubble constant (unit as

kilometers per second per million light years) in Equation (1.1) quantifies the universe's rate of expansion. The expansion is usually treated as an analogy of raisin bread with the galaxies similar to raisins that are moving away with space as the space expands rather than moving away through space [52]. With Doppler effect at play, redshift is observed and used as a measure of the distance. A galaxy at a redshift of three, for example, corresponds to a distance of about 12 billion light years.

$$H_0 = \frac{velocity}{distance} \tag{1.1}$$

The Doppler effect occurs when the observer is moving relative to the source of the waves. The received frequency is increased (compared to the emitted frequency) during the approach, it is identical at the instant of passing by and it is decreased as they are moving away. Equation (1.2) shows the relation between the observed frequency f and emitted frequency f_0. Doppler effect is of great use in astronomy and results in either redshift or blueshift. It is used to measure the radial velocity - speed at which stars and galaxies are approaching or receding away from Earth. Since blue light has a higher frequency than red light, spectral lines (aka. spectroscopic measurements) of an approaching astronomical light source exhibit a blueshift. While those of a receding astronomical light source exhibits a redshift.

$$f = (\frac{v + v_r}{v + v_s})f_0 \tag{1.2}$$

Electromagnetic radiation is classified into types according to the frequency of the wave. In order of increasing frequency and thus, decreasing wavelength i.e., Speed of the wave = Wavelength * Frequency (also shown in Figure 1.9 [53], [54]): radio waves, microwaves, infrared radiation, visible light(red through blue), ultraviolet radiation, X-rays and gamma rays. Redshift occurs when electromagnetic radiation - usually visible light - emitted or reflected by an object is shifted towards the red end of the electromagnetic spectrum due

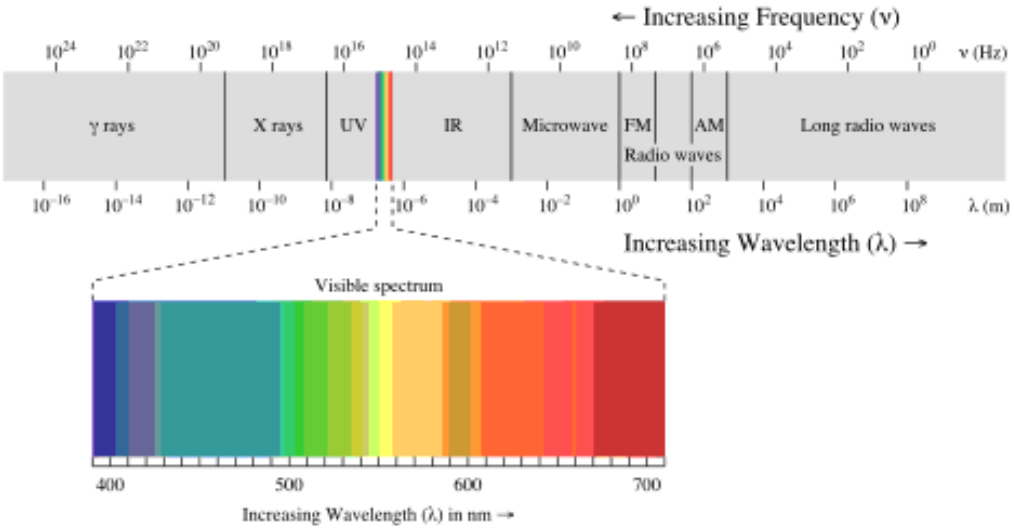


Figure 1.9: Electromagnetic Spectrum

to the Doppler effect. It is associated with an increase in the wavelength of electromagnetic radiation received by a detector compared to the wavelength emitted by the source. Conversely, a decrease in wavelength is called blue shift.

Even though Hubble's measurements were made almost a century ago, we have only measured the velocities and distances of a small fraction of the galaxies we can see, and so we have only small amount of data on whether the rate of expansion is the same in all places and in all directions in the universe. The redshift distance relation thus continues to help us map the universe in space and time. With fast, efficient and robust redshift predictors, we can develop a more timely, detailed and accurate understanding of the structure and the dynamics of the universe. There are presently two main ways to measure redshift - Spectroscopic redshift and Photometric redshift.

Most large telescopes have spectrometers, which are used to measure the velocities of astronomical objects from the Doppler shift of their spectral lines. A spectrum (the plural is "spectra") measures how much light an object emits as a function of wavelength. The spectra of stars and galaxies almost always show a series of discrete lines that form when certain atoms or molecules emit or absorb light. These lines are unique for each

element and always have the same spacing. These "spectral emission and absorption lines" always appear at the same wavelengths, so they make a convenient marker for redshift or blueshift. Spectroscopy technique is used to observe the frequency (or wavelength) of characteristic spectral lines to see how far the lines were shifted from their usual position. If the astronomers look at a galaxy and see one line at a longer wavelength than it would be on Earth, they would know that the galaxy was red-shifted and moving away from us. If they see the same lines at shorter wavelengths, they would know that the galaxy was blue-shifted and moving toward us.

Spectroscopic redshift measurements give us accurate and precise readings. Spectroscopic redshifts are important for calibrating photometric and photomorphic redshifts estimation methods. However, they are quite time intensive. For example, at $z \approx 0.5$, a $200kms^{-1}$ quality redshift requires the better part of a night's observing on a 4-m class telescope [55]. The rate at which objects are being identified via photometric surveys far exceeds the rate at which the spectroscopic redshift measurements can keep pace in determining their distance. We need methods and techniques that can take advantage of the color data and measurements from the wavelengths such as radio, infrared, etc as well as morphology-based properties to determine the photomorphic redshift associated with the object.

1.6 Photometric Redshift

With the availability of the powerful telescopes [1] [2] [3] that can image the vast realms of the universe, and the VO tools that can render the data easily accessible, it has literally opened up a whole new world to explore and study. This helps us to develop a better understanding of the structure and dynamics of the universe. [58] It enables identification of the new yet-unseen astronomical bodies similar to how the SDSS discovered the most distant quasars, sub-stellar objects and celestial dwarfs. Equally or even more importantly, it also helps answer science questions to a large extent that are mathematically intractable. The tough questions can be answered with even higher confidence using statistical tests and machine learning techniques. The redshift estimation is definitely one such problem that can benefit from these peta-bytes of captured survey data and can be estimated by the application of the novel machine learning techniques on the same.

There are several photometric surveys similar to the SDSS that have been implemented in the past decade or are in the process of being implemented using different filter systems. Other examples include University of British ColumbiaNASA (UBC-NASA) [59], Calar Alto Deep Imaging Survey (CADIS) [60], Classifying Objects by Medium-Band Observations in 17 Filters (COMBO-17) [61], Advance Large Homogeneous Area Medium Band Redshift Astronomical (ALHAMBRA) [62], DES [37], LSST and Panoramic Survey Telescope And

[1] The invention of a new telescopic system, the wide-field reflector, by Estonian optician Bernhard Schmidt (1879-1935) was one of the most important stages in the development of the observational technique. A correcting plate mounted in front of a reflector's objective allows compensating for most aberrations of the main mirror. The best known Schmidt telescope is the 1.2-m installed at the Mount Palomar Observatory in California in the 1950s. [16]

[2] Without adaptive optics, stars and galaxies viewed at high magnification will dance, distort and blur like stones seen at the bottom of a stream. With adaptive optics, they will remain steady and sharp, allowing telescopes on the ground to routinely equal or exceed the clarity obtained by NASA's Hubble Space Telescope. This capability has allowed current-generation telescopes to carry out high-resolution studies of objects ranging from moons in the outer Solar System to stars at the centre of the Milky Way. And now it is enabling the construction of telescopes measuring 2040 metres across, as much as four times the diameter and 16 times the light-gathering power of any now in existence. [56]

[3] The LSST uses a novel, three-mirror, modified Paul-Baker design, with an 8.4-meter primary mirror, a 3.4-m secondary, and a 5.0-m tertiary, along with three refractive corrector lenses to produce a flat focal plane with a field of view of 9.6 square degrees. In order to maintain image quality during operation, the deformations and rigid body motions of the three large mirrors must be actively controlled, using a set of curvature wavefront sensors located in the four corners of the LSST camera focal plane, to minimize optical aberrations, which arise primarily from forces due to gravity and thermal expansion. [57]

Rapid Response System (Pan-STARRS) [63] among others. These surveys represent powerful alternatives to the deep spectroscopic surveys such as DEEP2 or BOSS. Photometric surveys suffice for those scientific goals which only require limited redshift accuracy and low resolution spectral information. [64]. Big data catalogs are expected from large photometric surveys such as the LSST [4], the EUCLID [5] or the Wide-Field Survey Infrared Telescope [6]. They increase the urgency of the need for fast and reliable photo-z methods that are capable of processing large volumes of data in minutes to days instead of years. The improved methods will facilitate higher levels of analysis and model refinement for downstream data products. [65]

Photometric redshift uses the brightness of the object viewed through various standard filters, each of which lets through a relatively broad spectrum of colors to determine the redshift. The observed color of the galaxy depends on its redshift since the expansion of the universe causes farther galaxies to be more redshifted and they are on average younger than the nearer ones. The redshift estimate is then mapped to the distance of the observed object through Hubble's law. The technique relies upon the spectrum of radiation being emitted by the object having strong features that can be detected by the relatively crude filters. The technique was developed in the 1960s, but was largely replaced in the 1970s and 1980s by spectroscopic redshifts. The technique has gained popularity and is increasingly becoming an important technique as a result of large sky surveys conducted in the late 1990s and early 2000s which have detected a large number of high-redshift objects.

Photometric redshifts were originally determined by calculating the expected observed data from a known emission spectrum at a range of redshifts. The technique of photometric spectroscopy provides a method to determine at least qualitative characterization of a redshift. This is particularly important in the absence of sufficient telescope time to determine a spectroscopic redshift for each object. The photometric information is thus, useful to select spectroscopic targets for follow-up detailed investigation and thereby, optimize the

[4]http://www.lsst.org/lsst
[5]http://sci.esa.int/euclid
[6]http://wfirst.gsfc.nasa.gov

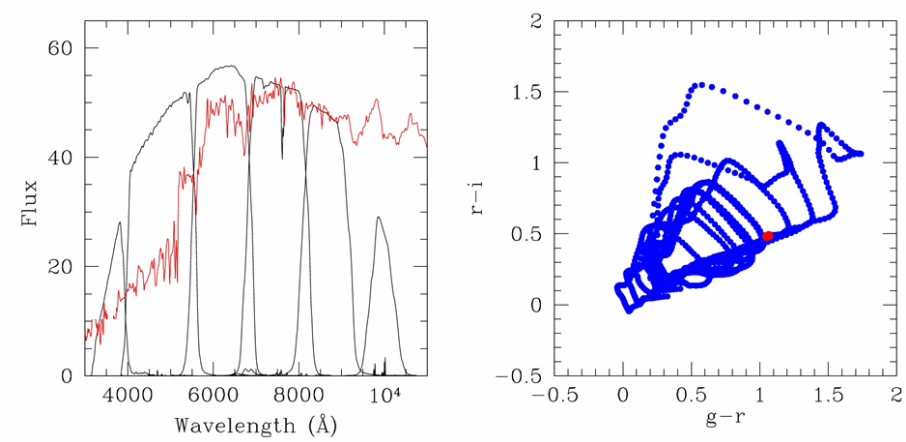


Figure 1.10: Photometric Degeneracy [66]

survey science. In time-scale, while color data can be captured in an exposure of minutes, spectroscopy will take hours instead. Spectroscopy cannot be used for very faint objects where photometric redshift will need to be used.

Photometric redshift techniques traditionally involve plotting lines of constant redshift and varying spectral type known as iso-z lines using magnitude. These plots are known as color-color diagram. The photometric redshift is estimated from the diagram based on the position of an object determined by it's colors. The drawback of this approach has been the degeneracies associated with it. Objects with same color profiles can have different redshift based on it's galaxy type. Refer Figure 1.10 - the degeneracies will occur where the different iso-z lines intersect. In other words, when an algorithm is based on the color-color diagram or related logic, since the algorithm is using color information alone, different photometric redshifts can be estimated if the data occurs at one of the many possible intersection. This severely impacts the predictive accuracy.

An "ideal" photometric redshift estimator should determine the return value as close as possible to the spectroscopic redshift. For example, the photometric redshift estimator should be able to determine that the three objects tagged as *lensed galaxy* (refer Figure 1.12) have exactly same photometric redshift since they are in fact, the same sky object. [7].

[7]This occurs due to gravitational lensing (refer Figure 1.11)

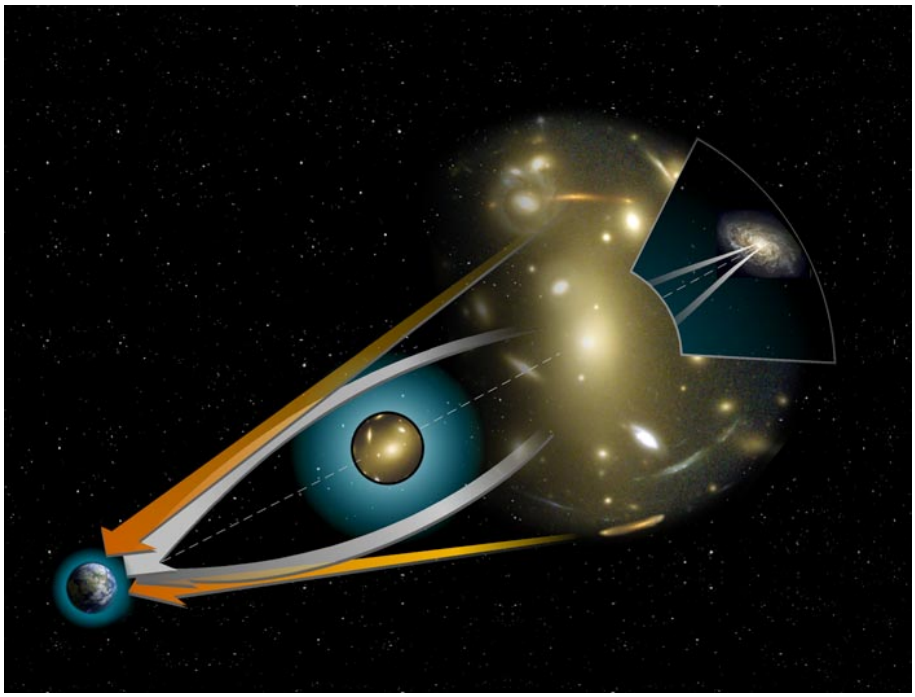


Figure 1.11: Gravitational Lensing Process

Due to above degeneracy issue arising when using only color information, we need to move beyond and consider other measured or calculated attributes to use as degeneracy resolvers as part of the redshift estimator logic.

1.7 Scope of this Work

The "deep" sky surveys are ground-based surveys that takes images of the sky - image that include objects more deeper into the sky that possible in previous sky surveys. From the image, the relative positioning of the objects is not clearly obvious. As we go deeper into the sky, the number of objects being captured by the survey is increasing mani-fold. Spectroscopic measurement of redshift is time-consuming and thus, not feasible to used as a technique for all the objects. Additionally, spectroscopic measurements are not possible for faint objects from the deeper regimes of the sky. It should be used only when detailed investigation is needed for specific objects. The accuracy of the redshift estimates improves

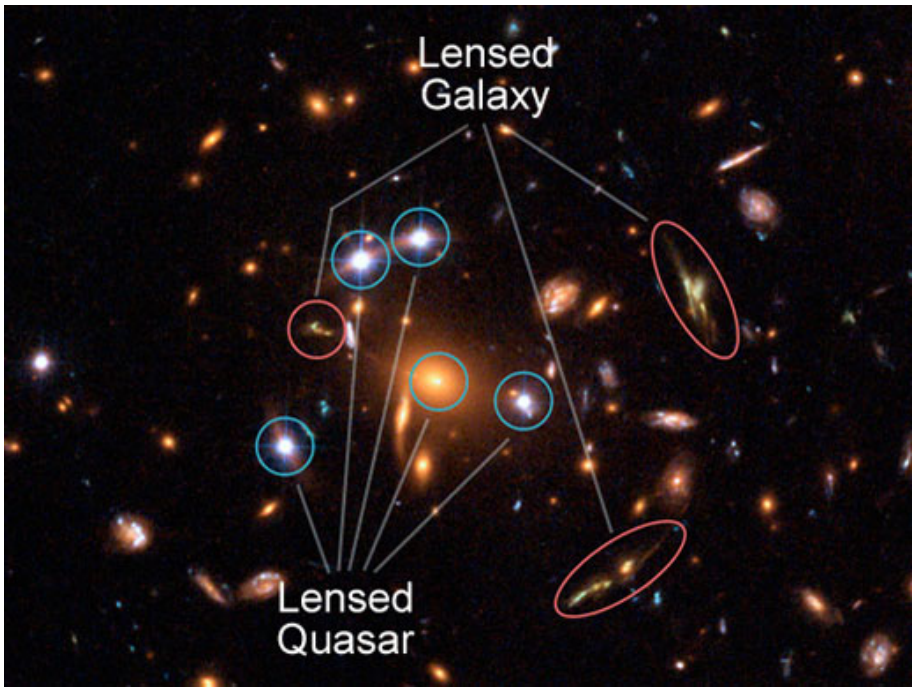


Figure 1.12: Gravitationally Lensed Galaxy

our confidence in the relative positioning of the multitudes of sky objects in the vast expanse of the universe. Alternative approaches are needed to determine redshift from the captured images.

The photometric redshift techniques use distance dependent attribute such as color to model their relationship and use them as estimators. We need to include multiple photometric (apart from color) and morphological attributes of the sky objects in the estimation process and make the estimation process robust and scalable to the degeneracy problem that plagues the existing photometric methods. Data-mining algorithms such as association rule learners, support vector machines, decision trees, neural networks are known to address this type of problem in other domains. This thesis evaluates machine learning and statistical techniques to address the current need. The results are compared and contrasted with the past related work. The past related work has largely been based on template fitting as well as machine learning techniques, primarily Artificial Neural Networks (ANN) and Random Forests (RF). Multiple learning algorithms are investigated with particular

emphasis on the required characteristics to return the robust measures of redshift. The data-set being used for the analysis will include data from multiple parts of the electromagnetic spectrum for the different sky objects. It will also include other distance-independent photometric attributes such as Point Spread Function (PSF) flux, Petrosian flux and Fiber flux. Additional distance-independent morphological attributes that were considered in the study include the Petrosian radius, the concentration index (ratio of the Petrosian radius containing 90 percent of the flux to the Petrosian radius containing 50 percent of the flux) and ratio of Fiber flux to Petrosian flux. This transitions the redshift estimation problem from a two-dimension color-color problem "photometric redshift" to a multi-dimensional problem "photomorphic redshift".

Problem Statement Mathematically speaking, we have n functions which are constraints on the value of z, say $f_1(z), f_2(z), \cdots, f_n(z)$. We have the data set that can be assumed to adhere to these constraints. The problem to be solved is finding the "most representative" model that best captures the constraints and avoids over-fitting so that the model is generalized enough for unseen data. In other words, independent estimators $f_1(z), f_2(z), \cdots, f_n(z)$ lead to the same answer (aka z) and thus, address the degeneracy concern that plague existing photometric estimation techniques. The ideal case will be as shown in Figure 1.13. Machine learning and statistical techniques are used to develop the possible models from the training dataset (a subset of the large dataset - for example, two-third of the entire sample) and study the predictive accuracy of the model against the test dataset (mutually exclusive subset different from the training set - for example, one-third of the entire sample when two-third of the entire sample is used as the training set). There has been extensive study in the field to use color and certain types of magnitude as redshift predictor. The present study adds other types of magnitude and morphology to the mix of attributes and studies the impact on accuracy using various predictive modeling methods. For the ideal solution, the predictive performance will be as shown in Figure 1.14b; these two types of plots are presented for all the investigated methods in the rest of the chapters.

$Color = f_1(z); Radius = f_2(z); Concentration = f_3(z)$

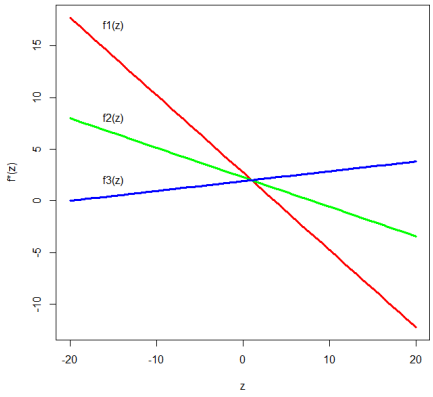
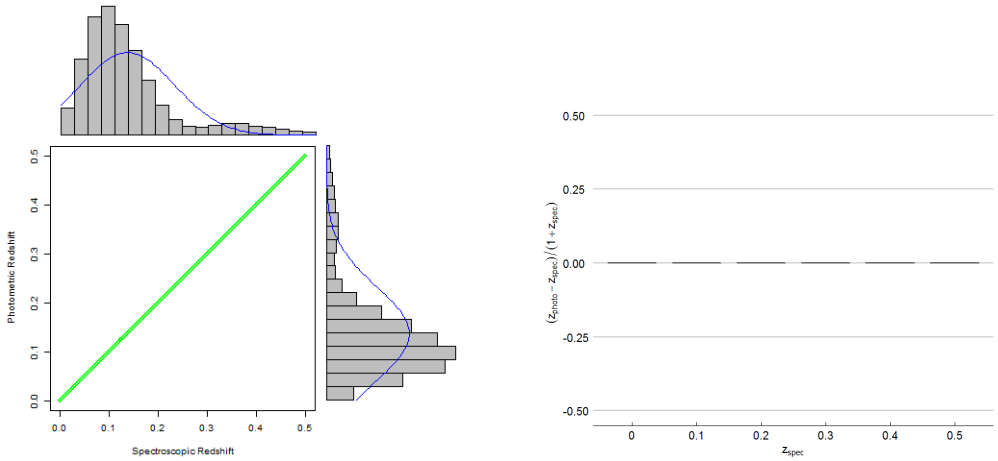


Figure 1.13: Ideal Solution for the Problem Statement

$$\Rightarrow \text{Find the z} = \begin{cases} = f_1^{-1}(Color) \\ = f_2^{-1}(Radius) \\ = f_3^{-1}(Concentration) \end{cases}$$



(a) Spec-z vs. Photo-z

(b) Spread of estimates

Figure 1.14: Predictive Performance of Ideal Photo-z Estimator

It is important to note that photometric redshifts typically cannot be used directly for

cosmological analysis, unless the estimators error distributions can be quantified precisely. The standard approach to quantify, or calibrate, the photometric error distributions is to use a small subsample of galaxies with known redshifts. Spectroscopic samples used to train photo-zs need to be locally (in the space of observables) representative subsamples of the photometric samples. For calibration of the photo-z error distributions, however, the spectroscopic sample must be globally representative. The ideal spectroscopic survey should satisfy the following properties - span a large area to beat down sample variance, and has to have tens of thousands of galaxies; span the same range of redshifts, galaxy types, and other observational selection parameters as the photometric survey; and extremely accurate redshifts. Existing and upcoming photometric surveys will have to learn to deal with very incomplete spectroscopic samples for photo-z calibration.

Chapter 2 provides the background of the precursors of the photomorphic redshift estimation technique. The precursor redshift estimate methods are the spectroscopic and the photometric redshift estimation techniques. The history of the astronomy domain in terms of the need and importance of redshift estimation is discussed. The versions of estimation techniques are categorized into color, template fitting and machine learning approaches. The study of galaxies and their redshift estimation is steadily becoming dependent on panchromatic studies that use multi-wavelength measurements from one or a combination of sky surveys. Chapter 3 elaborates the data collection process and the data sources from the SDSS CAS job server and other research studies provided online links such as the MegaZLRG DR6. The metrics that will be used in comparing the predictive accuracy of the experiments is discussed here as well. Chapter 4 presents the photomorphic redshift estimation technique using the Generalized Linear Model (GLM). Chapter 5 discusses the Bayesian network approach to photomorphic redshift estimation. Chapter ch:Future Work looks at the potential of the photomorphic redshift and the possible directions that the current work can take. Chapter 7 concludes with the main findings related to the relevance of the machine learning techniques in redshift estimation.

Chapter 2: Background

2.1 Spectroscopic Redshift

Almost every galaxy in the sky has a redshift in its spectrum. In general, a measurement of distance requires assumptions to be made about cosmological parameters while a redshift measurement does not. Spectroscopic redshifts (spec-z) can reach an accuracy of better than 10^{-3}, however this process is costly and time consuming. Each galaxy must be examined individually and observed for sufficient time that enough light is collected and a clear spectrum established. Modern multi-object spectrographs expedite this process by using multiple optical fibers to collect light for up to 4,000 galaxies simulataneously. However even these cutting edge, high-throughput machines are limited to observing $\sim 60,000$ galaxy spectra per observing night. [38]

Numerous redshift surveys have been done in the past as outlined in the survey paper [55]. The broad categories of redshift surveys include:

- Pencil-beam surveys (KOSS, Durham-AAT-SAAO, photometric and spectroscopic redshifts)
- 2-D Surveys by slices
- 3-D Surveys (CfA, Pisces-Perseus, SSRS, IRAS [1])
- Sparse Surveys with sparse sampling to survey huge volume with small number of observations

[1]The Infra-Red Astronomy Satellite (IRAS) was the first attempt to map the full sky at infra-red wavelengths. This could not be done from ground observatories because large portions of the infra-red spectrum is absorbed by the atmosphere. [26]

- Targeted Surveys (binaries and Groups, isolated galaxies, dwarfs, emission-line galaxies, zones of avoidance due to extragalactic extinction where radio and infrared techniques are useful, superclusters and voids).

Measuring a spectroscopic redshift involves the following four steps:

1. Obtain the spectrum of the object that shows spectral lines. From the pattern of lines, identify which line corresponds to which atom, ion, or molecule.

2. Measure the shift of any one of those lines with respect to its expected wavelength λ_{rest}, as measured in a laboratory on Earth. The measured redshift does not depend on which line you choose.

3. Apply a formula that relates the observed shift $\lambda_{observed}$ to velocity along the line-of-sight.

$$z = \frac{\lambda_{observed} - \lambda_{rest}}{\lambda_{rest}} \tag{2.1}$$

$$\Rightarrow 1 + z = \frac{\lambda_{observed}}{\lambda_{rest}} \tag{2.2}$$

Thus, if one can localize the spectral fingerprint of a common element such as hydrogen, then the redshift can be computed using the above simple arithmetic. Because of the spectrum shift, an identical source at different redshifts will have a different color through each pair of filters. At redshift $z = 0.0$, the spectrum is bright in the u and g filters, but dim in the i and z filters. At redshift $z = 0.8$, the opposite is the case. This suggests the possibility of determining redshift from photometry alone. Refer Figure 2.1. The scikit-learn DecisionTreeRegressor method models a 20-level decision tree and has a Root Mean Square (RMS) error of 0.22 when tested on 102,798 SDSS galaxies. About 1.5% of objects have redshift estimates which are off by greater than 1; termed as "catastrophic errors". Refer Figure 2.2. [67] [68]

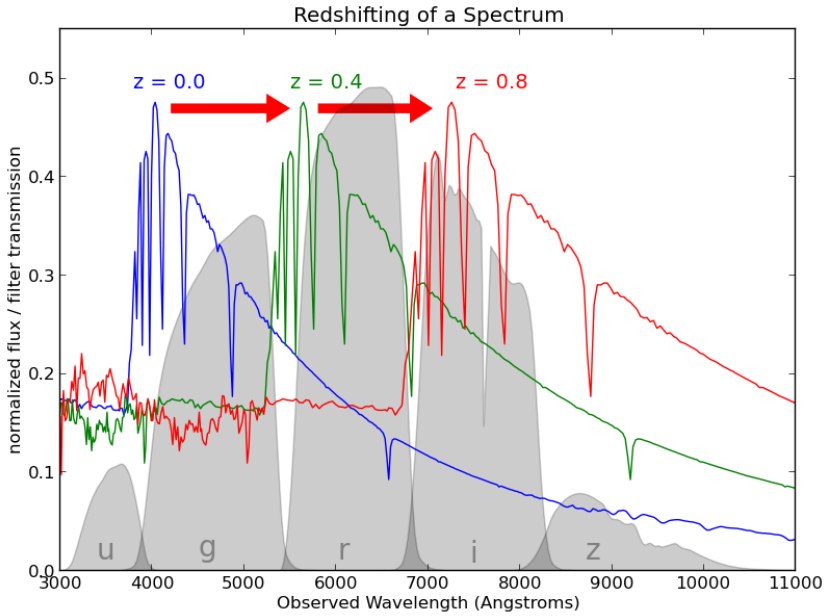


Figure 2.1: Spectrum of the star Vega (α-Lyr) at three different redshifts with SDSS ugriz filters as reference [67]

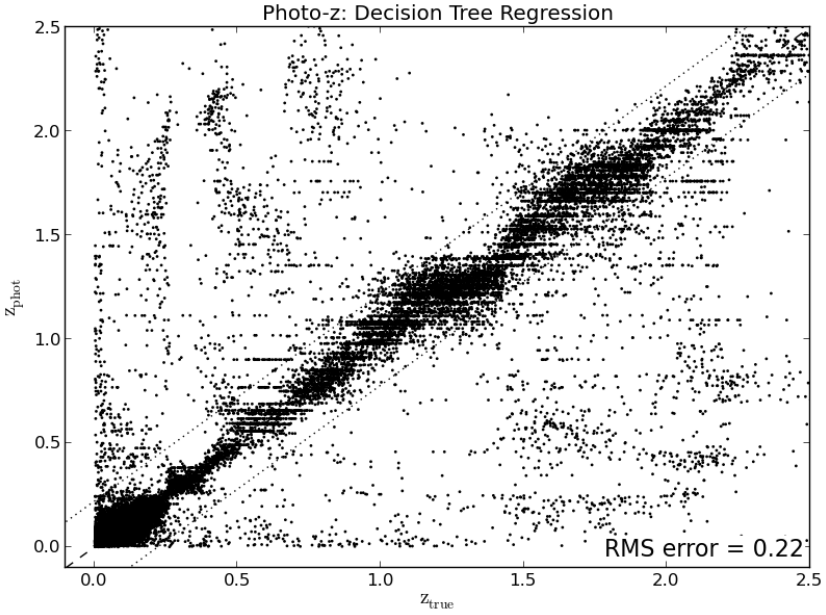


Figure 2.2: True and predicted redshifts of 102,798 SDSS galaxies using scikit-learn DecisionTreeRegressor [67]

The velocity of the galaxy with respect to us, in units of km/sec is equal to $c \times z$, where c is the speed of light, c = 3 x 105 km/sec. This definition of c only applies when z is small compared to 1.0. In the cosmological context, the redshift tells us the relative scale of the universe at the time the light left the galaxy. The redshift of any galaxy will have two components: a dynamic component and a cosmological component. However, from Earth we can measure only a single number, the redshift z. Without external arguments, we cannot distinguish the two types of redshift. The Cosmological Redshift is a redshift caused by the expansion of space. The wavelength of light increases as it traverses the expanding universe between its point of emission and its point of detection by the same amount that space has expanded during the crossing time. As a general rule, for nearby galaxies ($z < 0.001$), the cosmological component is small: the dynamic part prevails and we can think in terms of Doppler shifts (objects moving through space). For relatively distant galaxies ($z > 0.01$), the dynamic part is smaller than the cosmological part, and thinking in terms of Doppler shift velocities could be misleading. At intermediate redshift, $z \approx 0.003$, the two contributions to the measured redshift can be comparable in size. In this case, sorting out what is what is a challenge even to experts [69]. The third type - Gravitational Redshift - is a shift in the frequency of a photon to lower energy as it climbs out of a gravitational field. The uncertainties in studies of the evolution of galaxies are dominated by shot noise (statistically small samples of galaxies) rather than errors in redshift. If one could derive an estimate of the redshift of a galaxy from its photometric magnitudes then large, complete surveys could be realized. [70]

2.2 Photometric Redshift via Color

Photometric redshift estimation techniques use the difference of the magnitudes that represents color to interpolate and extrapolate redshift of unknown objects. It was first applied by Baum using nine passbands on elliptical galaxies in cluster 3C395 from three clusters out to maximum redshift of $z = 0.46$ and estimated it at $z = 0.44$. The possibility of measuring

very faint objects photoelectrically in the infrared, as well as shorter wavelengths, provided the opportunity of undertaking a stronger attack on the redshift-magnitude problem. [71] Color-color diagram involves plotting lines of constant redshift and varying spectral type, known as iso-z lines. Since most normal color-color diagrams are degenerate in a range of redshifts, color-shape diagrams are instead used. The shape measured whether the Spectral Energy Distributions (SEDs) turned up or down at both ends, that is, whether the spectrum was bowl shaped or humped.

Both the redshift and the bolometric magnitude of a galaxy are measured photoelectrically. The light from each galaxy is measured in a number of colors. When the results are plotted on a true energy scale, they yield a SED as shown in Figure 2.3. Figure 2.3 represents the mean of six bright elliptical galaxies in the Virgo cluster. A curve for similar ellipticals in another cluster at a greater distance will have about the same shape, but will be displaced towards the right and will fall at fainter magnitudes as shown in Figure 2.4 [2] for four elliptical galaxies at $z = 0.29$. The horizontal displacement yields the difference in redshift while the vertical displacement, after a $log(1+z)$ correction due to the logarithmic abscissa, yields the difference in bolometric magnitude. The study yielded an approximate relation $m(bol.) = constant + 5logz$ as shown in Figure 2.5. [71] One interpretation is that the color measured the first derivative with respect to wavelength of the spectrum and the shape measured the second derivative. The redshift of the galaxy is then found by finding the iso-z line closest to the point representing the galaxy. This method has been applied on a sample of 100 galaxies with known spectroscopic redshifts ranging from $z = 0.025$ to $z = 0.700$.

[72] reviews the early history of photometric redshifts. Though Baum (1962) [71] is noted for the first use of multicolor photometry for redshift estimation, nearly two decades later, Puschell et. al. (1982) [73] was the first to use the term "photometric redshift". This study estimated redshifts of faint radio galaxies via broadband photometry by using combination of near-infrared bands (JHK) along with optical bands (RI).

[2] Each point is the mean for four galaxies. The vertical bars represent the individual probable error.

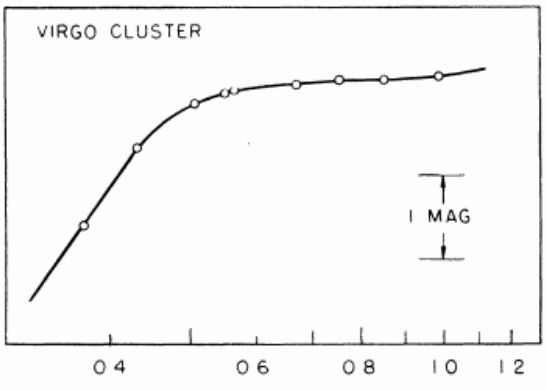


Figure 2. Mean spectral energy distribution curve for six bright elliptical galaxies in the Virgo Cluster. The ordinate represents relative energy per unit wave length expressed on a magnitude scale.

Figure 2.3: Mean SED for six elliptical galaxies in the Virgo Cluster [71]

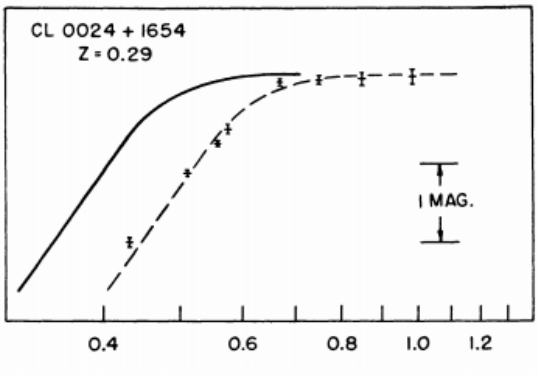


Figure 4. Mean spectral energy distribution for four elliptical galaxies in a cluster at $0^h\ 24^m$, $+16^\circ 54'$. These results show the red-shift to be 0.29. The vertical bars through the observed points represent the individual probable errors.

Figure 2.4: Mean SED of four elliptical galaxies with z=0.29 vs. Virgo Cluster [71]

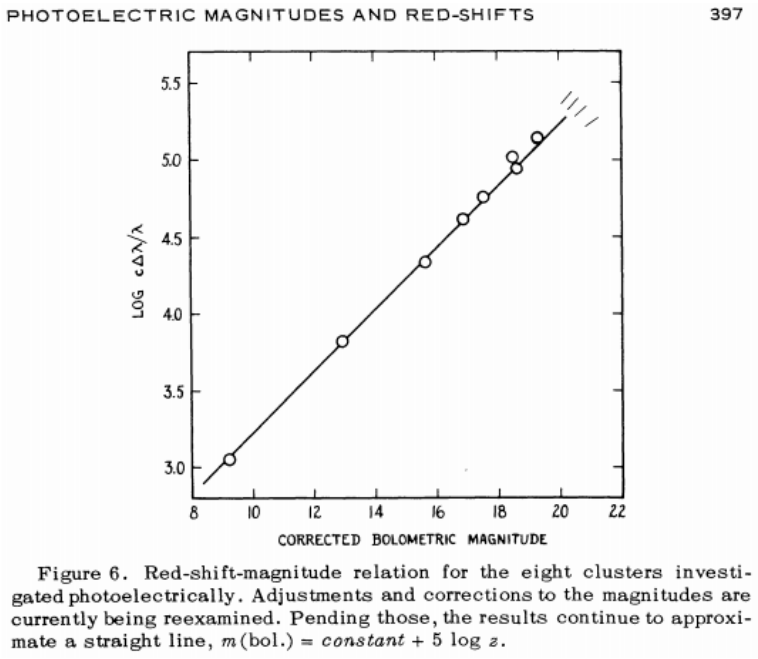


Figure 6. Red-shift-magnitude relation for the eight clusters investigated photoelectrically. Adjustments and corrections to the magnitudes are currently being reexamined. Pending those, the results continue to approximate a straight line, m(bol.) = *constant* + 5 log z.

Figure 2.5: Redshift-Magnitude relation for the eight clusters [71]

Butchins (1981 [74], 1983 [75]) used UK Schmidt plate BV R photometry of 3664 galaxies that reached $B \sim 22$. Because of the overlap in BV R colors of low-redshift ($z \leq 0.1$) early-type galaxies with higher redshift ($z \sim 0.4$ [3]) later-type galaxies, Butchins applied probablility constraints on luminosities and thus, it is one of the first use of Bayesian techniques in photometric redshift estimation.

Koo (1981 [76], 1985 [77][4], 1986 [78]) reached superior redshift accuracy $\delta z \leq 0.05$ in fainter limits ($B \sim 24$) and far less degeneracy by exploiting a set of filters (four broadband photographic filters UBRI) with much longer wavelength coverage. Loh and Spillar (1986 [79] [72]) was another similar effort where six medium bandpass filters were used. It is important to note here that these studies had been limited by the lack of redshift surveys to the limit of the photometric data. Thus, it required redshifts to be derived by matching the observed galaxy colors with those predicted from SEDs and assumed galaxy evolution

[3]The range of redshift $z = 0.0$ to $z = 0.4$ is chosen because above $z = 0.4$ the uncertainties of the ultra-violet part of the galaxy spectrum become significant in the blue. [74]

[4]Poor Person's Redshift Machine for Galaxies

models.

As part of their study, (Koo [77] [78] and Kron [80]) demonstrated that the distribution of galaxies in the multidimensional flux space U, B_J, R_F, and I_N are nearly planar. The finite thickness of this plane carries significant physical information, not just random scatter. The I_N magnitude is most strongly correlated with the R_F magnitude. This is expected as the spectral energy distributions of normal (i.e. non starburst) galaxies are a continuous monotonic function. [70] found that the position of a galaxy within this plane (out to $B_J \leq 22.5$ and $z < 0.3$, 370 galaxies in the photometric sample of SA57 and SA68 galaxies) is determined by its redshift, luminosity, and spectral type. Linear, quadratic, third and fourth-order polynomials were fitted. Redshifts for galaxies were estimated to an accuracy better than $\delta z = 0.05$. The dispersion was attributed to the photometric uncertainties within the photographic data. It was expected at this time that using high signal-to-noise photometric data, one can achieve an intrinsic dispersion of less than $\delta z = 0.02$. Using deep CCD photometry, [81] quantified the photometric-redshift relation within the standard AB magnitude system reaching the asymptotic intrinsic dispersions ($\sigma_z \approx 0.016$ for $z < 0.4$). This empirical relation had a measured dispersion of $\sigma_z \approx 0.02$ for $z < 0.4$. Redshifts can be reliably estimated for objects from broadband photometry out to $z \sim 0.8$.

Another color-based method is the "permitted redshifts" - colors of galaxies are plotted as a function of redshift from the Bruzual and Charlot models (a library of stellar flux spectra calculated by Kurucz and Buser from theoretical model atmospheres. The catalog consists of 1434 files, each representing a metal-line blanketed flux spectrum for a theoretical stellar model atmosphere.). Each available color (with its associated uncertainty) of a galaxy defines a "permitted redshift" range on the corresponding color-redshift diagram. The intersection of the permitted redshift ranges for all the colors determines the redshift. This method was used to discover a cluster of galaxies at $z > 0.75$ by looking for an excess in the redshift distribution in the field of a gravitationally-lensed quasars. The above method was also used to determine the redshift distribution of the Hubble Deep Field.

All galaxy spectra have a large Lyman break; short-ward of 912Å, the continuum drops

dramatically. When this break is redshifted into and past the U filter, the U flux is greatly reduced or non-existent, resulting in very red ultra-violet colors. In the ultra-violet dropout techniques, an exact redshift of a galaxy is not determined. Rather, the redshift is determined to be in the redshift range where the Lyman break is in or just past the U filter. Since U filters typically have a central wavelength of 3000Å, this works out to a redshift of $z > 2.25$. In practical terms, redshifted template galaxy spectra are used to determine a locus on a color-color plot where most galaxies lie in a particular redshift range. Those galaxies whose measured colors lie within the locus are deemed to be in that redshift range. Clearly, this method is a lot simpler than that the "permitted redshifts" as only two colors are considered. It is also a lot less precise as the redshift is not very constrained. This "Lyman-break ultra-violet dropout" technique is ideally suited for pre-selecting galaxies at high redshift for spectroscopic confirmation.

2.3 Photometric Redshift via Template-fitting

A different class of photometric redshift methods rely on a χ^2 fitting of a library of template SEDs to the observed data points (refer Equation 2.3), and differ mainly in how the SEDs are derived and on how they are fitted to the data. This is termed as "Template-fitting methods". This is employed by HYPERZ, Le PHARE and CanadaFrance Deep Fields-Photometric Redshift Survey (CFDF-PRS). [82] [83] [84] KCorrect is one of the implementation of template-fitting method. Chi-square is a statistical test commonly used to compare observed data with data we would expect to obtain according to a specific hypothesis. It is used to determine if the deviations (differences between observed and expected) were a result of chance, or were they due to other factors at work. The chi-square test tests the null hypothesis that there is no significant difference between the expected and observed result. The observed SED of a given galaxy is compared to a set of template spectra where $F_{obs,i}, F_{temp,i}$ and σ_i are the observed and template fluxes and their uncertainty in filter i, respectively, and b is a normalization constant.

$$Minimize \quad \chi^2(z) = \sum_{i=1}^{N_{filters}} [\frac{F_{obs,i} - b * F_{temp,i}(z)}{\sigma_i}]^2 \tag{2.3}$$

A method very similar to this is the linear regression method. Linear regression requires a training set of a large number of galaxies with multi-color photometry and spectroscopic redshifts. Redshift is assumed to be a linear or quadratic function of the magnitudes M_i of the galaxies as shown by Equation (2.4) where N is the number of filters. For improved accuracy, a large number of spectroscopic redshifts must have been measured before the technique can used. This method produces small dispersions, even when the number of filters available is small, and it has the advantage that it does not make any assumption concerning the galaxy spectra or evolution, thus bypassing the problem of our poor knowledge of high redshift spectra. However, this approach is not flexible: when different filter sets are considered, the empirical relation between magnitudes and redshifts must be recomputed for each survey on a suitable spectroscopic subsample. Moreover, the training set is constituted by the brightest objects, for which it is possible to measure the redshift. Thus, this kind of procedure could in principle introduce some bias when computing the redshifts for the faintest sources, thus error-prone, because there is no guarantee that we are dealing with the same type of objects from the spectrophotometrical point of view. Also, the redshift range between 1.4 and 2.2 had been hardly reached by spectroscopy up to now, because of the lack of strong spectral features accessible to optical spectrographs. Thus, no reliable empirical relation can be found in this interval.

$$z = \alpha_0 + \sum_{i=1,...,N} \alpha_i M_i + \sum_{i=1,...,N, j=1,...,N} \alpha_i M_i M_j \tag{2.4}$$

Including near-IR JHK photometry strongly reduces the error bars within the $1.2 \leq z \leq 2.2$ range, without significantly improving the uncertainties in z_{phot} outside this interval. If the filter Z is considered in addition to the five optical filters, the resulting dispersion

at low redshift become smaller up to $z_{model} \simeq 1.5$, but the degeneracy at $z_{model} = 1.5 \cdots 3$ still remains, even if less dramatic. In general, the reasons of failures can be ascribed to many effects, such as a wrong photometry (systematic errors when measuring magnitudes or underestimated photometric errors) leading to a highly unlikely fit, or a probability function with significant secondary peaks, because of degeneracy among the fit parameters, or a relatively flat probability function due to a lack of sufficient photometric information. [82]

Astronomers spread the light of the galaxy and draw the intensity as a function of the light's wavelength (usually binned into few thousand wavelength bins) called the spectral energy distribution (SED). Comparing the observed wavelengths of well known spectral lines of various elements (H, O, Na, Ca, etc.) to the theoretical rest frame values one can get the redshift. Since galaxies can be very faint, lot of time needed to get a reasonably good signal-to-noise spectrum even with the largest telescopes. In comparison, observing the galaxies in just a few bands (broad-band filters e.g., five *ugriz* of SDSS or six *ugrizy* of LSST) is much faster. This very low resolution spectrum (also called photometry) carry some information on the redshift of the galaxy, but the fine (few Angstrom wide) spectral line is smeared out several hundred times.

Refer Figure 2.6 for an analogy of the less detailed (thus, incorrect visual representation of the distribution) histogram of 100,000 data points when the bin sizes are larger in width; similar to the usage of broad filters for photometric surveys. This is in contrast to real and correct distribution representation when bin size is small; similar to the detailed measurements of spectroscopic measurement. Thus, narrower filters improve the spectral resolution, but strongly reduce the total system throughput.

[64] explores how photometric redshift performance depends on the number of filters n_f. The fewer the filters, the more prone the system is to color-redshift degeneracies; these make it impossible to unambiguously determine the redshift for a galaxy, even if observed at relatively high S/N. Adding near-IR observations improves the performance of low n_f systems, but the system which maximizes the photometric redshift completeness is formed

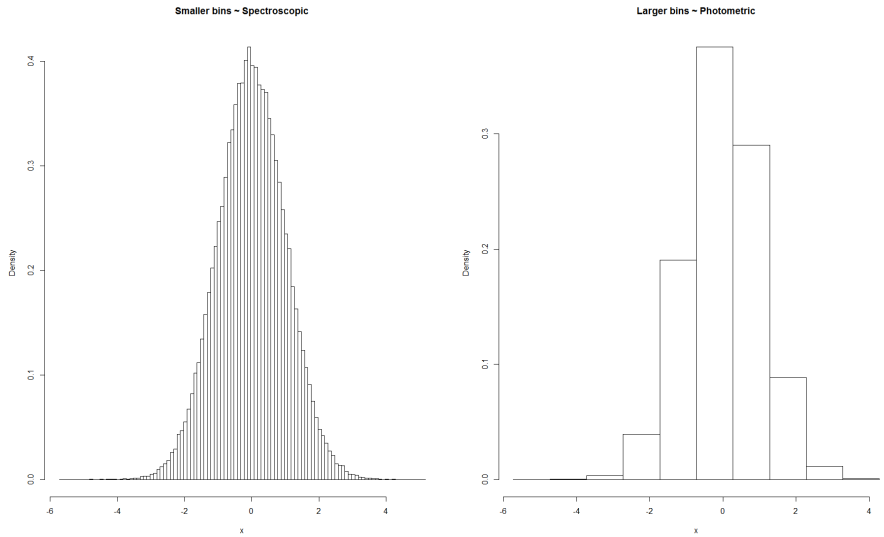


Figure 2.6: Spectroscopic vs. Photometric Analogy Example

by 9 filters with logarithmically increasing bandwidth (constant resolution) and half-band overlap, reaching $\sim$ 0.7 mag deeper, with 10% better redshift precision, than 45 filter systems. A system with 20 constant-width, non-overlapping filters reaches only ~ 0.1 mag shallower than 45 filter systems, but has a precision almost 3 times better, $\delta z = 0.014(1+z)$ vs. $\delta z = 0.042(1+z)$. BPZ photo-z estimation software is used in this study.

The accuracy with which redshifts can be determined is sensitive to the star formation history of the galaxy, for example the effects of age, metallicity and ongoing star formation. Whereas spectroscopic redshifts use sharp absorption and/or emission lines to determine the rest wavelength of the spectrum accurately, it is also possible to exploit the overall characteristic shape of the SED to estimate the redshift of a galaxy. This photometric redshift approach can be applied to broad-band images provided they have sufficiently high signal-to-noise ratio and adequately sample the important features of the SED. In particular, the $4000\mathring{A}$ spectral break and the Balmer and Lyman series limits are important features that arise in almost all galaxy spectra. Although precise redshifts cannot be determined by this method, estimates of (or limits on) z are obtained.[85] focus more closely on the interrelation

between star formation history and redshift estimation and photometric redshift determination explicitly takes into account the degeneracies implied by these variations. It addresses possible degeneracies in plausible values of galaxy type and redshift by storing a probability map for each galaxy, which can be used to estimate a range of acceptable redshifts rather than reducing the observed data to a single best bet estimate of galaxy type and redshift.

Thus, given a set of spectroscopic redshifts z_{spec} and colors C, the training-set method will try to fit a surface $\hat{z} = z(C)$ to the data. This is based on a very strong assumption: that the surface $\hat{z} = z(C)$ is a function defined on the color space, where each value of C corresponds to one and only one redshift. Although this functionality of the redshift/-color relationship cannot be taken for granted in the general case, it seems to be a good approximation to the real picture at $z < 1$ redshifts and bright magnitudes. [81]

Galaxies with the same value of (C) may have slightly different redshifts, and it seems to be assumed implicitly that this scatter is the factor limiting the accuracy of the method. Refer Table 2.1 for an example of how X3 is the suitable degeneracy resolver for Y-prediction using X1 and X2. When the colors of a galaxy do not exactly coincide with one of the spectra, χ^2 or the maximum-likelihood method will assign the redshift corresponding to the nearest template in the color space. The color/redshift degeneracies happen when the line corresponding to a single template self-intersects or when two lines cross each other at a point corresponding to different redshifts [these cases correspond to "bendings" in the redshift/color relationship z = z(C)] as shown in Figure 2.7. It is obvious that the frequency of such crossings will rise with the extension of the considered redshift range and with the number of templates included. Moreover, the presence of color/redshift degeneracies is also increased by random photometric errors. By applying a simple PCA analysis to the HDF-N photometric sample, it can be shown that the information contained in the seven UBVIJHK filters for the HDF galaxies can be condensed using only three parameters, the coefficients of the principal components of the flux vectors. [86]

The training-set method somewhat alleviates the degeneracy problem by introducing an additional parameter in the estimation: the magnitude, which in some cases breaks the

degeneracy. However, color/redshift degeneracies may also affect galaxies with the same magnitude, and the training-set method does not even contemplate their possibility. It may be known from previous experience that one of the possible redshift/type combinations is much more likely than any other, given the galaxy magnitude, angular size, shape, etc. In that case, and since the likelihoods are not informative enough, Bayesian probability states that the best option would be the one more likely a priori. [86]

Table 2.1: Degeneracy Resolution of $X1, X2 \Rightarrow Y$ using $X3$

X1	**X2**	**X3**	**Y**
x1	x21	-	y1
x1	x22	-	y1
x1	x21	x31	y1
x1	x22	x32	y1

[87] derives empirical color-redshift relations for $z \leq 4$ galaxies in the Hubble Deep Field (HDF) using a linear function of three photometric colors (U-B, B-V, V-I). The dispersion between the estimated redshifts and the spectroscopically observed ones range from $\sigma_z = 0.03 \cdots 0.1$ for $z \leq 2$ galaxies, and from $\sigma_z = 0.14 \cdots 0.25$ for $z \geq 2$ galaxies. The estimated redshifts are consistent with those derived from spectral template fitting methods. The advantage of these color-redshift relations is that they are simple and easy to use and do not depend on the assumption of any particular spectral templates; they provide model independent redshift estimates for $z \leq 4$ galaxies.

[88] is an analysis of photometric redshifts with Bayesian priors on physical properties of galaxies. They construct model templates of galaxies using a stellar population synthesis code and apply Bayesian priors on physical properties such as stellar mass and star formation rate. These physical priors are a function of redshift. and they help reduce the degeneracy and deliver significantly improved photometric redshifts. *This is an important finding, particularly for our present study. We are performing study of similar concept, but in machine learning context.* They simultaneously measure redshifts and physical properties

Left: *V-I* vs. *I-K* for the templates used in § 4 in the interval $1 < z < 5$. The size of the filled squares grows with redshift, from $z = 1$ to $z = 5$. If these were the only colors used for the redshift estimation, every crossing of the lines would correspond to a color/redshift degeneracy. *Right*: The same color-color relationships, "thickened" by a 0.2 photometric error. The probability of color/redshift degeneracies increases greatly.

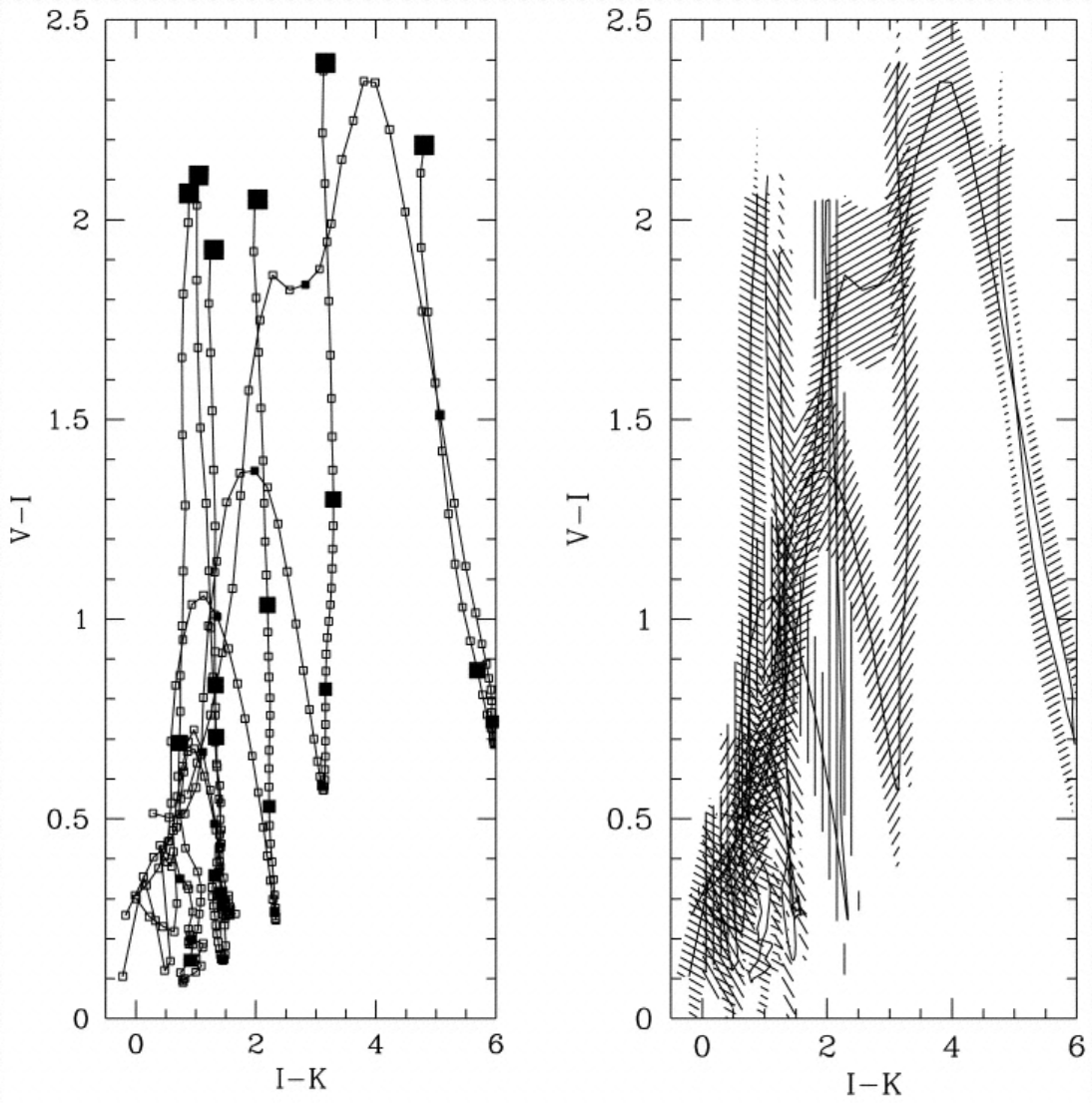


Figure 2.7: Color/Redshift Degeneracies [86]

of galaxies in a fully self-consistent manner, unlike the two-step measurements with different templates often performed in the literature. One may rightly worry that the physical priors bias the inferred galaxy properties, but we show that the bias is smaller than systematic uncertainties inherent in physical properties inferred from the SED fitting and hence is not a

major issue. This will be used in the on-going Hyper Suprime-Cam survey. This algorithm is in the SED-fitting category.

2.4 Photometric Redshift via Machine Learning

When we look at an object in the picture, the shape of the object alone does not tell us the distance of the object from the camera. Is it a small red ball close to the camera or a basketball at a larger distance? To help us determine the separation distance, we need more information. Color is distance dependent and yet it exhibits a degeneracy behavior when used as the only estimator of redshift. There have been past studies similar to this work to investigate the use of other measured attributes of the objects. The past studies have largely focused on attributes that are dependent on distance to determine, if aided with the large volumes of data, they can help us in developing more robust estimators.

Approximate computing appears to be our best answer by applying intelligent computation on the massive amounts of data that are being and will be produced in the world. It may be argued that approximate computing makes it difficult to reason logically about the results produced by a program. However, such reasoning is often difficult even in traditional computingreal numbers cannot always be represented precisely, the order of access to shared variables by multiple processors is often unpredictable and may lead to non-deterministic results, and it is virtually impossible to eliminate all potential sources of errors, hardware and software, from any system. Unlike traditional computing, approximate computing grants that errors will occur and transfers the responsibility for tolerating errors to the runtime, compiler, or even the application itself. This approach is applicable where the solution space is such that knowledge of the past behavior for a range of inputs is a good predictor of the behavior of the program on some new input. As an accelerator for the approximable region, a learning engine (such as digital neural network) has been proposed to be used in order to deliver an approximately correct result. For problems that can tolerate approximate answers, the expected benefit should be impressive performance gains for a tolerable loss of quality. [89] [90] The Approximation Criteria is that we cannot compute z exactly,

but rather estimate such that $|\widetilde{z} - z|$ is as close to zero as possible.

One of the most recent study (December 2014, [91]) provides the PhotoWeb that estimates the photometric redshifts of individual galaxies and their equivalent distance with sub-megaparsec accuracy. It uses the cosmic web as a constraint over the photo-z estimates. The redshift errors for individual galaxies of the PhotoWeb are of the order of $\Delta z \simeq 0.0007$, compared to errors of $\Delta z \simeq 0.02$ for current photo-z techniques. The mean redshift error is of the order of $5 \times 10^{-5} - 5 \times 10^{-4}$ compared to mean errors in the range $\Delta z \simeq 0.001 - 0.01$ for the best available photo-z estimates in the literature. The current photo-z techniques based on the spectral energy distribution of galaxies and their projected clustering produce redshift estimates with large errors. The large error is due to the poor constraining power of the attributes under consideration. The cosmic web, on the other hand, provides the strongest constraints on the position of galaxies. The network of walls, filaments and voids occupy ∼10% of the volume of the Universe, yet they contain ∼95% of galaxies. The cosmic web is a cellular system with well-defined boundaries. It sets a restricted set of intermittent positions that a galaxy can occupy along a given line-of-sight. Using the information in the density field computed from spectroscopic redshifts, the possible locations of a given galaxy can be narrowed down along the line of sight from a single broad probability distribution to one or a few narrow peaks.

[70] [86] [92] provide an extensive review and comparison between different photo-z methods. Standard photo-z techniques compute redshift estimates for each galaxy independently taking into account only their SED. They typically provide redshift estimates with errors of $\triangle z \sim 0.01 - 0.02$, equivalent to $\sim 40 - 80 Mpc$. [91]

[84] derives photometric redshifts using the neural network method ANNz as well as five other publicly available photometric redshift codes (HyperZ, SDSS, Le PHARE, BPZ [86] and ZEBRA) for $\sim$ 1.5 million Luminous Red Galaxies (LRGs) from SDSS DR6. Refer Figure 2.2 for a comprehensive list of photo-z methods till date. The codes such as Le PHARE which use new observed templates perform best in the lower redshift bins. All codes produce reasonable photometric redshifts, the 1σ scatters ranging from 0.057 to 0.097,

if averaged over the entire redshift range as shown in Figure 2.9. The redshift ranges used in this study were defined with the following as boundaries [0.2, 0.4, 0.5, 0.6, 0.7, 0.9]. The density plot of spectroscopic versus photometric redshift for ANNz is shown in Figure 2.8. The comparison of the photo-z estimates using ANNz versus other algorithms is shown in Figure 2.10. The comparision metrics used in this study were:

- 1σ scatter between spec-z and photo-z defined as $\sigma_z = \sqrt{(z_{phot} - z_{spec})^2}$
- Bias defined as $b_z = z_{phot} - z_{spec}$
- 1σ scatter around the mean photometric redshift in each bin defined as $\sigma_{z2} = \sqrt{(z_{phot} - \bar{z}_{phot})^2}$
- 1σ scatter around the mean spectroscopic redshift in each bin is defined as $\sigma_{z3} = \sqrt{(z_{spec} - \bar{z}_{spec})^2}$

Note: Code and config files can be found at [93].

Table 2.2: Software Packages for Photo-z estimation

Code	**Method**
HyperZ [84]	Template
BPZ [86]	Template and Bayesian Priors
ANNz	Neural Networks
ZEBRA	Template, Bayesian and Hybrid
KCorrect	Model Templates
LePHARE	Template
EAZY	Template

The importance of the photo-z technique is growing not only with the desire to gain a greater understanding of galaxy evolution but also in weak gravitational lensing, where redshift estimates can reduce contamination from intrinsic alignments, and allow the possibility of 3D lensing studies. In the case of weak-lensing studies in particular, there is a strong motivation to measure the lensing signal at the faintest possible magnitudes. But, of course,

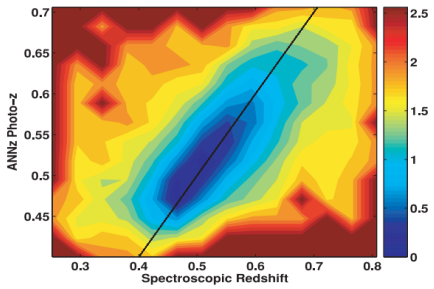


Figure 2. Density plots of spectroscopic versus photometric redshift for each of the public photo-z codes described in Section 3.1. The plots are colour-coded and the scale is exponential. A colour difference of 1 is equivalent to the density being decreased by a factor of e. The solid black lines show where the spectroscopic redshift is equal to the photometric redshift.

© 2011 The Authors, MNRAS **417,** 1891–1903
Monthly Notices of the Royal Astronomical Society © 2011 RAS

Figure 2.8: Density plots of spectroscopic versus photometric redshift for ANNz [84]

Figure 3. The 1σ scatter on the photometric redshift around the true spectroscopic redshift, defined as per equation (6) for each of the public photo-z codes described in Section 3.1, is in the left-hand panel and σ_{68} as a function of the spectroscopic redshift for each of the public photo-z codes described in Section 3.1 is in the right-hand panel.

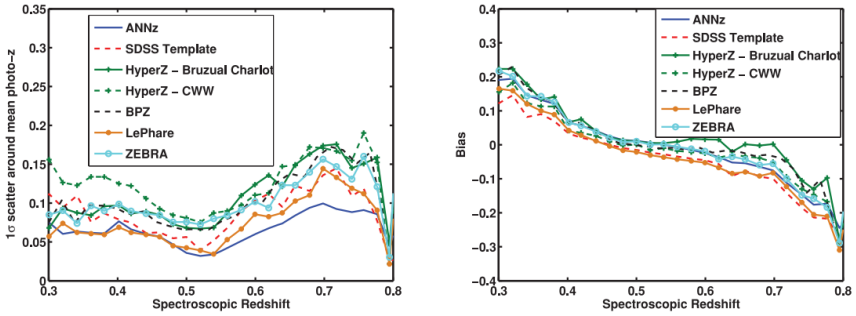


Figure 2.9: 1σ scatters for Photo-z estimation [84]

at faint magnitudes, the photometric measurement errors become significant and cause increased redshift errors. In recognition of the difficulty of obtaining reliable photometric redshifts at faint magnitudes, it is common, instead, to assume a statistical distribution for redshifts which may be calculated, given knowledge of the evolving galaxy luminosity function (GLF). [83]

Models typically assume that all data fits in memory, and that running time is accurately modeled as the number of basic instructions the algorithm performs. However in large-scale modern scientific experiments-related applications such as our study, data too large to fit

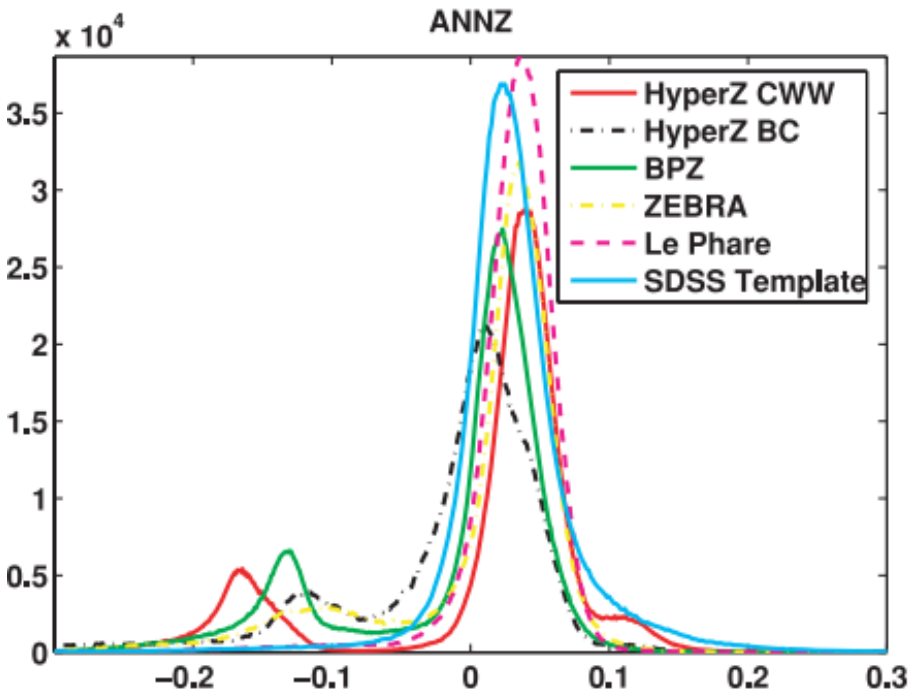


Figure 2.10: Histogram of the difference between photo-z estimate for all pairs of code vs. ANNz [84]

in memory must be analyzed. This consideration has led to the development of several models for processing such large amounts of data: the external memory model and cache-obliviousness where one aims to minimize the number of blocks fetched from disk; property testing where it is assumed the data is so massive that we do not wish to even look at it all and thus aim to minimize the number of probes made into the data; and massively parallel algorithms operating in such systems as MapReduce and Hadoop. [94]

[95] assessed the performance of photometric redshift estimator using approximately 15000 galaxies against the spectroscopic redshifts available from other surveys. The photometric data was collected via the sky survey DES during late 2012 and early 2013. Empirical photo-z methods using Artificial Neural Networks and Random Forests yielded the best performance in the performed tests. Additionally, neural networks have been found to be better estimators than template fitting methods based on the analysis on SDSS data (ANNz [84]).

Cool objects, such as small dwarf stars, give off most of their thermal radiation as infrared light. Therefore, the sky survey 2MASS could see these cool objects, which are otherwise invisible in visible-light surveys such as SDSS. In contrast, a hot, bright astronomical object that gives off a great deal of energy will emit x-rays. Due to these varying characteristics of the sky objects and the availability of large catalogs from sky surveys covering the different parts of the electromagnetic spectrum, analysis of data involving multi-wavelength measurements looks promising in providing a robust method for redshift estimation.

[96] estimates distance for $\sim$ 30 million galaxies from the SDSS DR4/DR5 data using two different ML-based approaches. MLP-based Neural network is used to categorize nearby ($z < 0.25$) vs. distant ($z > 0.25$) objects. Once categorized, two separate MLPs are used to work in the two distinct redshift regions. This improves the generalization capability compared to only one neural net since it is based on the different galaxy distribution in the two redshift intervals (the Main Galaxy (MG) sample in the nearby region, and the Luminous Red Galaxies (LRG) in the distant one). A hierarchical approach is proposed to partition the photometric parameter space using only the statistical properties of the data themselves. It starts from a preliminary clustering performed using an unsupervised clustering algorithm Probabilistic Principal Surfaces (PPS) [5]. It then makes use of the Negative Entropy concept and of a dendrogram structure to agglomerate the clusters found in the first phase.

[92] studies the size distribution of galaxies and its dependence on their luminosity, stellar mass, and morphological type using a sample of about 140,000 galaxies from the Sloan Digital Sky Survey (SDSS). Luminosity, size, circular velocity (or velocity dispersion), and morphological type are the most basic properties of a galaxy. Observed galaxies cover large ranges in these properties. Faint red galaxies have sizes quite independent of their luminosities. Clearly, the study of the distribution of galaxies with respect to these properties and the correlation among them is crucial to our understanding of the formation and evolution

[5]Nonlinear extension of principal components, in that each node on the PPS is the average of all data points that projects near/onto it.

of the galaxy population.

[97] covers the earliest detection ever of Type Ia supernova (2011fe [98] [98]) that has led to unparalleled observations of the initial stages of the stellar explosion and characterization of the nature of the stars that formed it. The early detection was possible due to real-time classification of astronomical time-series using machine learning-based computational framework that raised the supernova candidate event to the top of list of possible new transients. Type Ia supernovae have similarities that allow astronomers to use them as standards when comparing the distances of objects in the sky, however little is known about the stars that produce them or how they behave when they explode. The Palomar Transient Factory (PTF) is the wide-field survey that scans the skies for these transients. This is considered as an early example of peta-scale astronomical surveys with automated processing of massive data streams. This result is an excellent illustration of the power of machine learning to assist not only with data collection and reduction, but with the tasks of discovery and inference as well.

[99] explores the use of Generalized Linear Models (GLMs) for exploratory data analysis and robust regression. Logit and probit regression techniques are studied for handling binary/binomial data. It is used to explore the conditions of star formation activity and metal enrichment in primordial minihaloes from cosmological hydro-simulations including detailed chemistry, gas physics, and stellar feedback. They identify vast potential of GLMs and extended GLMs for the astronomical community in their possible application to a plethora of astronomical problems, such as: photometric redshift estimation (gamma distributed data), globular cluster counts (Poisson distributed data), or galaxy morphological classification (multinomial distributed data). The flow-chart for this approach is as shown in Appendix C.

[65] continues the study of [99] to explore the use of GLM based on principal components in estimating the photometric redshifts of galaxies from their multi-wavelength photometry. Using the gamma family with a log link function redshifts are predicted for the PHoto-z Accuracy Testing (PHAT) simulated catalogue and a subset of the galaxy data from SDSS

DR10. The overall performance of all existing photo-z codes were, to first order, consistent and displayed catastrophic errors ranging from 5 - 9%. This is considered good in terms of photo-z estimates [84]. The generated fits resulted in the catastrophic outlier rates to be as low as $\sim 1\%$ for simulated and $\sim 2\%$ for real data. The most dominant parameters affecting the predictive accuracy were the size of the training set and the number of principal components used. Adoption of the gamma family was based on the two important characteristics of the data: (i) a non-negative and continuous measurement, and (ii) heteroscedasticity, i.e., the variance of the photo-z measurements changes according to the redshift. The R-package CosmoPhotoz developed as part of the mentioned study includes two dataframes, PHAT0train and PHAT0test, containing 161042 and 8478 objects respectively. This dataset consists of 12 variables (11 bands and the redshift).

PhotoRApToR (Photometric Research Application To Redshift) is a Java/C ++ based desktop application with capabilities to solve non-linear regression and multi-variate classification problems. It is specialized for photo-z estimation. It embeds a machine learning algorithm, namely a multi-layer neural network trained by the Quasi Newton learning rule, and special tools dedicated to pre- and post-processing data. [100]

2.5 Panchromatic Studies of Galaxies

The SDSS astrometry is very accurate ($\sim$ 0.1 arcsec) [101]. It can be used for panchromatic studies of galaxies aided by recent surveys at wavelengths outside the optical range ($0.3 - 1\mu m$). It additionally offers rich optical information which includes high-quality spectra and photometry combined with mid/far-IR wavelength range that offers important observational constraints for models of galaxy formation and evolution. [101] investigates the the panchromatic properties of 99,088 galaxies ($0.01 \leq z \leq 0.30$) selected from the SDSS DR1 spectroscopic sample. These galaxies were positionally matched to sources detected by ROSAT (X ray), GALEX (UV), 2MASS (near-IR), IRAS (mid/far-IR), GB6 (radio 6 cm), FIRST (radio 20 cm), NVSS (radio 20 cm), and WENSS (radio 92 cm) sky surveys.

Strong correlation is identified between the detection fraction at other wavelengths and optical properties such as flux, colors, and emission-line strengths. UV-IR broad-band SEDs for different types of galaxies were constructed using GALEX, SDSS, and 2MASS data and it was found that they form a nearly one-parameter family. For example, the SDSS u- and r- band data, supplemented with redshift, can be used to predict K-band magnitudes measured by 2MASS with an rms scatter of only 0.2 mag. IR-radio correlation study shows that the slope may be different for different galaxy types (AGN vs. star-forming galaxies) and is related to the H_α/H_β line strength ratio.

[42] discuss selection of QSO candidates from the combined SDSS and GALEX catalogues. The SDSS has produced the largest collection of QSOs to date. The SDSS filters alone do not allow a clean separation of QSOs from other blue objects. The GALEX UV surveys when combined with optical data better enables the identification of QSOs problem. XMM-Newton Distant Cluster Project (XDCP) initiated in 2003 is a sample of spectroscopically confirmed X-ray luminous high-redshift galaxy clusters comprising 22 systems in the range $0.9 < z \leq 1.6$. [102] The data includes X-ray properties and luminosity-based total mass estimates. Distant cluster candidates were followed-up with moderately deep optical and near-infrared imaging in at least two bands to photometrically identify the cluster galaxy populations and obtain redshift estimates based on colors of simple stellar population models.

[103] discovered sixty-eight Type 2 AGN candidates in the two Great Observatories Origins Deep Survey (GOODS) fields by using a previously known correlation between X-ray luminosity and X-ray-to-optical flux ratio. Thirty-one of those candidates qualify as QSO 2, that is, optically obscured quasars. The analysis involved X-ray and optical data catalogues. By going ~ 3 magnitudes fainter than previously known Type 2 AGN, a region of redshiftpower space has been sampled that was so far unreachable with classical methods.This is an example of statistical identification of sources using multiwavelength information.

The 6dFGS [6] sky survey represents a survey of redshifts and peculiar velocities of galaxies. The galaxies were selected mainly from the extended source catalog (XSC) 2MASS survey catalog. Selecting galaxies in the IR spectral range, where the effect of interstellar absorption inside the Milky Way is much smaller than in the optical range, allows much better studies of the object distribution at low galactic latitudes.

2.6 Data Mining Applications of Astronomy

Data Mining is the search for knowledge in the vast amounts of data that is too complex or too large to be analyzed by traditional techniques. It has lead to new research fields such as knowledge discovery and data warehouse. The term alludes to digging for gold (read "knowledge") in mines of data. It is the nontrivial extraction of implicit, previously unknown and potentially useful information from data. Alternative perspective is that it is the transfer of a set of data into an other state of aggregation that allows the user to potentially benefit from it [104].

Different types of knowledge mining which is a result of data mining and knowledge discovery based on certain prior knowledge of the domain. The result is usually in a form which is intuitive and easily-comprehensible for a human being. The knowledge extraction involves generalization, specialization and derivation. Using prior (P) and background knowledge (BK), the consequence is determined.

- Inference $P \cup BK \models C$
- Deduction Given P and BK, derive C
- Induction Given C and BK, hypothesize P
- Analogy If P' $\sim$ P, hypothesize C' $\sim$ C

Penn State Center for Astrostatistics maintains StatCodes - a meta-site with links to

[6] http://www-wfau.roe.ac.uk/6dFGS/

public domain software implementing statistical methods [105]. [106] covers statistical methods needed to efficiently analyze complex data sets from astronomical surveys such as the PanSTARRS, DES, and the upcoming LSST using Python code and example data sets from the SDSS. [107] provides C-based toolkit for data preparation prior to usage in analysis.

[108] publicly released in 2010 a blinded mix of simulated SNe, with types (Ia, Ib, Ic, II) selected in proportion to their expected rate as part of Supernova Photometric Classification Challenge. A spectroscopically confirmed subset was provided for training. The goals of this challenge were to (1) learn the relative strengths and weaknesses of the different classification algorithms, (2) use the results to improve classification algorithms, and (3) understand what spectroscopically confirmed sub-sets are needed to properly train these algorithms. Several different classification strategies resulted in similar performance, as reported in the result paper [109]. The most stable figure of merit versus redshift (for unconfirmed SNe) has $C_{FoM-Ia} = 0.3 - 0.45$ at all redshifts. The largest variation is $0.1 < C_{FoM-Ia} < 0.6$. Comparing the best figure of merit (vs. redshift) for each strategy shows that three strategies yield similar results: selection cuts, Bayesian probabilities and statistical inference. For all of the entries, the classification performance was significantly better for the spectroscopic training subset than for the unconfirmed sample. The degraded performance on the unconfirmed sample was in part due to participants not accounting for the bias in the spectroscopic training sample.

Some of the interesting examples of the current research in the astronomy domain involving solutions based on statistics and data-mining techniques include [110]:

- The distance problem (e.g., Photometric Redshift estimators)
- Star-Galaxy separation
- Cosmic-Ray detection in images
- Supernova detection and classification
- Morphological classification (galaxies, AGN, gravitational lenses among others)

- Class and Subclass Discovery (brown dwarfs, methane dwarfs among others)

[111] presents the data release for Galaxy Zoo 2 (GZ2), a citizen science project with more than 16 million morphological classifications of 304,122 galaxies drawn from the SDSS. Morphology is a powerful probe for quantifying a galaxys dynamical history; however, automatic classifications of morphology (either by computer analysis of images or by using other physical parameters as proxies) still have drawbacks when compared to visual inspection. GZ2 uses classifications from volunteer citizen scientists. While the original Galaxy Zoo project identified galaxies as early-types, late-types, or mergers, GZ2 measures finer morphological features. These include bars, bulges, and the shapes of edge-on disks, as well as quantifying the relative strengths of galactic bulges and spiral arms. The majority ($\geq 90\%$) of GZ2 classifications agree with those made by professional astronomers.

Chapter 3: Data Preparation

3.1 Telescope Measurements

Astronomers collect light and other radiation from celestial objects and use this information to interpret and develop a better understanding of the universe. The telescopes use filters and measure magnitudes. Magnitude is a number that measures the brightness of a star or galaxy. In magnitude, higher numbers correspond to fainter objects, lower numbers to brighter objects; the very brightest objects have negative magnitudes. When you say that a star has a certain magnitude, you must specify the color that the magnitude refers to. For example, SDSS measures magnitudes in five different colors by taking images through five color filters. A filter is a kind of screen that blocks out all light except for light with a specific color. The SDSS telescope's filters are green (g), red (r), and three colors that correspond to light not visible to the human eye: ultraviolet (u), and two infrared wavelengths (i and z); symbolized by u, g, r, i, and z on the SkyServer. These filters were chosen to view a wide range of colors, while focusing on the colors of interesting celestial objects. Color is symbolized by subtracting the magnitudes: u-g, g-r, r-i, and so on. Note that all these quantities involve magnitude, so they decrease with increasing light output.

Luminosity of an object is the amount of energy it emits per second. Apparent brightness or flux is the total energy received per second on each square meter of the observer's telescope.

$$F = \frac{L}{4\pi d_2} \tag{3.1}$$

The physical property that magnitude actually measures is radiant flux - the amount of light that arrives in a given area on Earth in a given time. Since color is measured by

magnitude, a star's color also depends on how much light arrives at Earth. Radiant flux is the physical basis for color. The definition of magnitude m in terms of radiant flux F is as shown in Equation (3.2).

$$m = -\log_{2.51} \frac{F}{F_{Vega}} \tag{3.2}$$

The star Vega in the northern hemisphere constellation Lyra is used as the standard for the magnitude system, so F_{Vega} means the amount of light arriving at Earth in a given time from Vega. This definition means that Vega's magnitude is set at zero through all filters. This does not mean that Vega looks the same through all filters; it just means that astronomers have agreed to use Vega as the zero point for the magnitude scale. Similar to how the freezing point of water is used as the zero point for the Celsius temperature scale. There's nothing special about Vega that made astronomers choose it as the zero point. The negative sign in the definition ensures that brighter stars have smaller magnitudes. So if Earth receives less light from a certain star than from Vega (through a given filter), that star's magnitude will be positive. If Earth receives more light from a certain star than from Vega, that star's magnitude will be negative.

For galaxy photometry, measuring flux is more difficult than for stars, because galaxies do not all have the same radial surface brightness profile and have no sharp edges. In order to avoid biases, SDSS needed to measure a constant fraction of the total light, independent of the position and distance of the object. Petrosian magnitude has this property. Two galaxies that have the same surface brightness profile shape but different central surface brightness have the same fraction of their flux represented in the Petrosian magnitude. Petrosian magnitude is a modified form of Petrosian (available as petroMag field in the Galaxy view of DR10 data) that measures the galaxy fluxes within a circular aperture whose radius is defined by the shape of the azimuthally averaged light profile. The Petrosian magnitudes should recover essentially all of the flux of an exponential galaxy profile and about 80% of

the flux for a de Vaucouleurs profile [1]. [112] Note: Identical galaxies seen at two different (luminosity) distances have fluxes related exactly as the inverse square of distance (in the absence of K-corrections).

Petrosian Radius is independent of distance and insensitive to reddening due to dust in the foreground [113] [112]. More detailed information about the Petrosian measurements is mentioned in Appendix B. Petrosian concentration index C is the ratio of Petrosian ninety-percent radius r_{90} to the Petrosian half-light radius r_{50}.

Based on photometric data, a sample of galaxies were selected for spectroscopic measurement in SDSS covering the wavelength range 3800Å- 9200Å. The main galaxy sample consisted of galaxies with r-band Petrosian magnitudes $r \leq 17.77$ and r-band Petrosian half-light surface brightnesses $\mu_{50} \leq 24.5$ mag $arcsec^{-2}$. These cuts resulted in about 90 galaxy targets per square degree, with a median redshift of 0.104. About 6% of galaxies that satisfy the selection criteria were not observed because they had a companion closer than the 55 arcsec minimum separation of spectroscopic fibers. The uniformity and completeness of the galaxy sample makes it ideal for studies of large-scale structure and the characteristics of the galaxy population in the local universe. [114] The spectroscopic measurement of redshift is used to compare against and estimate the predictive accuracy of the various photometric redshift algorithms.

Colors of galaxies reflect their dominant stellar populations and thus correlate with morphology. [115] and [116] demonstrated a tight correlation between the $u - r$ color, concentration of the galaxys light profile, and morphology. [116] used data for 456 bright galaxies recorded during the commissioning phase of the SDSS to examine the statistical properties of color indices, scale lengths, and concentration indices as functions of morphology for the SDSS photometric system. $u' - g'$, $g' - r'$, and $r' - i'$ colors of SDSS galaxies match well with those expected from the synthetic calculation of SED of template galaxies and with those transformed from $UBVR_CI_C$ color data of nearby galaxies. The agreement

[1] de Vaucouleurs' law (also called the de Vaucouleurs profile) describes how the surface brightness of an elliptical galaxy varies as a function of apparent distance from the center.

is somewhat poor,however, for the $i' - z'$ color band, with a discrepancy of $0.1 - 0.2$ mag. The half-light radius of galaxies depends slightly on the color bands. The inverse concentration index, defined by the ratio of the halflight Petrosian radius to the 90% light Petrosian radius, correlates tightly with the morphological type; this index allows us to classify galaxies into early (E/S0) and late (spiral and irregular) types. [115] studied 147,920 galaxies from SDSS and found that the distribution of galaxies in the $g* - r*$ versus $u* - g*$ color-color diagram is strongly bimodal and the two peaks correspond roughly to early- (E, S0, and Sa) and late-type (Sb, Sc, and Irr) galaxies. The colors of galaxies are correlated with their radial profiles, as measured by the concentration index and by the likelihoods of exponential and de Vaucouleurs' profile fits.

The principal goal of the 6dF is to study large-scale deviations in the velocity of galaxies from the homogeneous Hubble expansion. The distribution of such deviations provides the unique means to study mass distribution in the universe independent of the assumptions that galaxies follow the true mass distribution. For about 15,000 early-type galaxies evenly distributed over the southern sky, z-independent distances will be determined using the Fundamental Plane method (the fundamental plane is a three-parameter relation between photometric and kinematic characteristics of galaxies). Then, by comparing these distances with those derived from the observed values of z, it will be possible to estimate the peculiar velocities of galaxies arising due to inhomogeneities in mass distribution. [16]

3.2 Data Retrieval

The data is downloaded from Sloan Digital Sky Survey (SDSS) [2] SkyServer. Data Release 10 (DR10) is used for the experiments in the present study. DR11 and DR12 are the latest

[2]Funding for the SDSS has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Aeronautics and Space Administration, the National Science Foundation, the U.S. Department of Energy, the Japanese Monbukagakusho, and the Max Planck Society. The SDSS web site is http://www.sdss.org/. It is managed by the Astrophysical Research Consortium (ARC) for the Participating Institutions. The Participating Institutions are The University of Chicago, Fermilab, the Institute for Advanced Study, the Japan Participation Group, The Johns Hopkins University, Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, University of Pittsburgh, Princeton University, the United States Naval Observatory, and the University of Washington.

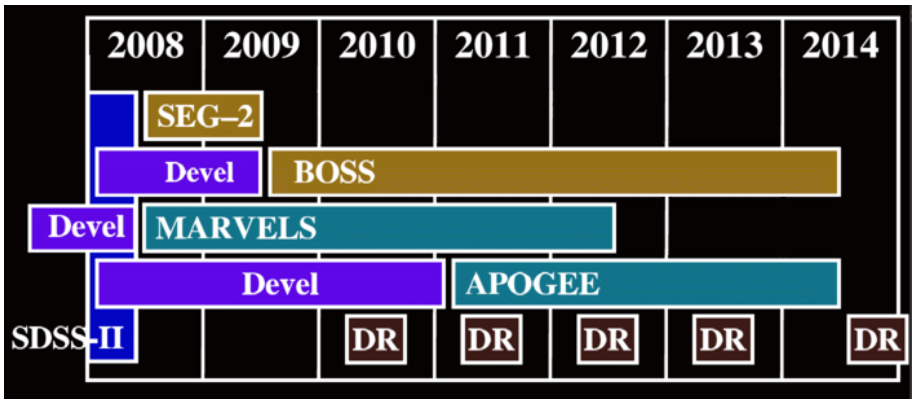


Figure 3.1: SDSS Surveys Timeline

release; made available simultaneously on December 2014. DR10 includes photometric redshift estimations for all galaxies. SDSS-III [117] consisted of four surveys executed on the same 2.5m telescope: the Apache Point Observatory Galactic Evolution Experiment (APOGEE), the Baryon Oscillation Spectroscopic Survey (BOSS), the Multi-Object APO Radial Velocity Exoplanet Large-area Survey (MARVELS), and the Sloan Extension for Galactic Understanding and Exploration 2 (SEGUE-2) [118].

BOSS focused on mapping the Universe on the largest scales, creating the largest volume three-dimensional map of galaxies ever created. SEGUE-2 and APOGEE focused on the structure and evolution of our own Milky Way galaxy. MARVELS searched very nearby stars for evidence of "exoplanets" surrounding them. The BOSS and SEGUE-2 programs required "dark" time when the Moon was less than 60% illuminated, or below the horizon. The APOGEE and MARVELS programs were executed during the remaining "bright" time. The MARVELS and BOSS spectroscopic surveys began in 2008 and 2009, and APOGEE began in 2011. SDSS-III collected data through the summer of 2014. The survey time-line and data releases schedule is shown in Figure 3.1 and Figure 3.2 respectively. All the data releases are available online and accessible through the CASJobs online interface with a SQL-like query language.

Astronomy-based data and manipulation results are handled in XML format. VOTable

Date	Data Release	APOGEE	BOSS	MARVELS	SEGUE-2
2011 Jan	DR8		Final imaging		Final spectra
2012 Jul	DR9		Spectra (up to 2011 Jul)		
2013 Jul	DR10	Spectra (up to 2012 Jul)	Spectra (up to 2012 Jul)		
2014 Dec	DR11	Spectra (up to 2013 Jul)	Spectra (up to 2013 May)		
2014 Dec	DR12	Final spectra	Final spectra	Final radial velocities	

Figure 3.2: SDSS Data Releases Dates

is the specific XML format for the exchange of VO data tables. The VO Registry provides a simple Google-like interface to search for data collections and catalogs of interest. A search for quasars returns 228 results today and this includes catalogs, images, cone search services, data-retrieval services, and publications along with metadata about each of these search results. The results can easily be filtered based on what the user needs. VO Inventory, VIM (VO Integration and Mining), and Datascope tools help the user find datasets that contain information relevant to a certain position in the sky and perform further analysis on the results found by combining related data from multiple sources. Various paths exist for the flow of data through these and other available VO tools.

Another XML format commonly used in astronomy data is the FITS format used for image-related data. FITS stands for "Flexible Image Transport System". It is endorsed by NASA and the International Astronomical Union. It is much more than just another image format and is used for the transport, analysis, and archival storage of scientific data sets. It has support for multi-dimensional arrays - 1D spectra, 2D images and 3D+ data cubes. The format includes header keywords which provide descriptive information or metadata about the image data.

3.3 Feature Selection

Actual mining of data only makes up 10% of the time required for the complete knowledge discovery process, preprocessing takes the rest. [107] Selecting the dataset for the experiments is crucial to the overall success of the investigation being undertaken. If the training set is a good representative of the problem under investigation, then the performance of the learning algorithm will naturally improve and converge. Additionally, the "good" sample can be used to impute the missing values with much higher confidence. Where possible and relevant. the performance of the different estimation methods should be tested against a bad representation of data or unseen data inorder to determine if any of the methods are able to overcome the influence of "bad" data-points.

SDSS data includes different types of magnitude measurements in the ugriz bands - Fiber, PSF, Model and Petrosian (refer fiberMag_*, psfMag_*, modelMag_* and petroMag_* attributes respectively in Table 3.1). All magnitudes are based on the concept of using circularized brightness profiles extracted for a predefined set of radii. There are multiple types of magnitudes because depending on an objects brightness profile, they have different noise properties. Fiber magnitudes reflect the flux contained within the aperture of an spectroscopic fiber in each band. It assumes an aperture appropriate to the SDSS spectrograph (3" in diameter). [119] For galaxy photometry, measuring flux is more difficult than for stars, because galaxies do not all have the same radial surface brightness profile, and have no consistently distinct edges.

In order to avoid biases, SDSS measures a constant fraction of the total light, independent of the position and distance of the object. To satisfy these requirements, the SDSS adopted a modified form of the Petrosian (1976) system, measuring galaxy fluxes within a circular aperture whose radius is defined by the shape of the azimuthally averaged light profile. The images of overlapping galaxies are deblended using a robust code that conserves flux. Under most conditions, the $cmodelMag_*$ magnitude is a reliable estimate of the galaxy flux that accounts for the effects of local seeing and thus, less dependent on local seeing variations. However, for nearby galaxies, that is, galaxies bright enough to be included

in the SDSS spectroscopic sample, have relatively high signal-to-noise ratio measurements of their Petrosian magnitudes. Since these magnitudes are model-independent and yield a large fraction of the total flux, roughly constant with redshift, Petrosian magnitude is the measurement of choice for such objects. Thus, cModel_* attribute is not included in the present study. [119] Appendix B discusses the associated Petrosian radius readings with Petrosian magnitude measurements. Associated with Petrosian magnitude, there are three radius that are measured - Petrosian radius, Petrosian radius containing 50% of Petrosian flux and Petrosian radius containing 90% of Petrosian flux (refer petroRad_*, petroR50_* and petroR90_* attributes respectively in Table 3.1).

SpecObj and *Galaxy* are the database views from SDSS DR10 used in the data retrieval query as shown in Appendix A. **SpecObj** view has the spectroscopic data of sky objects that are properly filtered based on data-cleanliness from the *SpecObjAll* table. Sprectroscopic redshift measurements are available in this table to compare the accuracy of the photometric redshift estimates. *PhotoObjAll* is the full photometric catalog quantities for SDSS imaging. This table contains one entry per detection, with the associated photometric parameters. *PhotoObjAll* table has the following views:

- *PhotoObj*: all primary and secondary objects; essentially this is the view you should use unless you want a specific type of object.
- *PhotoPrimary*: all photo objects that are primary (the best version of the object).
  - *Star*: Primary objects that are classified as stars.
  - **Galaxy**: Primary objects that are classified as galaxies.
  - *Sky*:Primary objects which are sky samples.
  - *Unknown*:Primary objects which are no0ne of the above
- *PhotoSecondary*: all photo objects that are secondary (secondary detections)
- *PhotoFamily*: all photo objects which are neither primary nor secondary (blended)

Note: The calculated columns from the data of the SDSS tables are shown in Table 3.3.

Table 3.1: Relevant Columns from Galaxy View of SDSS DR 10 [120]

Column Name	Type	Length	Unit	Description
specObjid	bigint	8		Unique database ID
u	real	4	mag	Shorthand alias for modelMag_u
g	real	4	mag	Shorthand alias for modelMag_g
r	real	4	mag	Shorthand alias for modelMag_r
i	real	4	mag	Shorthand alias for modelMag_i
z	real	4	mag	Shorthand alias for modelMag_z
psfMag_u	real	4	mag	PSF magnitude (u)
psfMag_g	real	4	mag	PSF magnitude (g)
psfMag_r	real	4	mag	PSF magnitude (r)
psfMag_i	real	4	mag	PSF magnitude (i)
psfMag_z	real	4	mag	PSF magnitude (z)
fiberMag_u	real	4	mag	Flux in 3 arcsec diameter fiber radius (u)
fiberMag_g	real	4	mag	Flux in 3 arcsec diameter fiber radius (g)
fiberMag_r	real	4	mag	Flux in 3 arcsec diameter fiber radius (r)
fiberMag_i	real	4	mag	Flux in 3 arcsec diameter fiber radius (i)
fiberMag_z	real	4	mag	Flux in 3 arcsec diameter fiber radius (z)
petroMag_u	real	4	mag	Petrosian Magnitude (u)
petroMag_g	real	4	mag	Petrosian Magnitude (g)
petroMag_r	real	4	mag	Petrosian Magnitude (r)
petroMag_i	real	4	mag	Petrosian Magnitude (i)
petroMag_z	real	4	mag	Petrosian Magnitude (z)
petroRad_u	real	4	arcsec	Petrosian Radius (u)
petroRad_g	real	4	arcsec	Petrosian Radius (g)
petroRad_r	real	4	arcsec	Petrosian Radius (r)
petroRad_i	real	4	arcsec	Petrosian Radius (i)
petroRad_z	real	4	arcsec	Petrosian Radius (z)
petroR50_u	real	4	arcsec	Petrosian Radius containing 50 percent of Petrosian flux (u)
petroR50_g	real	4	arcsec	Petrosian Radius - 50 percent (g)
petroR50_r	real	4	arcsec	Petrosian Radius - 50 percent (r)
petroR50_i	real	4	arcsec	Petrosian Radius - 50 percent (i)
petroR50_z	real	4	arcsec	Petrosian Radius - 50 percent (z)
petroR90_u	real	4	arcsec	Petrosian Radius containing 90 percent of Petrosian flux (u)
petroR90_g	real	4	arcsec	Petrosian Radius - 90 percent (g)
petroR90_r	real	4	arcsec	Petrosian Radius - 90 percent (r)
petroR90_i	real	4	arcsec	Petrosian Radius - 90 percent (i)
petroR90_z	real	4	arcsec	Petrosian Radius - 90 percent (z)

Table 3.2: Relevant Columns from SpecObj View of SDSS DR 10 [120]

Column Name	Type	Length	Description
specObjid	bigint	8	Unique database ID
zWarning	int	4	Bitmask of warning values; 0 means all is well; e.g., zWarning = dbo.fSpecZWarning('OK')
sourceType	varchar	128	Type of object e.g., 'GALAXY'
z	real	4	Final Redshift
zErr	real	4	Redshift Error

Table 3.3: Calculated Columns from Galaxy View columns of SDSS DR 10 [120]

Column Name	Formula	Description
fp_u	$\frac{fiberMag\_u}{petroMag\_u}$	$\frac{FiberMagnitude}{PetrosianMagnitude}$
fp_g	$\frac{fiberMag\_g}{petroMag\_g}$	$\frac{FiberMagnitude}{PetrosianMagnitude}$
fp_r	$\frac{fiberMag\_r}{petroMag\_r}$	$\frac{FiberMagnitude}{PetrosianMagnitude}$
fp_i	$\frac{fiberMag\_i}{petroMag\_i}$	$\frac{FiberMagnitude}{PetrosianMagnitude}$
fp_z	$\frac{fiberMag\_z}{petroMag\_z}$	$\frac{FiberMagnitude}{PetrosianMagnitude}$
pR_u	$\frac{petroR50\_u}{petroR90\_u}$	$\frac{Radiuscontaining50percentofPetrosianflux}{Radiuscontaining90percentofPetrosianflux}$
pR_g	$\frac{petroR50\_g}{petroR90\_u}$	$\frac{Radiuscontaining50percentofPetrosianflux}{Radiuscontaining90percentofPetrosianflux}$
pR_r	$\frac{petroR50\_r}{petroR90\_u}$	$\frac{Radiuscontaining50percentofPetrosianflux}{Radiuscontaining90percentofPetrosianflux}$
pR_i	$\frac{petroR50\_i}{petroR90\_u}$	$\frac{Radiuscontaining50percentofPetrosianflux}{Radiuscontaining90percentofPetrosianflux}$
pR_z	$\frac{petroR50\_z}{petroR90\_u}$	$\frac{Radiuscontaining50percentofPetrosianflux}{Radiuscontaining90percentofPetrosianflux}$
C_u	$\frac{1}{pR\_u}$	$\frac{Radiuscontaining90percentofPetrosianflux}{Radiuscontaining50percentofPetrosianflux} = Concentration$
C_g	$\frac{1}{pR\_g}$	$\frac{Radiuscontaining90percentofPetrosianflux}{Radiuscontaining50percentofPetrosianflux} = Concentration$
C_r	$\frac{1}{pR\_r}$	$\frac{Radiuscontaining90percentofPetrosianflux}{Radiuscontaining50percentofPetrosianflux} = Concentration$
C_i	$\frac{1}{pR\_i}$	$\frac{Radiuscontaining90percentofPetrosianflux}{Radiuscontaining50percentofPetrosianflux} = Concentration$
C_z	$\frac{1}{pR\_z}$	$\frac{Radiuscontaining90percentofPetrosianflux}{Radiuscontaining50percentofPetrosianflux} = Concentration$
ug	$u - g$	$Color1$
gr	$g - r$	$Color2$
ri	$r - i$	$Color3$
iz	$i - z$	$Color4(zismodelMag\_z)$

3.4 Photometric Redshift and SDSS

Two alternative methods for photometric redshift estimation are used in DR10 [121] - kd-tree nearest neighbor fit (KF) is stored in the table *Photoz* and random forests (RF) is stored

in the table *PhotozRF*. The estimators use colors and inclination angle (*expAB_r*[3] in the *PhotoObj* table) of each galaxy. Although using inclination angle does not significantly improve the overall estimation, it does remove a systematic bias. The training set contains more than 850,000 galaxies from the DR8 spectroscopic catalog (average r magnitude 17.3), and an additional 14,000 galaxies from other spectroscopic redshift surveys that include deeper (up to redshift of 1) and fainter (average r magnitude 20.75) galaxies. The RMS of the estimation errors for the two parts of the training set are 0.018 and 0.103, respectively. The more than fivefold increase of error for the faint subset is mostly due to the larger photometric errors in their associated measurements. Both KF and RF provide an explicit estimate of the redshift errors (zErr) and it has been found to be reliable and unbiased.

The query to lookup the data from *Photoz* and *PhotozRF* along with their associated spectroscopic redshift (spec-z) from *SpecObj* is as shown in 3.4. [122] This includes a total of 1,788,471 records. The distribution of the spectroscopic redshift (mean = 0.32, 1st quartile = 0.11, 3rd quartile = 0.51, max = 7.05) and photometric redshift data (KF: mean = 0.31, 1st quartile = 0.11, 3rd quartile = 0.51, max = 1.0; RF: mean = 0.31, 1st quartile = 0.11, 3rd quartile = 0.49, max = 0.88) are similar for the two photo-z tables. Refer Figure 3.3 for the distribution of the estimates from the kd-tree algorithm. Catastrophic error is present in all the ranges including $z < 0.3$ (median of spec-z) but increases for $z > 0.7$. This can be explained by the impact of the lack of adequate amount of training data on the predictive accuracy of the algorithm.

The estimates are bound to $z < 1.0$ for higher redshifts ($spec - z > 1.0$). Degeneracy in photo-z estimation is prevalant in both the algorithms in all the ranges, irrespective of the amount of training data in the given range. This can be noted from the spread of the boxplots for spec-z vs. photo-z as shown in Figure 3.4 for KF and Figure 3.6 for RF. Looking in detail at the spec-z vs. photo-z specifically for the range with largest number of records (thus, the impact due to lack of training data on the predictive accuracy of the algorithm is minimized), that is $z = (0, 1.0]$, the degeneracy is still noticeable - refer Figure

[3]This attribute is not used in the present study.

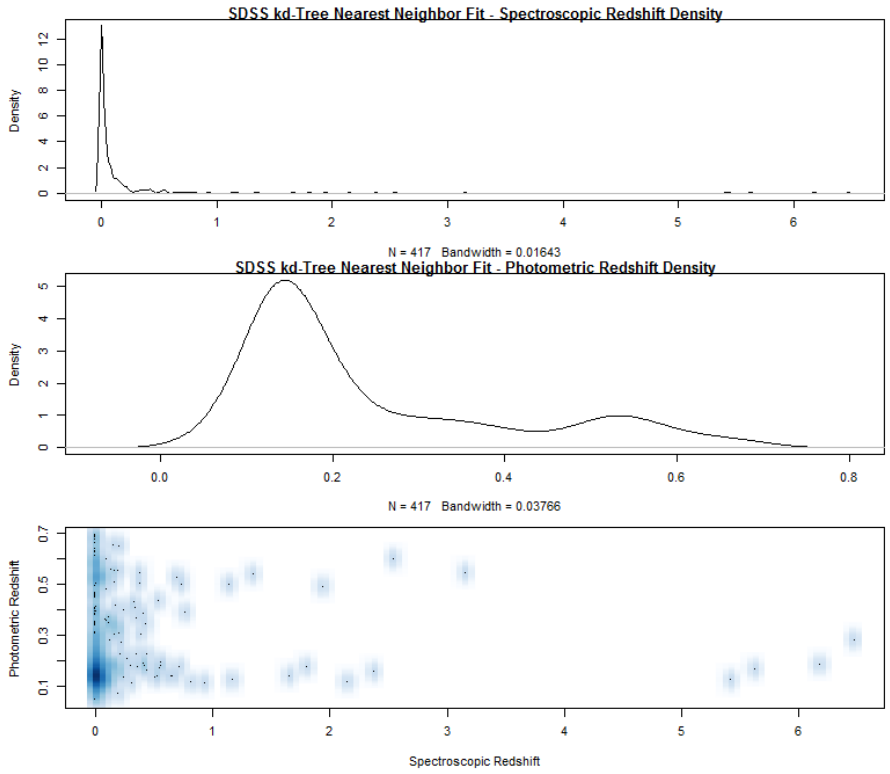


Figure 3.3: SDSS kd-tree Photo-z Data Distribution

3.5 for KF and Figure 3.7 for RF.

```
select sz.bestObjID, pz.nnCount, sz.z as zSpec, sz.zErr as zSpecErr,
        pz.z as zPhoto, pz.zErr as zPhotozErr
from Photoz pz, SpecObj sz
where pz.objID = sz.bestObjID;

select sz.bestObjID, sz.z as zSpec, sz.zErr as zSpecErr,
         pz.z as zPhoto, pz.zErr as zPhotozErr
from PhotozRF pz, SpecObj sz
where pz.objID = sz.bestObjID;
```

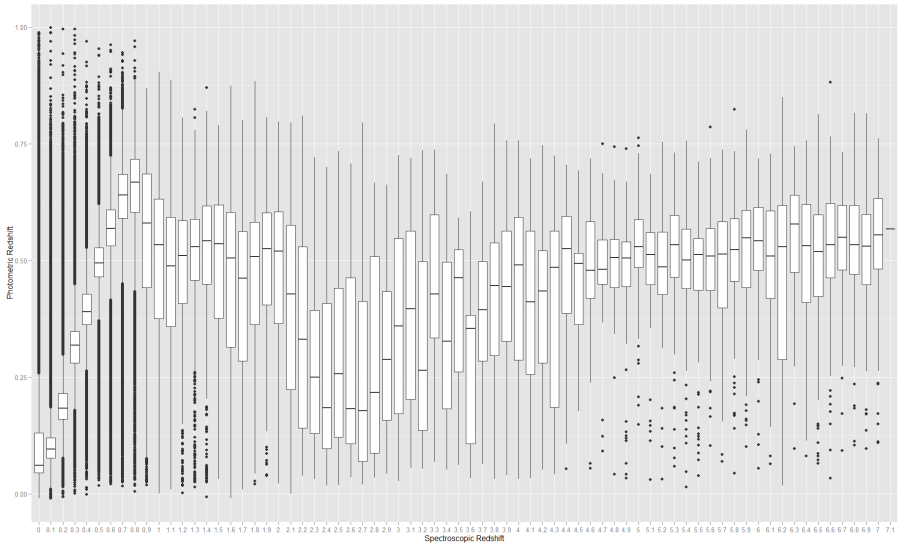


Figure 3.4: SDSS kd-tree Spec-z vs. Photo-z Degeneracy

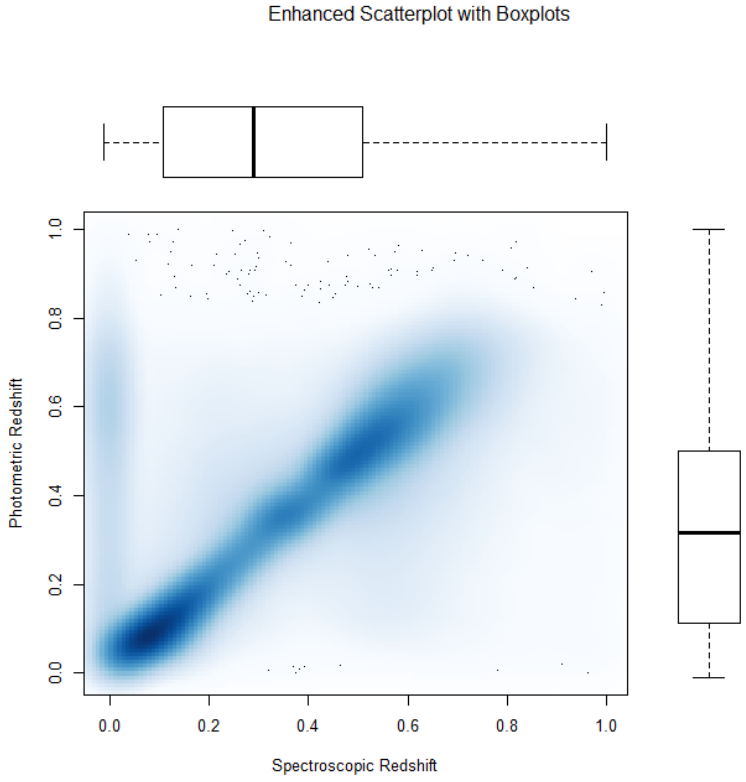


Figure 3.5: SDSS kd-tree Spec-z vs. Photo-z

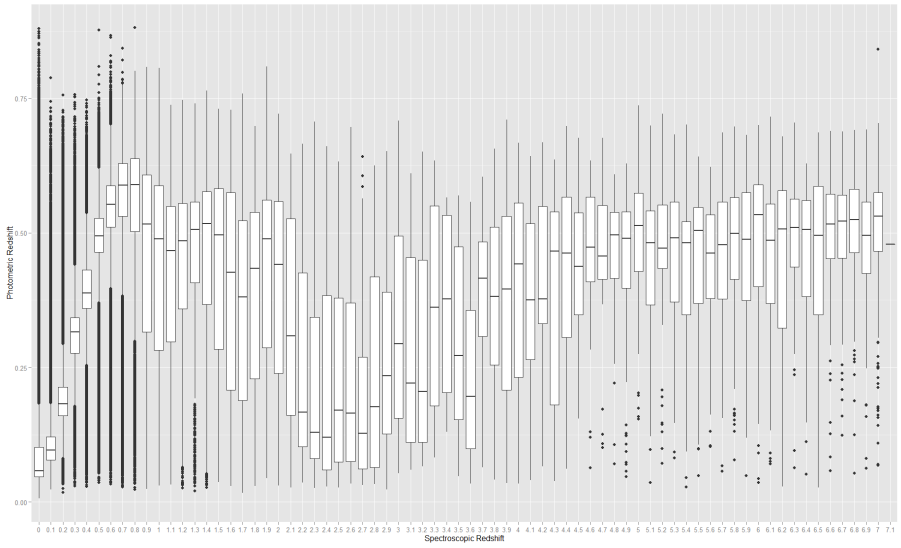


Figure 3.6: SDSS Random Forest Spec-z vs. Photo-z Degeneracy

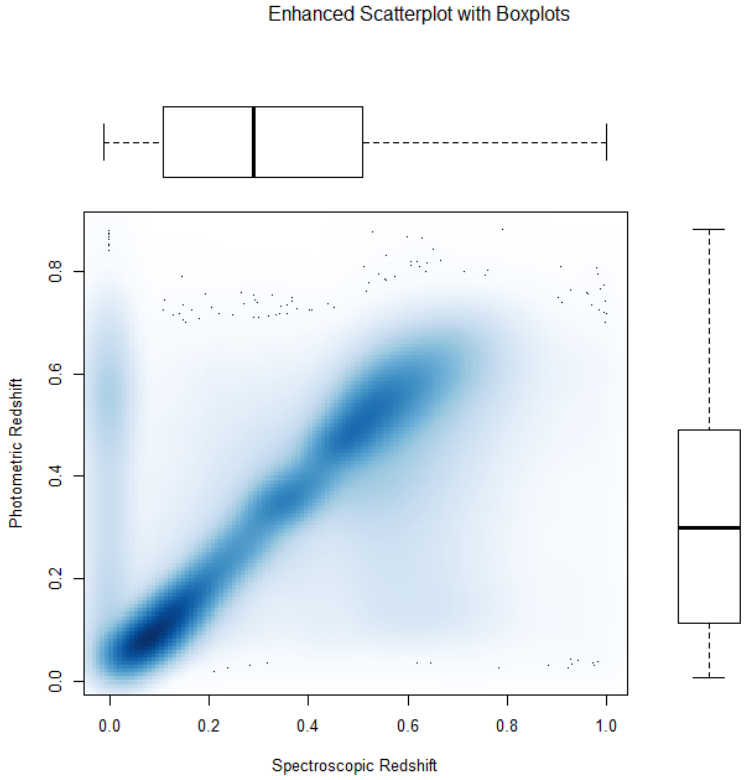


Figure 3.7: SDSS Random Forest Spec-z vs. Photo-z

3.5 Accuracy of Previous Solutions

MegaZ-LRG DR6 Catalogue is available online. [93] This catalogue was generated as a result of a comparative study of six photometric redshift methods applied to 1.5 million luminous red galaxies. [84] The six methods are ANNz, BPz Bayesian, BPz ML, Hyper-z, LePhare, ZEBRA. The methods were applied to data available in SDSS DR6 and compared with the photometric redshift (photo-z) estimate available in SDSS. The catalog lists the object id alongwith the photo-z estimates from the different methods and any applicable configuration parameters.

Since the study compared the estimates against the SDSS photo-z estimate, the predictive performance was relative to the SDSS photo-z. As part of the present study, we study the performance against the spectroscopic redshift. This study was particularly important after the analysis in Section 3.4 improved our understanding of the catastrophic error and degeneracy involved in the two different SDSS photo-z estimates (kd-tree and random forest).

The catalgue was split into multiple files and only required redshift attributes were loaded. This enabled faster and smooth loading of the data with the data cleanup being applied on one subset at a time, thereby reducing the memory footprint of the data merging process. SDSS CAS jobs was used to find the spectroscopic redshift from an older data release, that is, DR6 (DR10 is the current release). Since DR6 is not as integrated with the existing CAS jobs as compared to DR10, the table join with the object id from MegaZ-LRG was not straight-forward. The query steps are outlined in 3.5.

```
Upload the object ids from MegaZ-LRG DR6 into MyTable (associated with user profile↩
     database MyDB)
MyDB:  select min(objID) from MyTable; --587722952230174848
MyDB:  select max(objID) from MyTable; --588848901539103104

DR6:  select count(*) from SpecObj where bestObjID >= 587722952230174848 and ↩
    bestObjID <= 588848901539103104; -- 880427
```

```
DR6:  select bestObjID, z, zErr from SpecObj where bestObjID >= 587722952230174848 ↩
    and bestObjID <= 588848901539103104; -- 880427 in table MyDB.DR6_SpecObj

MyDB: select spct.z, spct.zErr, myID.objID from MyDB.DR6_SpecObj spct, MyDB.MyTable↩
     myID where spct.bestObjID = myID.objID;
```

The data is largely concentrated in the redshift range of (0 - 0.5) with measurements spanning the wider range of (0 - 5). The distribution of the data in the range of z = 0-0.5 is shown in Figure 3.8. All the six methods have similar predictive accuracy performance with degeneracy in all the ranges of redshift. The catastrophic error was particularly bad in the z = 0 - 0.2, strangely where the largest amount of data is available. The error analysis across the different redshift ranges for ANNz is shown in Table 3.4; the numbers are closely similar in all six methods.

Table 3.4: MegaZ-LRG DR6 Catalogue - ANNz Photo-z Error Analysis

Range	**Count**	**Mean z-Spec**	**Mean z-Photo**	**RMS**	**Catastrophic Error %**
(0-0.1]	1272	0.051	0.509	0.463	100.00
(0.1-0.2]	947	0.138	0.512	0.379	100.00
(0.2-0.3]	340	0.261	0.487	0.241	100.00
(0.3-0.4]	475	0.359	0.499	0.157	68.63
(0.4-0.5]	1160	0.446	0.478	0.063	10.17

The methods estimated the redshift based on Dereddened model magnitudes and de Vaucouleurs magnitude. It is likely these measurements were not as robust in the earlier data releases as compared to the magnitude measurements in more recent releases such as DR8 or DR10. Thus, when these measurements were used to predict redshift, it resulted in odd and strange patterns leading to catastrophic errors as shown for ANN-z in Figure 3.9. The availability of training data in a given range has no effect on the accuracy associated with the learner in that range. Figure3.10 shows the degeneracy with the ANN-z estimates.

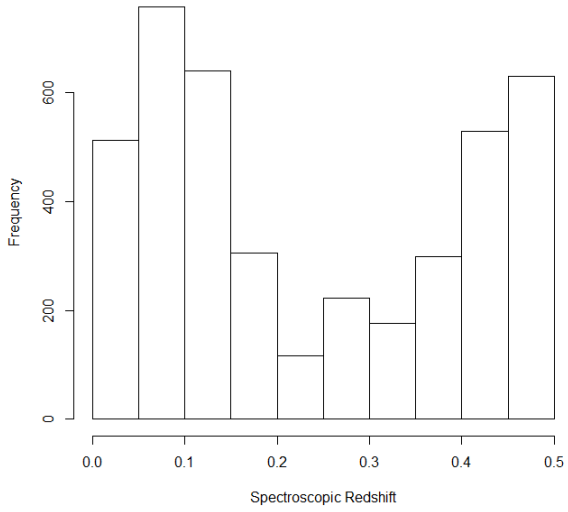


Figure 3.8: MegaZ-LRG DR6 Spectroscopic Redshift Distribution $z = 0 - 0.5$

3.6 Sampling Methods

Sampling method is important to get a training set and test set that is a good representative of the overall distribution. Various methods have been studied inorder to enable this. The bootstrap brings to bear various desirable features in the massive data setting, notably its relatively automatic nature and its applicability to a wide variety of inferential problems. It can be used to assess bias, to quantify the uncertainty in an estimate (e.g.,via a standard error or a confidence interval), or to assess risk. However, these virtues are realized at the expense of a substantial computational burden. Bootstrap-based quantities typically must be computed via a form of Monte Carlo approximation in which the estimator in question is repeatedly applied to resamples of the entire original observed dataset.

Bag of Little Bootstraps (BLB) applies the bootstrap to each small subset, where in the resampling process of each individual bootstrap run, weighted samples are formed such that the effect is that of sampling the small subset n times with replacement, but the computational cost is that associated with the size of the small subset. This has the effect that,

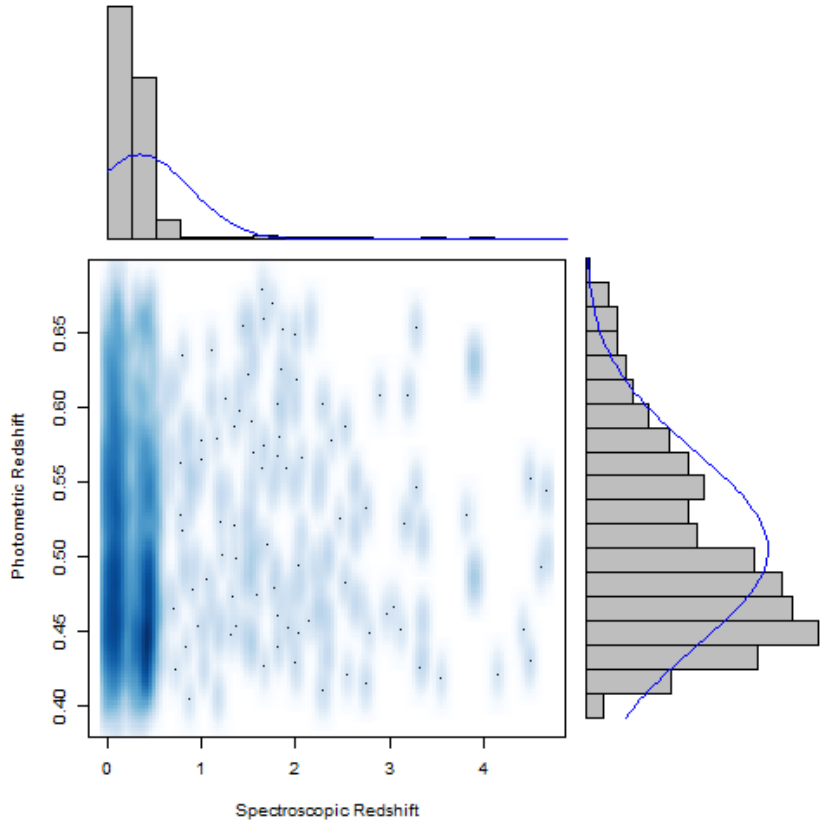


Figure 3.9: Catastrophic Error Analysis - ANNz in MegaZ-LRG DR6

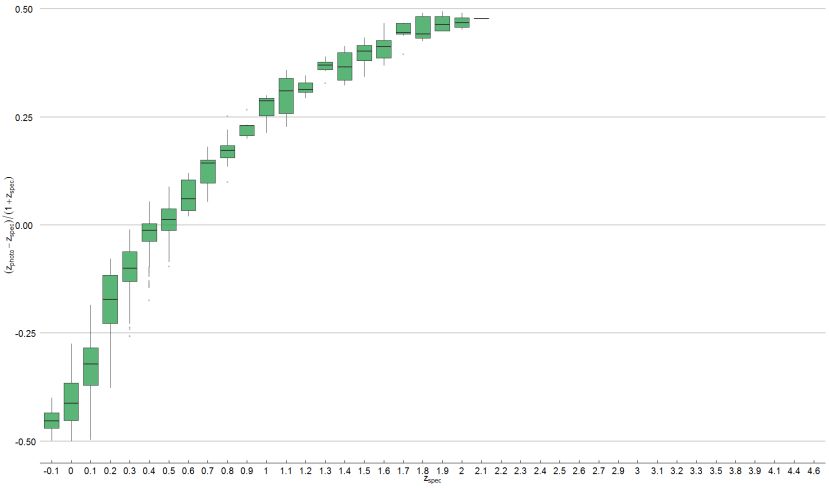


Figure 3.10: Spec-z vs. Photo-z - ANNz in MegaZ-LRG DR6

despite operating only on subsets of the original dataset, BLB does not require analytical rescaling of its output, unlike the m out of n bootstrap method. Overall, BLB has a significantly more favorable computational profile than the bootstrap, as it only requires repeated computation of the estimator under consideration on quantities of data that can be much smaller than the original dataset. As a result, BLB is well suited to implementation on modern distributed and parallel computing architectures which are often used to process large datasets. Also, our procedure maintains the bootstraps generic applicability, favorable statistical properties (i.e., consistency and higher-order correctness), and simplicity of implementation. Finally, as we show in experiments, BLB is consistently more robust than alternatives such as the m out of n bootstrap and subsampling.[123] It is necessary to develop data driven methods of selection of m which lead to reasonable results over situations where both the bootstrap works and where it doesn't. [124]

Cross-validation is generating multiple subset of the available data and applying the learning method against the subset and studying the predictive accuracy across the multiple subsets. In the ideal case, the predictive accuracy is independent of the training subset. In the real world, this is rarely true - there will be range of accuracy for the different subsets but that range is desired to be narrow. The data is split 90%-10% in all the experiments in the present study using the 700,777 galaxies dataset from SDSS. This results in the training set consisting of 630699 observations and the test set consisting of 70078 observations. The predictive accuracy vs. additional time involved due to a certain sampling method tradeoff analysis need to be performed. The reported results are based on random sampling and making sure the distribution of the training and test subset match as shown in Figure 3.11.

The Bayesian Network R package *bnlearn* provides a *cpdist* query to derive the posterior distribution based on priors (aka evidence). It uses logical sampling to estimate the distribution and the number of sample points can be provided as a parameter. The reported data are based on a sampling size of 50,000 data points. If the distribution is all null, the search is repeated for a maximum of ten runs. The execution moves to the next test data point if a distribution is identified.

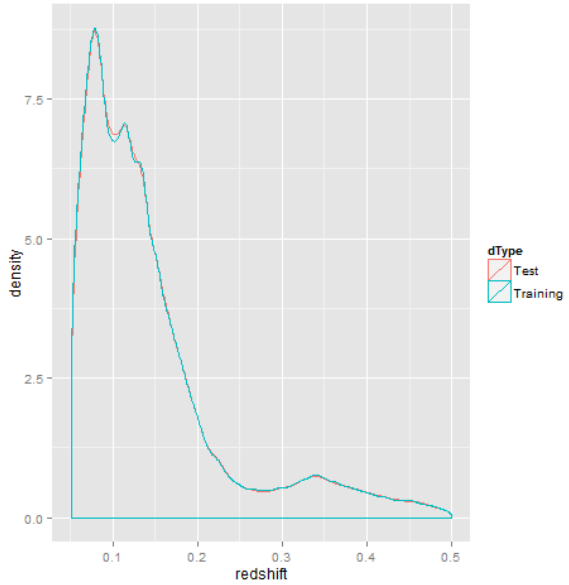


Figure 3.11: Training and Test Dataset Distribution

3.7 Predictive Accuracy Measures

Occam's Razor states that among competing hypotheses that predict equally well, the one with the fewest assumptions should be selected. Other, more complicated solutions may ultimately prove to provide better predictions, butin the absence of differences in predictive abilitythe fewer assumptions that are made, the better. It is important to be able to consistently measure the predictive accuracy of an estimation method. Since spectroscopic redshift is the true redshift and we have that measured for the 700,777 galaxy records being used in our study, we should use it in comparisons. The photometric redshift should be as close to the spectroscopic redshift as possible and not diverge far.

The upcoming LSST ugrizy filter system will involve redshift estimation from $0 < z < 6$ and the following are the required metrics:

- RMS scatter in uncertainties, $\frac{\sigma z}{(1+z)} < 0.05$
- Fraction of 3σ outliers below 10%

- Bias, $e_z = \frac{zphot - zspec}{(1+zspec)} < 0.003$

- Uncertainties on σz must be known to better than 1%.

Based on the above guidelines, the metrics for the different estimation methods were measured in the redshift ranges that are more heavily populated than others. The metrics are count of records, mean spectrocopic redshift, mean photometric redshift, mean error, mean absolute error, bias (also termed as catastrophic error percent). The ranges are from "(0-0.1]" through "(0.1-0.2]" to "(0.4-0.5]". The Bayesian method also includes a certainty value associated with the estimate. The certainty is the posterior probability associated with the estimated value.

3.8 Software Used

The following software were used in running the experiments and performing data analysis:

- RStudio Desktop v0.98.1062 using R version 3.1.1 (2014-07-10)

- R package Shiny 0.10.2.1 is a web application framework for R. Using two files ui.r and server.r, a dynamic UI that responds to user input and performs data analysis based on the input can be developed. Long term goal is to provide a web application for a part of the experiments of the present study.

- Revolution Analytics R provides the parallel framework. An initial attempt has been made at parallelizing the estimation process. Other parts of the workflow should be considered for parallelization to improve the time required for the different computations.

- R Bayesian Network package bnlearn [125], discretization, RgraphViz was used extensively for Bayesian Network-based method used in this study.

- BRMLtoolbox [126] is a Matlab-based package for Bayesian Network and Machine Learning implementations.

- BayesiaLab Software Package has been in development since 2001 and it provides a neat interface to perform the Bayesian Network data analysis. They provide demo license use only for a month.
- RapidMiner, QlikView and QlikSense, GoogleVis were explored for potential use in data analysis and visualization.
- R package Sweave embeds the R code in LaTeX documents [127][128][129][130]; it generates tex file from .snw file. R package Stangle extracts the R code from .snw file into an R source file.
- R package ggplot2, dplyr, reshape2 for visualization tools
- R package smoteboost as a boosting technique for high redshift regions with insufficient data was explored.

Chapter 4: Generalized Linear Model (GLM) Photomorphic Redshift

4.1 Generalized Linear Model (GLM)

Ordinary linear regression predicts the expected value of a given unknown quantity (the response variable, a random variable) as a linear combination of a set of observed values (predictors). This implies that a constant change in a predictor leads to a constant change in the response variable (i.e. a linear-response model). This is appropriate when the response variable has a normal distribution. In cases where the response variable is expected to be always positive (similar to redshift measurement) and varying over a wide range, constant input changes lead to geometrically varying, rather than constantly varying, output changes. General linear models are not suited for situations where there are restrictions on Y (e.g., binary, count or strictly positive data) or when the variance depends on the mean. The GLMs are a generalisation of this framework, capable of handling both scenarios.

In a generalized linear model (GLM), each outcome of the dependent variables, Y, is assumed to be generated from a particular distribution in the exponential family, a large range of probability distributions that includes the normal, binomial, Poisson and gamma distributions, among others. The procedure relies on one main feature of the PDF: within the chosen family, a distribution should be uniquely identified through one single parameter (called location or mean). Determining this parameter is the ultimate goal of the GLM methodology. All GLMs share a similar structure and are characterized by the following:

- A random response component whose mean μ is to be estimated. The response variable, Y, is assumed to be theoretically derived as a random sample of an underlying single parameter PDF belonging to the GLM family of distributions. The goal of

modelling Y is to find an unbiased estimate of the mean parameter which better describes the data.

- A systematic (or linear) component built from the explanatory variables, X (sometimes called covariates), and their associated slope coefficients. Their multiplication produces a linear predictor for each observation.

- A link function which defines how the mean is associated with the explanatory variable. The link function linearises the relationship between the mean response and predictors (X_i)

[65] used GLM modeled on principal components in estimating the photometric redshifts of galaxies from their multi-wavelength photometry. The magnitudes of galaxies can be strongly correlated across different broadband filters. Principal Component Analysis is performed to avoid this multicollinearity and the principal components (PCs) of the observed magnitude are set as the explanatory variables. Additionally, the PCs also optimize the use of computational resources, as the calculation time required increases non-linearly with the number of explanatory variables. A set of photo-z packages which can be used on any multi-wavelength data set was developed in R and Python. The R package CosmoPhotoz provided by [65] has been used to study the impact of using morphology attributes in addition to color and magnitude measurements on the predictive accuracy of photo-z estimation. The following four subsets of attributes were investigated to compare and contrast their relative impact - color (four), color and magnitude (twenty), color and morphology (twenty) and six attributes with highest correlation (Ratio of Fiber to Petrosian magnitude in r-band and g-band, Petrosian magnitude in u-band and g-band, and Fiber magnitude in g-band and u-band) from the entire set of fourty-five attributes.

The model was based on two-third of the $z \leq 0.5$ data of 700777 galaxies, that is, 467184 galaxies. The test was on the remaining one-third of the galaxies, that is, 233593 galaxies. Test was also run on data from the unseen range of $z > 0.5$ in the modeling phase; this consists of 1730 galaxies with maximum $z = 5.4$. The catastrophic error rate in [65] was

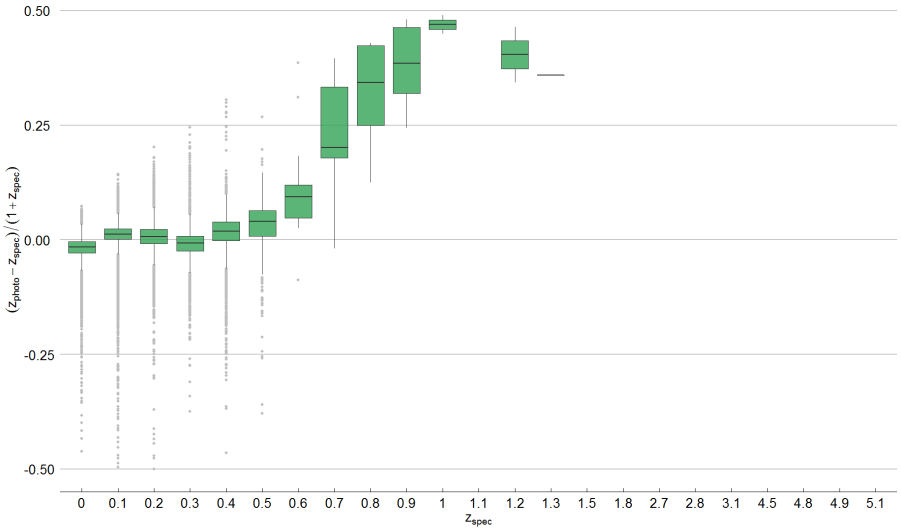


Figure 4.1: GLM - Color - Training set = Half dataset

found to reach $1-5\%$, comparable with current techniques involving template fitting or ML. Thus, if the catastrophic error on the different ranges with different sets of attributes will give us a better idea of how applicable GLMs are for photo-z estimation.

4.2 Color

This study was performed with three as the number of principal components. The model trained with half the data performed worse than the model trained with two-third of the data. Refer Figure 4.1 vs. Figure 4.2 to observe the larger spread of estimates in the regions with lesser amount of training data. It is more noticeable in $z > 0.7$.

Note: All the experiments were run using spectroscopic redshift with precision three, for example, spec-z = 0.314. Better performance of these method runs are expected when using spectroscopic redshift with precision one and two since there will be larger amount of data per redshift value. This will also naturally lead to higher certainty for a estimate.

In our study with the galaxy data from SDSS as retrieved by using the query A, the catastrophic error rate is $0.2-0.3\%$ in the $z = 0-0.2$ region with the majority of the

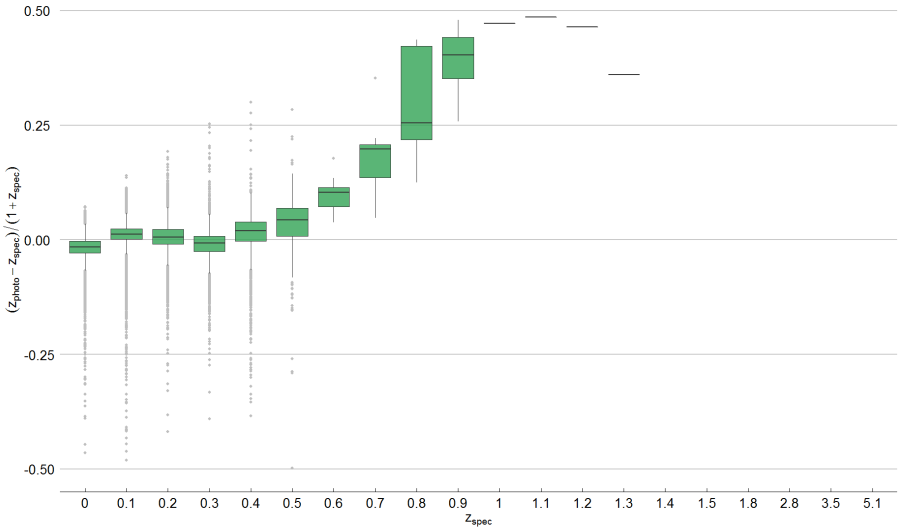


Figure 4.2: GLM - Color - Training set = Two-third dataset

data. It is worse at $2.8 - 3\%$ in $z = 0.3 - 0.5$ region with comparatively lesser amount of data than the earlier region. This error rate is yet again better than $1 - 5\%$ of the existing other techniques. The catastrophic error rate is worse than existing standards in the region $z = 0.4 - 0.5$ that has less than half the number of records in the immediate earlier region of $z = 0.3 - 0.4$. The GLM model is better fit when more data is available and the fit is quantified by the percent of objects being estimated with catastrophic error. For unseen data in training of $z > 0.5$ range, the model does not converge and is thus not able to estimate for objects with $z > 1.3$. Refer the error summary table in Table 4.1. The comparative test performance when dealing with data in similar range as training data range $z = 0 - 0.5$ is shown in Figure 4.3 while the test performance when dealing with data in beyond training data range of $z > 0.5$ is shown in Figure 4.4. The performance does not improve if two-digit precision is used for redshift instead of three-digit precision.

Since the performance for the range "(0.4-0.5]" is not as good as the other ranges, a closer look was taken at the data distribution. It was found that the data could be split into two different distribution when comparing color $u - g$ vs. color $g - r$ as shown in Figure 4.5. It was found that if a different model was formulated for $z > 0.33$, it lead to better

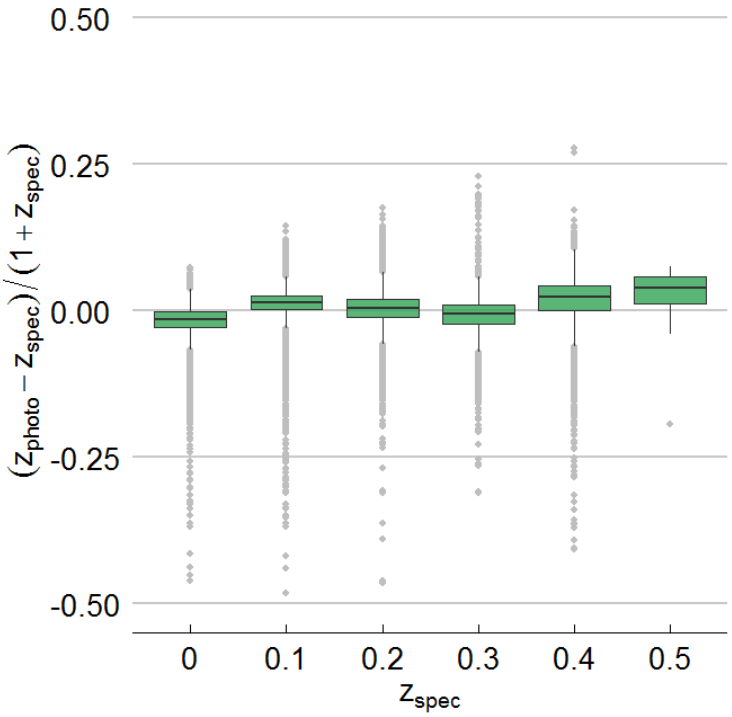


Figure 4.3: GLM - Color - Test $z \leq 0.5$

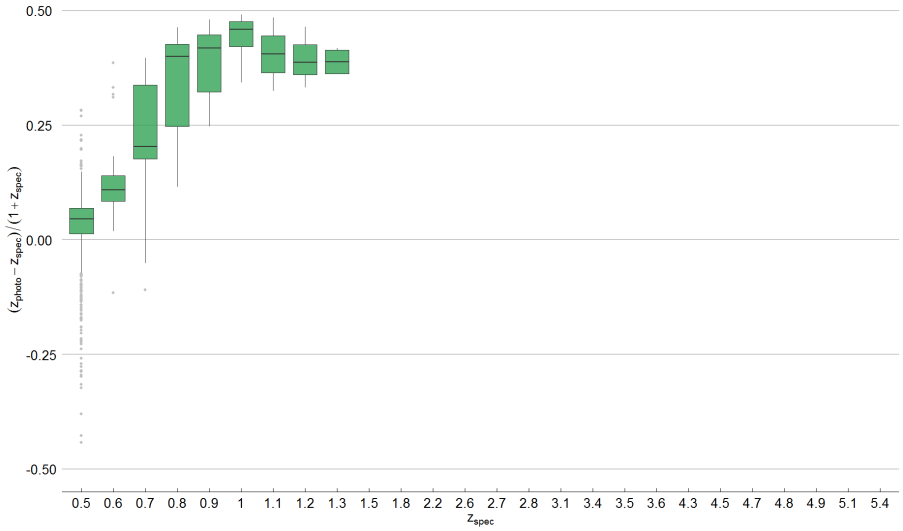


Figure 4.4: GLM - Color - Test $z > 0.5$

Table 4.1: CosmoPhotoz GLM Method - (Color data only) Photo-z Error Analysis

Range	**Count**	**Mean z-Spec**	**Mean z-Photo**	**Mean Error**	**RMS**	**Catastrophic Error %**
(0-0.1]	100485	0.064	0.080	0.021	0.048	0.29
(0.1-0.2]	93109	0.140	0.124	0.022	0.046	0.34
(0.2-0.3]	14982	0.239	0.232	0.028	0.043	2.80
(0.3-0.4]	12237	0.348	0.358	0.029	0.042	2.77
(0.4-0.5]	5585	0.441	0.415	0.052	0.072	8.31

performance in both the resulting ranges, that is, "(0.3-0.4]" and "(0.4-0.5]" as shown in Table 4.2. The improvement in catastrophic error rate for "(0.3-0.4]" reduced significantly from 2.77 to 0.37. The improvement for "(0.4-0.5]" reduced even more significantly from 8.31 to 0.27.

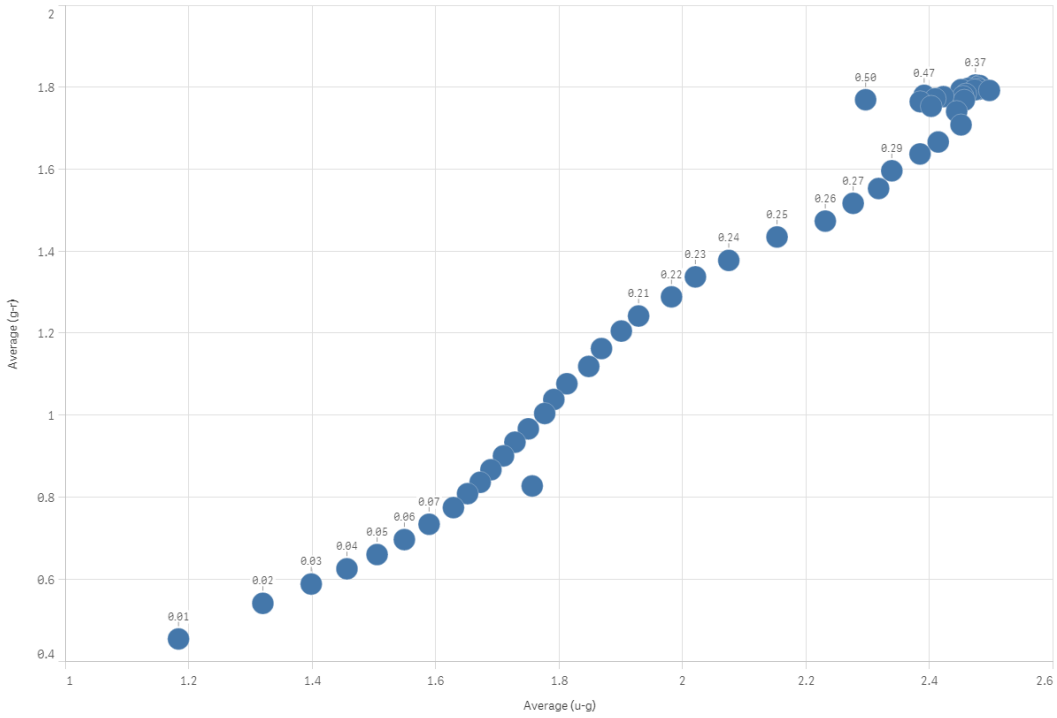


Figure 4.5: Mean redshift for $u-g$ vs. $g-r$, $z \leq 0.5$

Table 4.2: CosmoPhotoz GLM Method - (Color data only for for $z > 0.33$) Photo-z Error Analysis

Range	**Count**	**Mean z-Spec**	**Mean z-Photo**	**Mean Error**	**RMS**	**Catastrophic Error %**
(0.3-0.4]	9800	0.357	0.368	0.021	0.027	0.37
(0.4-0.5]	5579	0.442	0.423	0.024	0.03	0.27

4.3 Photometric Attributes

This study was performed with three as the number of principal components. This has noticeably better performance compared to when only color attributes are used to formulate the GLM model in all the ranges. Figure 4.6 shows how the spread of the estimates based on color and PSF magnitudes is much narrower across the different ranges compared against a similar graph for Color as shown in Figure 4.3. Refer Figure 4.7 and Figure 4.8 for color and Fiber magnitude, and color and Petrosian respectively. The catastrophic error rate in the "(0.4-0.5]" range is significantly better and this makes this model's estimates in all the ranges in the $1 - 5\%$ range of other current methods. Thus, our findings match the widely studied and reported understanding that color and magnitude together perform as better predictors of photo-z in contrast to using color alone. The color attributes were studied with PSF magnitudes in one run and Fiber Magnitudes in another run. Both the runs performance was very similar in performance and as shown in Table 4.3 and Table table:glm-colorfiber-caterr respectively. The Petrosian magnitudes had similar performance for the objects that could be estimated in it's model. However, a large number of objects could not be estimated due to divide-by-zero errors in their estimation.

When all the color (four) and magnitude (five each of PSF, Fiber and Petrosian corresponding to the ugriz band; total fifteen) attributes were combined together to generate GLM models, the catastropic error increased when three principal components were used. As the number of principal components were increased, the number of objects with the

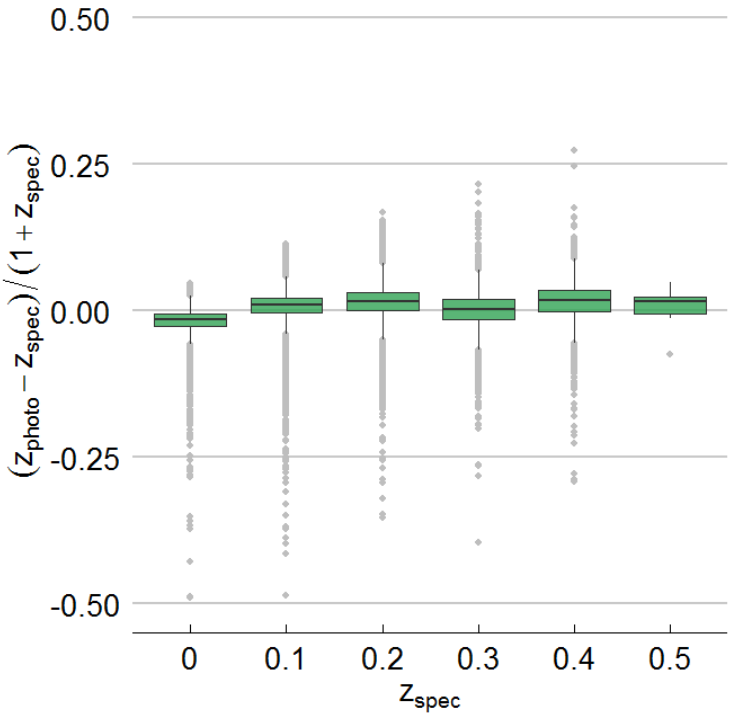


Figure 4.6: GLM - Color and PSF Magnitude - Test $z \leq 0.5$

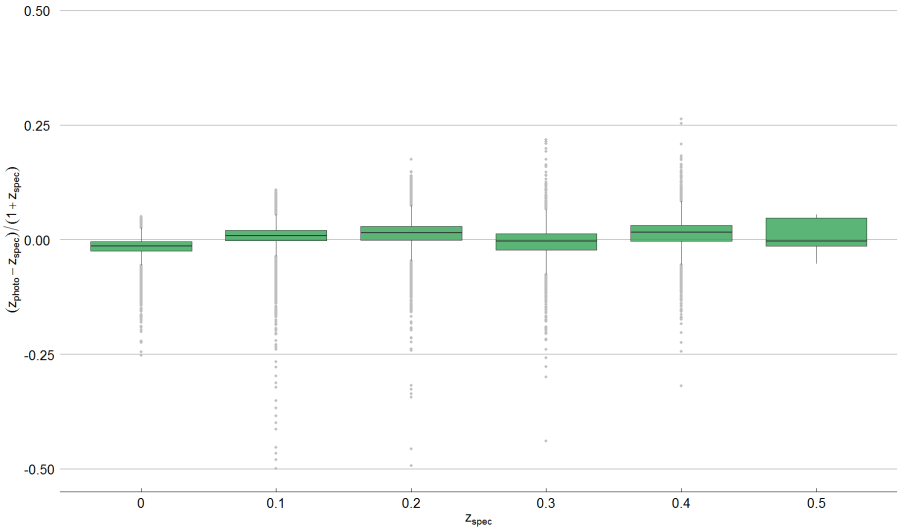


Figure 4.7: GLM - Color and Fiber Magnitude - Test $z \leq 0.5$

Table 4.3: CosmoPhotoz GLM Method - (Color and PSF Magnitude data only) Photo-z Error Analysis

Range	**Count**	**Mean z-Spec**	**Mean z-Photo**	**Mean Error**	**RMS**	**Catastrophic Error %**
(0-0.1]	71975	0.076	0.094	0.021	0.031	0.71
(0.1-0.2]	92920	0.14	0.13	0.022	0.041	0.48
(0.2-0.3]	15069	0.239	0.222	0.036	0.052	3.93
(0.3-0.4]	12067	0.348	0.347	0.03	0.042	2.31
(0.4-0.5]	5669	0.441	0.42	0.041	0.054	4.29

Table 4.4: CosmoPhotoz GLM Method - (Color and Fiber Magnitude data only) Photo-z Error Analysis

Range	**Count**	**Mean z-Spec**	**Mean z-Photo**	**Mean Error**	**RMS**	**Catastrophic Error %**
(0-0.1]	71597	0.076	0.093	0.02	0.041	0.43
(0.1-0.2]	93363	0.14	0.13	0.02	0.036	0.32
(0.2-0.3]	15217	0.239	0.223	0.033	0.047	3.2
(0.3-0.4]	12039	0.348	0.357	0.032	0.078	3.1
(0.4-0.5]	5498	0.441	0.421	0.039	0.052	4.04

catastrophic error reduced as more of the variance was accounted for in the additional principal components. The model generated with four to six principal components perform better than color and color-single type of magnitude in all the five ranges. Note when four principal components are used, the catastrophic error percent reduces significantly in the '(0.2-0.3]', '(0.3-0.4]' and '(0.4-0.5]' ranges. The spread of the estimation in the different ranges is narrower compared to any of the above attribute combinations. Refer Figure 4.9 for the estimates when six principal components were used. The error summary for these nineteen attributes using number of principal components = (3,4,5,6) are shown in Table 4.5, 4.6, 4.7 and 4.8 respectively.

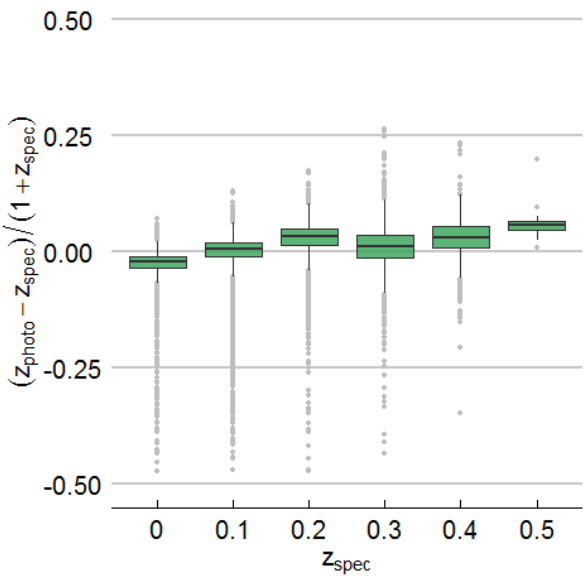


Figure 4.8: GLM - Color and Petrosian Magnitude - Test $z \leq 0.5$

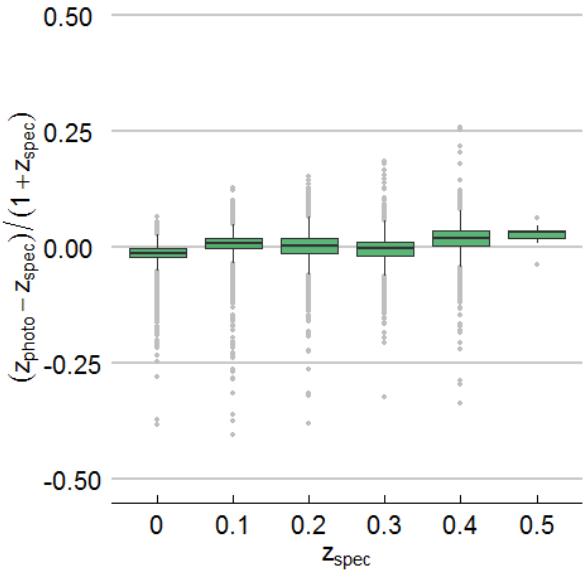


Figure 4.9: GLM - Color and PSF/Fiber/Petrosian Magnitude - Test $z \leq 0.5$

Table 4.5: CosmoPhotoz GLM Method - (Color and PSF/Fiber/Petrosian Magnitude data only) Photo-z Error Analysis - Number of Principal Components = 3

Range	**Count**	**Mean z-Spec**	**Mean z-Photo**	**Mean Error**	**RMS**	**Catastrophic Error %**
(0-0.1]	72887	0.076	0.097	0.024	0.041	0.6
(0.1-0.2]	92956	0.14	0.13	0.023	0.032	0.32
(0.2-0.3]	15126	0.239	0.214	0.042	0.053	6.43
(0.3-0.4]	12293	0.348	0.345	0.038	0.048	4.03
(0.4-0.5]	5503	0.441	0.416	0.049	0.063	10.52

Table 4.6: CosmoPhotoz GLM Method - (Color and PSF/Fiber/Petrosian Magnitude data only) Photo-z Error Analysis - Number of Principal Components = 4

Range	**Count**	**Mean z-Spec**	**Mean z-Photo**	**Mean Error**	**RMS**	**Catastrophic Error %**
(0-0.1]	73058	0.076	0.092	0.019	0.026	0.39
(0.1-0.2]	93165	0.14	0.131	0.02	0.027	0.27
(0.2-0.3]	15087	0.239	0.224	0.031	0.041	2.49
(0.3-0.4]	11903	0.348	0.353	0.025	0.035	1.5
(0.4-0.5]	5536	0.441	0.417	0.035	0.044	1.79

Table 4.7: CosmoPhotoz GLM Method - (Color and PSF/Fiber/Petrosian Magnitude data only) Photo-z Error Analysis - Number of Principal Components = 5

Range	**Count**	**Mean z-Spec**	**Mean z-Photo**	**Mean Error**	**RMS**	**Catastrophic Error %**
(0-0.1]	72810	0.076	0.09	0.018	0.024	0.17
(0.1-0.2]	93577	0.139	0.13	0.018	0.025	0.18
(0.2-0.3]	14980	0.239	0.235	0.029	0.11	2.41
(0.3-0.4]	12002	0.348	0.353	0.026	0.037	1.74
(0.4-0.5]	5458	0.441	0.417	0.041	0.055	3.37

Table 4.8: CosmoPhotoz GLM Method - (Color and PSF/Fiber/Petrosian Magnitude data only) Photo-z Error Analysis - Number of Principal Components = 6

Range	**Count**	**Mean z-Spec**	**Mean z-Photo**	**Mean Error**	**RMS**	**Catastrophic Error %**
(0-0.1]	71522	0.076	0.09	0.017	0.023	0.14
(0.1-0.2]	93403	0.139	0.13	0.017	0.023	0.12
(0.2-0.3]	15066	0.239	0.234	0.027	0.038	2.12
(0.3-0.4]	12047	0.348	0.352	0.025	0.035	1.31
(0.4-0.5]	5624	0.441	0.416	0.039	0.074	2.45

4.4 Photomorphic Attributes

The morphology attributes investigated are ratio of Fiber to Petrosian magnitude (fp_*), Petrosian radius and concentration index (five each corresponding to the ugriz band) combined with color (four). They are investigated one at a time as well as the all the attributes together.

When fp_* are combined with color, the generated GLM model using three principal components performs better compared to using color alone in all ranges, except $z = 0.4-0.5$. fp_* could not be used alone as the predictor since it resulted in numerous divide-by-zero errors while developing the model. When compared against color and magnitude as well, it similarly performed better in all ranges, except $z = 0.4 - 0.5$. Refer Table 4.9 for fp_* with three principal components error summary to compare against only-color-based error summary in Table 4.1 and color-magnitude error summary in Table 4.3 (PSF), 4.4 (Fiber) and 4.5 (all magnitudes with three principal components). Refer Figure 4.10 for the estimates when three principal components are used.

The peformance for the range $z = 0.4 - 0.5$ was improved when four principal components were used to generate the GLM model instead of three; error percent went down from 13% to 5.4%. The catastrophic error percent reduced in all the other ranges as well. The error summary for $z = 0 - 0.3$ was better than color and all magnitudes, even when color-magnitude data was modeled using four or five principal components. This hints that morphology likely covers the variance of the data more compactly than the magnitudes. Refer Table 4.10 for fp_* with four principal components error summary to compare against color-magnitude error summary in Table 4.6 (all magnitudes with four principal components). Five and six principal components were not studied since color and fp_* attributes are only nine attributes and it did not make sense to study more than four principal components. Refer Figure 4.11 for the estimates when three principal components are used.

The second morphology attribute Petrosian radius was not found to be a good predictor when used with or without color. The error summary shows that for $z = 0-0.2$ where there

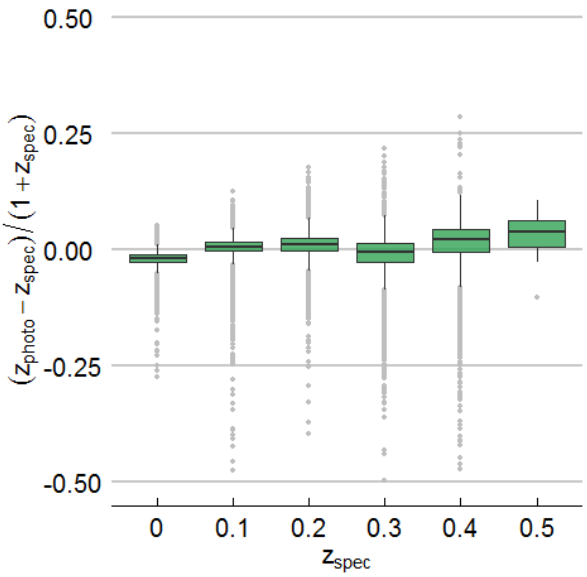


Figure 4.10: GLM - Color and Ratio of Fiber to Petrosian Magnitude - Test $z \leq 0.5$ - Number of Principal Components = 3

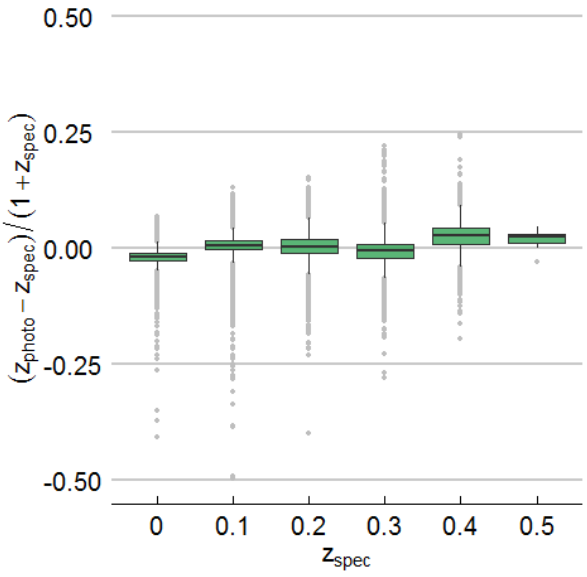


Figure 4.11: GLM -Color and Ratio of Fiber to Petrosian Magnitude - Test $z \leq 0.5$ - Number of Principal Components = 4

Table 4.9: CosmoPhotoz GLM Method - (Color and Ratio of Fiber to Petrosian Magnitude data only) Photo-z Error Analysis - Number of Principal Components = 3

Range	**Count**	**Mean z-Spec**	**Mean z-Photo**	**Mean Error**	**RMS**	**Catastrophic Error %**
(0-0.1]	46562	0.084	0.131	0.048	5.625	0.15
(0.1-0.2]	92980	0.14	0.131	0.016	0.026	0.15
(0.2-0.3]	15210	0.239	0.226	0.029	0.042	2.62
(0.3-0.4]	11917	0.348	0.362	0.037	0.06	5.24
(0.4-0.5]	5579	0.441	0.424	0.059	0.089	13.05

Table 4.10: CosmoPhotoz GLM Method - (Color and Ratio of Fiber to Petrosian Magnitude data only) Photo-z Error Analysis - Number of Principal Components = 4

Range	**Count**	**Mean z-Spec**	**Mean z-Photo**	**Mean Error**	**RMS**	**Catastrophic Error %**
(0-0.1]	46398	0.084	0.104	0.021	0.025	0.12
(0.1-0.2]	93314	0.139	0.131	0.016	0.023	0.16
(0.2-0.3]	15040	0.239	0.234	0.027	0.042	2.14
(0.3-0.4]	12086	0.348	0.358	0.028	0.039	2.31
(0.4-0.5]	5501	0.441	0.405	0.046	0.056	5.38

is large amount of data, it shows good performance but not as good as the prior-discussed predictors (slightly better than color and PSF magnitude in this range, but significantly worse in the other ranges). It needs to be investigated as to why the performance deteriorates for $z = 0.3 - 0.5$ - maybe it is likely due to bad data (incorrect measurements entirely or certain subsets of data). Refer Table 4.11 and Table 4.12 for error summary when three and four principal components respectively were used to generated the GLM model using color and Petrosian radius data. Refer Figure 4.12 and 4.13 for the spread of all the estimates in a given range when three or four principal components respectively are used to generate the GLM model. As z increases and the number of data points associated with that range reduces, the spread of the estimate increases.

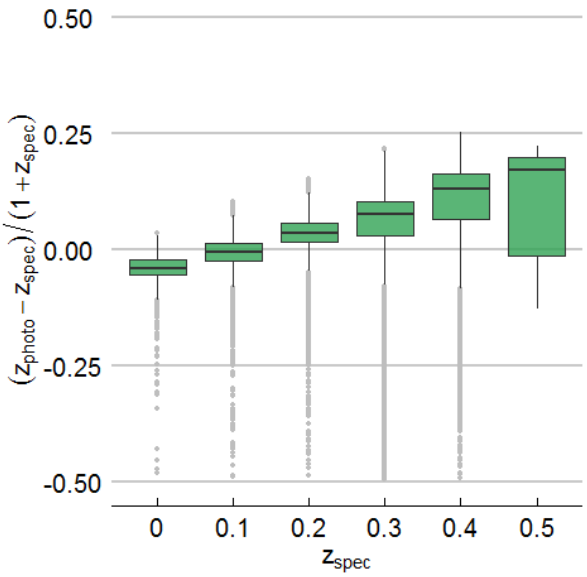


Figure 4.12: GLM - Color and Petrosian Radius - Test $z \leq 0.5$ - Number of Principal Components = 3

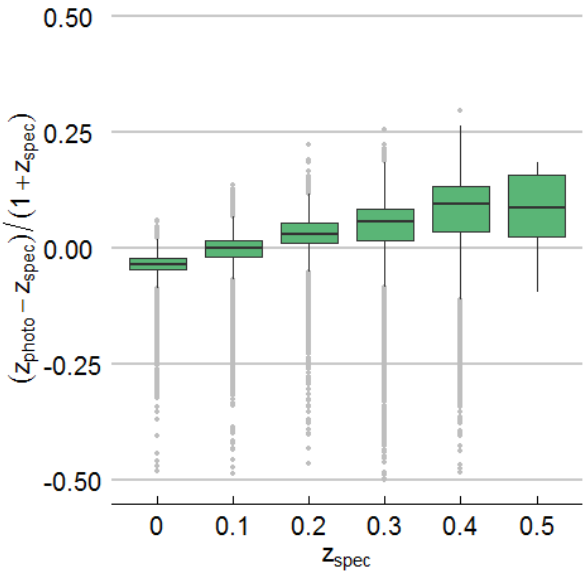


Figure 4.13: GLM - Color and Petrosian Radius - Test $z \leq 0.5$ - Number of Principal Components = 4

Table 4.11: CosmoPhotoz GLM Method - (Color and Petrosian Radius data only) Photo-z Error Analysis - Number of Principal Components = 3

Range	**Count**	**Mean z-Spec**	**Mean z-Photo**	**Mean Error**	**RMS**	**Catastrophic Error %**
(0-0.1]	46418	0.084	0.126	0.042	0.059	0.51
(0.1-0.2]	93219	0.139	0.143	0.026	0.037	0.35
(0.2-0.3]	14983	0.239	0.196	0.055	0.073	11.51
(0.3-0.4]	11997	0.348	0.287	0.125	0.159	60.57
(0.4-0.5]	5635	0.442	0.309	0.193	0.218	83.05

Table 4.12: CosmoPhotoz GLM Method - (Color and Petrosian Radius data only) Photo-z Error Analysis - Number of Principal Components = 4

Range	**Count**	**Mean z-Spec**	**Mean z-Photo**	**Mean Error**	**RMS**	**Catastrophic Error %**
(0-0.1]	46499	0.084	0.175	0.092	11.223	1.13
(0.1-0.2]	93107	0.14	0.14	0.025	0.188	0.66
(0.2-0.3]	15047	0.239	0.205	0.053	0.081	10.06
(0.3-0.4]	12026	0.348	0.302	0.103	0.217	42.11
(0.4-0.5]	5562	0.441	0.34	0.154	0.185	71.72

Concentration index is the third and last morphology-based attribute under consideration. It could not be used to estimate for $z < 0.1$. It could not be used alone as the predictor as well since it resulted in numerous divide-by-zero errors while developing the model. It has high catastrophic error rate in almost all the ranges and does not stand out as a good estimator by itself as shown in Table 4.13. It has approximately 20-25% catastrophic error rate for the range $z = 0.4 - 0.5$. The spread of the estimates increases for higher ranges as shown in Figure 4.15. The performance deteriorates further with using four principal components instead of three.

Table 4.13: CosmoPhotoz GLM Method - (Color and Concentration Index data only) Photo-z Error Analysis - Number of Principal Components = 3

Range	**Count**	**Mean z-Spec**	**Mean z-Photo**	**Mean Error**	**RMS**	**Catastrophic Error %**
(0.1-0.2]	30525	0.171	0.188	0.022	0.029	0.31
(0.2-0.3]	15005	0.239	0.232	0.025	0.043	1.57
(0.3-0.4]	12248	0.348	0.335	0.044	0.278	4.34
(0.4-0.5]	5551	0.442	0.41	0.068	0.088	21.06

4.5 Highest Correlation Attribute Subset

The Bayesian analysis in Section 5 identifies the following six attributes as highly correlated - Ratio of Fiber to Petrosian in g-band and r-band, Petrosian magnitude in u-band and g-band, and Fiber magnitude in u-band and g-band. Formulating the GLM model based on these six attributes show low catastrophic error percent in all the ranges, comparable in performance to the color and magnitude results. The model based on these six attributes outperformed in the range z=0.1-0.2 and 0.3-0.5 even when two principal components were used. Refer Table 4.14 and 4.15 for the performance in the different ranges for two and three principal components respectively. There was not any significant improvement on using three principal components but it would account for more variance in the data. The spread of the estimates in the different ranges can be set in Figure 4.15. The spectroscopic redshift vs. photometric redshift can be seen at Figure 4.16. This is the closest match compared to the rest of the attribute sets considered in this analysis without any distant outlier.

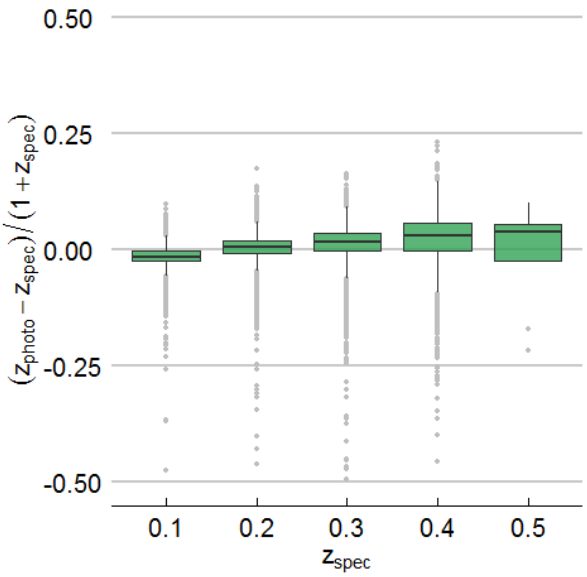


Figure 4.14: GLM - Color and Concentration Index - Test $z \leq 0.5$ - Number of Principal Components = 3

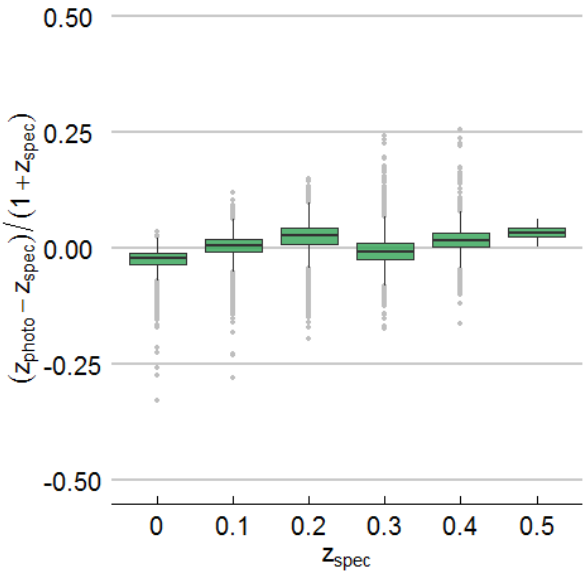


Figure 4.15: GLM - Six highest correlated attributes - Test $z \leq 0.5$ - Number of Principal Components = 3

Table 4.14: CosmoPhotoz GLM Method - (Six highest correlated attributes data only) Photo-z Error Analysis - Number of Principal Components = 2

Range	Count	Mean z-Spec	Mean z-Photo	Mean Error	RMS	Catastrophic Error %
(0-0.1]	46471	0.084	0.11	0.026	0.032	0.32
(0.1-0.2]	93448	0.14	0.132	0.021	0.027	0.12
(0.2-0.3]	15237	0.239	0.211	0.04	0.049	3.37
(0.3-0.4]	12120	0.348	0.36	0.032	0.316	1.33
(0.4-0.5]	5675	0.441	0.41	0.037	0.048	2.03

Table 4.15: CosmoPhotoz GLM Method - (Six highest correlated attributes data only) Photo-z Error Analysis - Number of Principal Components = 3

Range	Count	Mean z-Spec	Mean z-Photo	Mean Error	RMS	Catastrophic Error %
(0-0.1]	46594	0.084	0.11	0.026	0.032	0.32
(0.1-0.2]	93192	0.139	0.132	0.021	0.026	0.12
(0.2-0.3]	15320	0.239	0.208	0.042	0.05	3.28
(0.3-0.4]	12140	0.348	0.356	0.032	0.041	1.75
(0.4-0.5]	5675	0.441	0.414	0.036	0.046	2.36

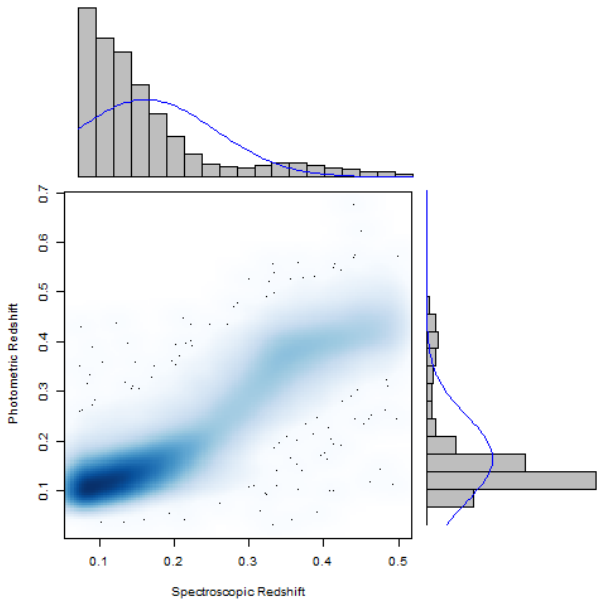


Figure 4.16: GLM - Six highest correlated attributes - Spec-z vs. Photo-z - Number of Principal Components = 3

Chapter 5: Bayesian Photomorphic Redshift

5.1 Bayesian Statistics

Bayes' Law is a simple method for updating beliefs in the light of new evidence. Suppose there is some statement A that you initially believe has a probability P(A) of being correct (what Bayesians call the "prior" probability). If a new piece of evidence, B, comes along, then the probability that A is true given that B has happened (what Bayesians call the "posterior" probability) is given by

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \Rightarrow Posterior \propto \text{Likelihood of observed data} \times \text{Prior} \qquad (5.1)$$

where P(B—A) is the likelihood that B would occur if A is true, and P (B) is the likelihood that B would occur under any circumstances. Note: This is in contrast to the "frequentist" approach, which views probability not as a degree of belief but as the relative frequency of events that can be repeated many times and it is the dominant statistical paradigm.

Bayesian statistics was invented in the 18th century by an English Presbyterian minister named Reverend Thomas Bayes, whose manuscript on the subject was published posthumously in 1763 by some accounts to calculate the probability of Gods existence. For decades, however, Bayesian analysis was too computationally intensive to carry out in many cases. The approach typically involves calculating high-dimensional integrals, whereas frequentist approaches more often involve optimization, which is easier from a computational standpoint. By the 1980s and 1990s, however, increases in computing power, combined

with the development of Markov chain Monte Carlo methods for calculating numerical approximations to high-dimensional integrals "liberated Bayesian inference, and made it much more prominent. Alan Turing's use of Bayesian statistics to help crack the Enigma encryption machines used by Germany in World War II. [131] It is being used in a variety of fields - from physics to cancer research, ecology to psychology.

Judea Pearl [132] originally coined the term Bayesian Networks, also known as Bayesian Belief Network (BBN), in the late 1970s. Belief networks play a central role in two uncertainty formalisms: probability theory where they are called Bayesian networks, causal nets or influence diagrams, and the Dempster-Shafer theory where they are rerferred as galleries, qualitative Markov networks or constraint networks. The four relationships that are seen as basic primitives of probability theory are: likelihood [1], conditioning [2], relevance [3] and causation [4]. Probabilities are summaries of knowledge that is left behind when information is transferred to a higher level of abstraction. Bayesian methods provide a formalism for reasoning about partial beliefs under conditions of uncertainty.[133]

Bayesian Networks (BN) are directed probabilistic graph models that are used to model variable dependency relationships. In other words, BNs are used to represent and approximate joint distributions over sets of variables. The inter-connections structure of the graphical model represents the dependencies among the set of variables. The goal of learning a graphical model is to learn both the graphical structure (qualitative knowledge) and the parameters of the approximate joint distribution (quantitative knowledge) from data. The graph structure of the same is a directed acyclic graph (DAG). Each variable is independent of its nondescendants in the network given its parents. When converting a directed graph to an undirected graph, we must add links between "unmarried" parents who share a common child (i.e., "moralize" the graph) to prevent us reading off incorrect independence statements. In a general form of the graph, the nodes can represent not only random variables

[1] "Tim is *more likely* to run than to walk"
[2] "*If* Tim is sick, he can't run"
[3] "Whether Tim runs *depends on* whether he is sick"
[4] Being sick *caused* Tim's inability to run"

but also hypothesis, beliefs and latent variables. Practitioners often follow frequentists' method to estimate the parameters of the network. Bayesian networks are well suited to dealing with incomplete data and capturing the prior knowledge of the domain. It is robust to model overfitting since the data is combined probabilistically with prior knowledge in the model. There exists over fifty learning algorithms for BN. [134] [135] [136] [137] [138] [139] [140] [141] [142] [126] [143] [144] [145] [146] [147]

This is a technique used by scientists to turn data into knowledge, evidence and predictions to aid domain modeling, risk analysis and decision support. It only depends on data to build the model without any assumption about the functional form. Bayesian calculations should be used not necessarily to replace classical frequentist statistics but to flag spurious results. One downside of Bayesian statistics is that it requires prior information and often scientists need to start with a guess or estimate. Things get murkier when statisticians use Bayes' rule to try to reason about one-time events, or other situations in which there is no clear consensus about what the prior probabilities are. If this initial belief is way off, we are likely to get bad inferences. It is however, a statistical approach to combine prior beliefs and experiences with new evidence.

In the early morning of June 1, 2009, Air France flight AF 447, carrying 228 passengers and crew, disappeared over a remote section of the Atlantic Ocean. All attempts to find the airplane was futile, including sonar searches. Bayesian inference approach started by constructing a probability map based on the initial data about the flight's disappearance, then used Bayes' Law to incorporate the evidence provided by the failures of the various search attempts. The wreckage was found within a week. Most statistical techniques cannot handle data that comes in different flavorssurface and underwater searches with different types of equipment, information about the plane's flight path, the drift model, and so forthbut Bayesian inference allows statisticians to easily combine many different types of measurements and data. Each measurement simply gets transformed into a likelihood function on the space of all possible locations for the airplane, representing the likelihood of obtaining that particular measurement if the airplane is in that particular spot. Bayes' Law

then uses this likelihood function to update the prior, resulting in the posterior distribution. Bayesian analyses recovered the lost U.S. nuclear submarine Scorpion and the wreck of the SS Central America, a steamship that sank off the Atlantic coast in 1857. [131]

The Coast Guard program Sarops (Search and Rescue Optimal Planning System) uses Bayesian statistics and it was used succesfully to find a missing person starting with the most sparse information about the time-frame in which the person went missing in 2013. Searchers added new information on prevailing currents, places the search helicopters had already flown and some additional clues found by the boats captain. The system could not deduce exactly where Mr. Aldridge was drifting, but with more information, it continued to narrow down the most promising places to search. [148]

The Bayesian approach was catapulted into the public eye when Nate Silver, on his FiveThirty Eight blog, used it to predict correctly the poll outcome of every state in the 2012 U.S. Presidential election. It is being used in a wide range of applications, including finding distant quasars, estimating HIV prevalence in different regions, and explaining the phenomenon that richer people tend to vote Republican while richer states tend to vote Democrat. Some other uses of Bayesian statistics and graphical network modeling include the following:

- Embedded in Microsoft Office products, including the Answer Wizard of Office 95, the Office Assistant of Office 97, and the Technical Support Troubleshooter applications. [141]

- The Vista system is a decision-theoretic system that has been used at NASA Mission Control Center in Houston. The system uses Bayesian networks to interpret live telemetry and provides advice on the likelihood of alternative failures of the space shuttle's propulsion systems. It also considers time criticality and recommends actions of the highest expected utility. [141]

- [149] [150] presents a Bayesian methodology "the most natural and unambiguous approach towards the aggregation problem while addressing uncertainty in the expert

judgment at the same time" for assessing relative accident probabilities and their uncertainty using paired comparison to elicit expert judgments. The approach is illustrated for a risk study of the Washington State Ferry, the largest passenger ferry system in the U.S. Epistemic uncertainty, i.e. uncertainty due to lack of knowledge of the system, results from uncertainties in input data to simulation models and truncating estimates affect the results (output) of the simulation models. Bayesian simulation analysis allows treatment of these uncertainties. [151] is a similar study for the San Francisco Bay ferries.

- Accelerated life testing (ALT) is the set of procedures used to reduce the time needed to obtain information related to life characteristics of an item, material or part of interest. [152] compares different ALT designs (fixed stress, profile ALT, progressive step-stress ALT and regressive ALT) within a single Bayesian inference framework.

- The continuous-time Bayesian networks (CTBNs) have been used to model social networks, cardiogenic heart failure, and stroke rehabilitation. A CTBN is a probabilistic graphical model in which nodes are discrete random variables, where the state evolves continuously over time. The probability law that governs the state transitions depends on the state of the node parents in the graph. [153]

- Determination of crystal structures using X-ray/neutron powder diffraction is an inverse problem of finding a disposition of atoms by fitting an experimental diffraction pattern with a model signal. The accuracy of the obtained structural parameters is often limited by systematic errors that affect intensities and shapes of diffraction peaks. A probabilistic method that accounts for systematic errors using Bayesian statistics and marginalization of error corrections was developed, without assuming any particular model for these errors. [154]

- [155] uses Bayesian networks (BNs) for discovering relations between genes, environment, and disease for a population-based study of bladder cancer in New Hampshire, USA. The R package *bnlearn* was used for the study.

- In the context of image segmentation, Bayesian inference is a tool for determining the likelihood that a particular object x is present in a scene, given that sensor data y (i.e., image data) is observed. In Solar System studies, Bayesian methods are used for automatically identifying various surface structures on the Sun - plage [5], network [6] and background components [7]. The boundaries produced by the Bayesian approach are smooth but were found to be very time consuming. [156] This indicates a scope for additional work to optimize the application of Bayesian networks and improve the overall performance.

- Global solar irradiation is considered as the most significant parameter in meteorology, solar conversion, and renewable energy applications, particularly for modeling the sizing and modeling of photovoltaic (PV) systems. Bayesian Neural Network (NN) approach performed better than the other examined models (NN and empirical models) for prediction of daily solar irradiation. [157] The dataset consists of daily solar irradiation, sunshine duration, air temperature and relative humidity from an meteorological database with measurements from 1998 to 2002 at Al-Madinah (Saudi Arabia). Other similar studies related to usage of BN for this issue are [158]

- The Bayesian network approach has been used for short-term solar flare level prediction in [159]. The performances of the two BN models generated by this study appear comparable with other methods. They do note that the comprehensibility of the Bayesian network models is better than the other methods.

- Earth systems models use mathematical descriptions as modeled processes of the environmental phenomena. They are sophisticated and involve multiple approximations that require parameters whose values are not derived as field measurements. Determining appropriate parameter values and estimating the related forecast uncertainties

[5] high-intensity, clustered regions that spatially coincide with active regions
[6] a lower-intensity cellular structure formed by the hot boundaries of convection cells
[7] lowest-intensity region formed by the cooler cell interiors

is a challenging task. These models provide reasonable descriptions of the current climate system, however, it is unclear how accurately they will respond to changes in external forcing. Model calibration has often been viewed not singularly as a question of optimization, but as the probabilistic description of a range of parameter sets (and therefore model forecasts) in which the modeler has confidence which lends well to a Bayesian formulation of model calibration. [160]

- [161] studies the potentials and limits of Bayesian Networks in dealing with uncertainty characterizing the definition and implementation of climate change adaptation policies. Applications of BN in earlier studies to ecological modeling, natural resource management, and climate change policy issues are reviewed.

- [162] utilizes a Bayesian Belief Network (BBN) approach to quantify the understanding of the complex physical, chemical, and biological processes that lead to eutrophication "an increase in the rate of supply of organic matter to an ecosystem" in an estuarine ecosystem (New River Estuary, North Carolina, USA). There were two main challenges - the discretization procedure and feedback relationships. Points to consider when discretizing are: the size of the available dataset, the interpretation goal of the node, the placement of the node with in the BBN (does it have any predecessor nodes?), the shape of the underlying distribution, the number of outliers, and the number of repetitive values for data points. The two established discretization techniques for empirical datasets are equal-interval and equal-frequency. The equal-interval method is unsuitable when the dataset is unevenly distributed or contains outliers, since it would result in sparsely populated bins. The equal-frequency method has shortcomings when dataset has repetitive values. Neither of these techniques preserve the original distribution of the data; hence, this paper discretized the BBN nodes by exploring a new approach called moment matching method, which focuses on matching lower statistical moments of the initial distribution (i.e. mean, variance, skewness,

kurtosis, etc.). This leads to a better representation of the underlying continuous distribution. Another important point to address while discretizing continuous variables is the number of intervals. Large number of intervals would improve representation of the underlying distribution but increase the size of the conditional probability tables due to increase in states of predecessor nodes; hence, an optimal number of intervals for each variable should be determined. Feedback relationships challenge was expected to be addressed in future work using Dynamic Object Oriented Bayesian Networks (OOBN) with each OOBN representing a time step.

- [163] combines information from divergent sources of data to classify the risk of desertification after a forest fire. Data consisted of satellite sensor images, topographic maps, geological maps etc, each one with its own resolution and accuracy, from the burned forests in the Mediterranean region. The effort is to incorporate the uncertainty in the input data in the network and present various methods by which the conditional probability matrices used by the network can be constructed.

Bayesian network does not contain any causal assumptions, i.e. no knowledge of the causal order between the variables, so the interpretation should be merely statistical (informational). Causal networks are Bayesian networks in which the correct probability model after intervening to fix any node's value is given simply by deleting links from the node's parents. This can be used for predictions based on various courses of action.

Bayesian inference is most useful in domains where experts can provide good models from which to construct the prior. For example, when a bridge is designed, we consider the structure and using the strength of materials available we calculate how many cars per hour might cross the bridge. We call this type of calculation a forward problem. Alternatively, we could start with the expected traffic flow, engineer in the material properties, and thus determine the design. We call this calculation an inverse problem. Solving inverse problems is often extremely difficult. The most important keys to extracting the maximum insight from a given data set are to sample the most appropriate solution parametersthose for

which prior information is most abundantand to carefully construct the prior to make it as informative as possible. Constructing efficient and informative priors is a very creative endeavor.[164]

[165] empirically evaluates algorithms for learning four types of Bayesian network (BN) classifiers - Naïve-Bayes, tree augmented Naïve-Bayes (TAN), BN augmented Naïve-Bayes (BAN) and general BNs, where the latter two are learned using two variants of a conditional-independence (CI) based BN-learning algorithm. A Naïve-Bayes BN is a simple structure that has the classification node as the parent node of all other nodes. No other connections are allowed in a Naïve Bayes structure since it assumes all the features are independent of each other. TAN partitions the attributes into disjoint groups and assumes independence only between attributes of different groups. Thus, this results in tree-like structures. BAN classifiers extend TAN classifiers by allowing the attributes to form an arbitrary graph. CI-based algorithms are competitive with (or superior to) the best known classifiers.

Belief networks are popular tools for encoding uncertainty in expert systems. These networks rely on inference algorithms to compute beliefs in the context of observed evidence. Additionally, belief networks are used by experts to encode selected aspects of their knowledge and beliefs about a domain. Once constructed, the network induces a probability distribution over its variables. One established method for exact inference on belief networks is the probability propagation in trees of clusters (PPTC) algorithm. [166] PPTC converts the belief network into a secondary structure, then computes probabilities by manipulating the secondary structure.

Some of the applications of Bayesian statistics in astronomy domain include the following among others:

- The New York University astrophysicist David Hogg credits Bayesian statistics with narrowing down the age of the universe. As recently as the late 1990s, astronomers could say only that it was eight billion to 15 billion years; now, factoring in supernova explosions, the distribution of galaxies and patterns seen in radiation left over from the Big Bang, they have concluded with some confidence that the number is 13.8

billion years. [148]

- Bayesian reasoning combined with advanced computing power has also revolutionized the search for planets orbiting distant stars, said Dr. Turner, the Princeton astrophysicist. In most cases, astronomers cant see these planets; their light is drowned out by the much brighter stars they orbit. What the scientists can see are slight variations in starlight; from these glimmers, they can judge whether planets are passing in front of a star or causing it to wobble from their gravitational tug. [148]

- [167] states "Our approach is probabilistic in the sense that we do not expect to succeed with the classification of every specific object, but to correctly classify most objects, while minimizing biases in the output sample, this by assigning each object its most likely type. The flexibility of the Bayesian framework allows us to do so by analyzing SNe from different surveys, with different depths and types of redshift information, to incorporate fully all the available information on each object, and to propagate correctly the unknowns."

- [168] Approximate Bayesian Computation (ABC) represents a powerful methodology for the analysis of complex stochastic systems for which the likelihood of the observed data under an arbitrary set of input parameters may be entirely intractable the latter condition rendering useless the standard machinery of tractable likelihood-based, Bayesian statistical inference [e.g. conventional Markov chain Monte Carlo (MCMC) simulation]. ABC is applied to a case study in the morphological transformation of high-redshift galaxies. First, a stochastic model for the competing processes of merging and secular evolution in the early Universe is developed. Secondly, through an ABC-based comparison against the observed demographics of massive galaxies at $1.5 < z < 3$ in the CANDELS/EGS [8] data set, posterior probability densities for the key parameters of this model is derived. Another astronomical problem readily amenable to ABC is that of inferring the age and mass of an unresolved star cluster

[8] Cosmic Assembly Near-IR Deep Extragalatic Legacy Survey (CANDELS)/Extended Groth Strip (EGS)

based on its broad-band SED.

- [169] presents a new method for inferring the metallicity (Z) and ionization parameter (q) of H II regions and star-forming galaxies using strong nebular emission lines (SELs). The Bayesian inference method derives the joint and marginalized posterior probability density functions for Z and q given a set of observed line fluxes and an input photoionization model.

5.2 Bayesian Statistics in Redshift Estimation

Bayesian networks are graphical representation of probabilistic relationships among a set of variables. In general, there are three main approaches for learning Bayesian networks from data - search-and-score, constraint-based and hybrid. The search-and-score approach attempts to identify the network that maximizes a score function indicating how well the network fits the data. One such score metric is the a posteriori probability of a network N given the data D and prior knowledge K, i.e., $argmax_N P(N|D,K)$. Algorithms in this category search the space of all possible structures for the one that maximizes the score using greedy, local, or heuristic search techniques, such as hill-climbing or simulated annealing. The computation of the full likelihood over both the parameter space and structure space is impractical for all but the smallest networks, requiring approximations such as the Bayesian Information Criterion (BIC) to be used. The BIC score (also known as the Schwarz Information Criterion and equivalent to the Minimum Description Length), can be written as shown in (5.2), where $p(D|\hat{\theta},G)$ is the likelihood of the data D according to estimated parameters $\hat{\theta}$ and structure G, N is the sample size of the dataset, and n_p is the number of parameters. The second term serves to penalize networks with many edges, thus the BIC will lead to a preference for simpler graphs. For large N, the highest scoring model often has parameters that are close to the maximum likelihood values.

$$BIC = log(p(D|\hat{\theta},G)) - \frac{n_p}{2} log(N) \tag{5.2}$$

When a greater tolerance for complex networks is desired (e.g., in the exploratory phase of analysis), the Akaike information criterion (AIC) provides an alternative scoring function as shown in (5.3)

$$AIC = log(p(D|\hat{\theta}, G)) - n_p \tag{5.3}$$

The AIC penalizes less harshly for the inclusion of additional edges (and their associated parameters). It is important to note that the maximum likelihood itself cannot be used as a score function, as without the inclusion of a penalty term it would always lead to selection of a completely connected network.

The K2 score, which corresponds to the Bayesian posterior for the special case of a uniform prior on both the structure and parameters. The contribution of each variable to the logarithm of the K2 score can be written as shown in (5.4).

$$log(K2(X_i)) = \sum_{j=1}^{q_i} \left(ln\left(\frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!}\right) + \sum_{k=1}^{r_i} ln(N_{ijk}!) \right) \tag{5.4}$$

where N_{ijk} represents the number of cases in the database in which the variable X_i took its kth value ($k = 1, 2, \cdots, r_i$), and its set of parents was instantiated as its jth unique combination of values ($j = 1, 2, \cdots, q_i$), and $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. The logarithm of the total K2 score is then the sum of the individual contributions. The K2 score is typically intermediate to the AIC and BIC in its penalization of network complexity.

The second approach for learning Bayesian networks is constraint-based. Algorithms following this approach estimate from the data whether conditional independence between the variables hold. Typically, this estimation is performed using statistical or information theoretic measures. The conditional independence constraints are propagated throughout the graph and the networks that are inconsistent with them are eliminated from further consideration. A sound strategy for performing conditional independence tests ultimately retains (and returns) only the statistically equivalent networks consistent with the tests.

Constraint-based methods can be more efficient than score-based approaches, especially when the number of samples is large. The score-based approach is generally preferred, particularly when dealing with small sample size and noisy data. The third approach of hybrid algorithms combine the two conventional methods to maximize their advantages. Typically, they start with a constraint-based algorithm to find the skeleton of the network and then employ a score-based method to identify the best set of edge orientations.

The search-and-score algorithms for learning Bayesian networks from data have two components: a scoring metric and a search procedure. The scoring metric computes a score reflecting the goodness-of-fit of the structure to the data, for example, BDe metric [136]. The search procedure tries to identify network structures with high scores. The search problem of identifying a Bayesian network (where each node is allowed at most K parents) that has a relative posterior probability greater than a given constant is NP-complete. [170] In other words, if the user overestimates k, the algorithm will take unnecessarily long to finish and may even be rendered intractable for large datasets. Additionally, if the user underestimates k, there is a risk of discovering a suboptimal network. Unfortunately, the effects of underestimating k affect quality in a non-local fashion.

Learning the most probable a posteriori Bayesian network from data is an NP-Hard problem. The study in 1996 using the condition that each node has at most k parents, for all $k \geq 2$, determined it to a NP-Complete problem [170]. It was later determined in a 2004 study that this is a NP-Hard problem with the condition that $k \geq 3$ [171]. Additionally, the problem of finding the optimal path graphical models is NP-Hard as well (applies for both directed path and undirected path). [172]

Problems are intractable when they "can be solved, but not fast enough for the solution to be usable"; that is, there is no known polynomial time solution. NP-complete problems are commonly said to be intractable; however, the reality is more complex. All known algorithms for solving NP-complete problems require exponential time in the worst case; however, these algorithms nevertheless solve many problems of practical importance astoundingly quickly, and are hence relied upon in a broad range of applications. In practice,

the difficulty of the problem of learning large Bayesian networks from data as perceived by the community is perhaps best captured in this relatively recent quote [173]:

> In our view, inferring complete causal models (i.e., causal Bayesian networks) is essentially impossible in large-scale data mining applications with thousands of variables

Given a Bayesian network with the joint-probability distribution associated with the various nodes, one can use it for inference. Inference can of two types - causal or predictive support (top-down, through parent nodes, generative models specifying how causes generate effects) or diagnostic support (bottom-up, through children nodes, going from effects to causes). Exact computation of conditional probabilities in belief networks is NP-hard. To prove it is NP-Hard [9], [174] transform a well-known NP-complete problem, called 3-Satisfiability (3SAT) to a Decision-problem version of Probabilistic Inference using Belief NETworks. Thus, research should be directed away from the search for a general, efficient probabilistic inference [10] algorithm, and toward the design of efficient special-case, average-case, and approximation algorithms. Exact inference methods include the cycle-cutset conditioning and variable elimination. [137] Cycle-cutset solves the difficulties of multiply connected (loopy) graphs by identifying nodes that, when removed, would reveal a singly connected subgraph as shown in Figure 5.1. In variable elimination, one simply picks any non-deleted node x in the graph, and then adds links to all the neighbours of x. Node x is then deleted. One repeats this until all nodes have been deleted. [126]

Many investigators in the AI community have tacitly assumed that algorithms for performing approximate inference with belief networks are of polynomial complexity. Indeed, special cases of approximate inference can be performed in time polynomial in the input size. [175] discovered that the general problem of approximating conditional probabilities with

[9] To prove that a problem Q' is NP-hard, it is sufficient to transform a known NP-complete problem Q to Q' and to show that this transformation can be done efficiently (i.e., in time that is polynomial in the size of Q).

[10] $P(S_1|S_2)$, where S_1 is either a single instantiated variable or a conjunction of instantiated variables, and S_2 is a conjunction of instantiated variables

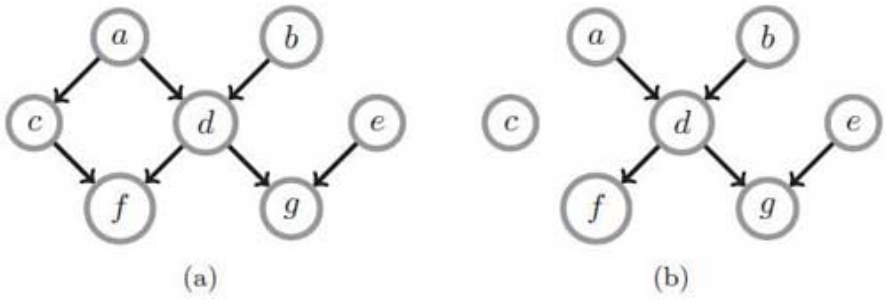


Figure 5.1: A multiply connected graph (a) reduced to a singly connected graph (b) by conditioning on the variable c. [126]

belief networks, like exact inference, resides in the NP-hard complexity class. Probabilistic inference using certain restricted types of belief networks can be performed efficiently. For example, the message passing algorithm can perform probabilistic inference using singly connected networks (also called polytrees) in time that is linear as a function of the size of the belief network. [134] [142]

Since the message passing approach can lead to double counting for undirected cycles, the resolution for that is to convert the BN into a tree, by clustering nodes together, to form what is called a junction tree, and use that instead. Probabilistic inference using multiply connected networks with all variables instantiated to specific values also requires only time linear in the size of the network. This takes advantage of decomposability that is a powerful technique in solving many kinds of network problems. However, inference using multiply connected networks containing uninstantiated variables appears to be much more computationally difficult. The existing algorithms have a time complexity that, in the worst case, is exponential as a function of the number of uninstantiated variables in the network. [174] [141] Generalized distributive law (GDL), also known as sum-product algorithm that can be visualized as a factor graph [11], includes as special cases the Baum-Welch algorithm, the fast Fourier transform (FFT) on any finite Abelian group, the Gallager-Tanner-Wiberg decoding algorithm, Viterbi's algorithm, the BCJR algorithm, Pearl's belief propagation

[11] $ab + ac = a(b + c)$

algorithm, the Shafer-Shenoy probability propagation algorithm, and the turbo decoding algorithm and is guaranteed to give exact answers only in the junction tree condition. [176] [177]

An alternative approach for undirected graphs via which the above mentioned message passing algorithm is prevented from visiting the same node twice is a Revised Polytree Algorithm. When this new algorithm is applied via cutset conditioning to general networks it obtains not just the significant improvement in speed, but also a much simpler form for combination. Furthermore, the revised algorithm requires only minor modifications to existing implementations of the Polytree Algorithm. [178]

Knowing that a problem is NP-hard is important, because it suggests that any attempt at a general, exact, efficient solution is unlikely to be successful. Thus, attempts to develop such an algorithm should be given very low priority. Alternative strategies should be sought that include average-case,special-case, and approximation algorithms. Approximation algorithms produce an inexact, bounded solution, but guarantee that the exact solution is within those error bounds. Monte Carlo simulation techniques is an approach for approximate inference to produce a point-valued probability estimate, plus a standard error of that estimate. The estimate improves as sampling proceeds. Other approaches include MCMC methods (Gibbs Sampling and Metropolis-Hastings algorithm), loopy belief propagation and variational methods [179] - the latter two approaches are based on the law of large numbers to approximate large sums of random variables by their means. Figure 5.2 shows the graphical models family and their uses.

Learning structure is harder than learning parameters. Learning a model or model parameters from data forces us to deal with uncertainty since with only limited data we can never be certain which is the correct model. Four cases can arise in the learning structure process [141] as shown in Table 5.1.

Most of the galaxies detected in very deep exposures are in practice inaccessible to spectroscopical analysis. The spectroscopical sample only comprises $\approx 20\%$ of the $I < 27$

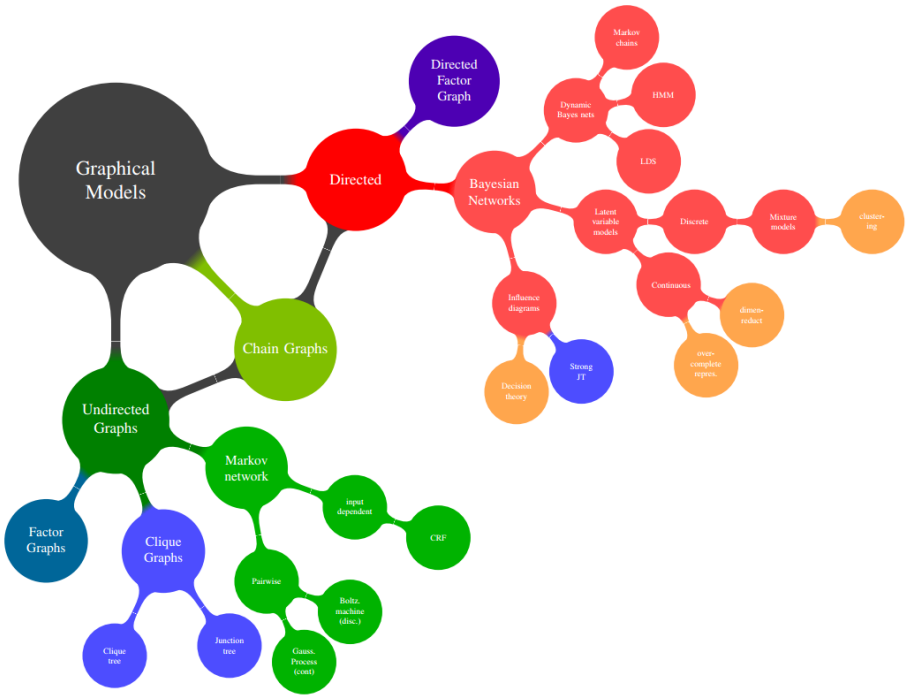


Figure 5.2: Graphical Models [126]

Table 5.1: Structure and Observability impact on BN structure Learning Method [141]

Structure	**Observability**	**Learning Method**
Known	Full	MLE
Known	Partial	EM (or Gradient Descent)
Unknown	Full	Search through Model Space
Unknown	Partial	EM + Search through Model Space

galaxies detected in the Hubble Deep Field North (HDF-N). In contrast, surprisingly accurate photometric redshifts were quickly obtained for most of the HDF-N galaxies using maximum-likelihood methodology. However, a significant fraction ($\approx 10\%$) of redshift estimates presented large, "catastrophic" errors ($\triangle z \geq 1.0$). With Bayesian statistical approach it is possible to obtain fast, inexpensive andmore importanthighly reliable photometric redshifts for $\approx 90\%$ of the $I < 27$ HDF-N galaxies. [86] discusses the use of prior probabilities

and Bayesian marginalization to facilitate the inclusion of relevant knowledge, such as the expected shape of the redshift distributions and the galaxy type fractions. Moreover, the Bayesian formalism can be easily generalized to deal with a wide range of problems that make use of photometric redshifts.

There is excellent agreement between the ≈ 130 HDF-N spectroscopic redshifts and the predictions of the method, with a rms error of $\delta z \approx 0.06(1 + zspec)$ up to $z < 6$ and no outliers nor systematic biases (refer Figure 5.3). If the method is further tested by estimating redshifts in the HDF-N but restricting the color information to the UBVI filters; the results are shown to be significantly more reliable than those obtained with maximum-likelihood techniques. [86]

Zurich Extragalactic Bayesian Redshift Analyzer (ZEBRA) is a more sophisticated Bayesian template-fitting photometric redshift code compared to its predecessor, BPZ. It includes photometry check mode that checks and corrects the photometry in certain filters, a template optimization mode to improve the standard set of templates in specified redshift bins using a training set of galaxies with spectroscopic redshifts, and the ability to calculate a prior self-consistently from the photometric catalogue when ZEBRA is run in its Bayesian mode. Bayesian mode of ZEBRA produces considerably better photometric redshifts than the maximum likelihood mode. [84] used a smoothing kernel to smooth the prior after every iteration.

The ALHAMBRA survey has observed 8 different regions of the sky using a photometric system with 20 contiguous $\sim 300\mathring{A}$ filters covering the optical range, combining them with deep $JHKs$ imaging. The observations were carried out with the Calar Alto 3.5m telescope. The photometric system was designed to maximize the effective depth of the survey in terms of accurate spectral-type and photo-zs estimation along with the capability of identification of relatively faint emission lines. The multicolor photometry and photo-zs were measured for $\sim$438,000 galaxies. The photometric redshifts were calculated with the BPZ2.0 software resulting in a catastrophic error or 1% , which includes new empirically calibrated templates and priors. Our photo-zs have a precision of $\delta_z/(1 + z_s) = 1\%$ for $I < 22.5$ and 1.4% for

BPZ photometric redshifts, plotted against the spectroscopical ones for the HDF-N. Some interpolation (three intermediate spectra) is performed between the main template spectra mentioned in the text, which slightly reduces (by 10%) the small-scale scatter. One of two outliers in the bottom panel of Fig. 3 is assigned a correct redshift by BPZ. The other is the only object discarded when a $p_{\Delta z} < 0.95$ threshold is applied (see text). The rms scatter around the continuous line is $\Delta z_b/(1+z_b) = 0.059$.

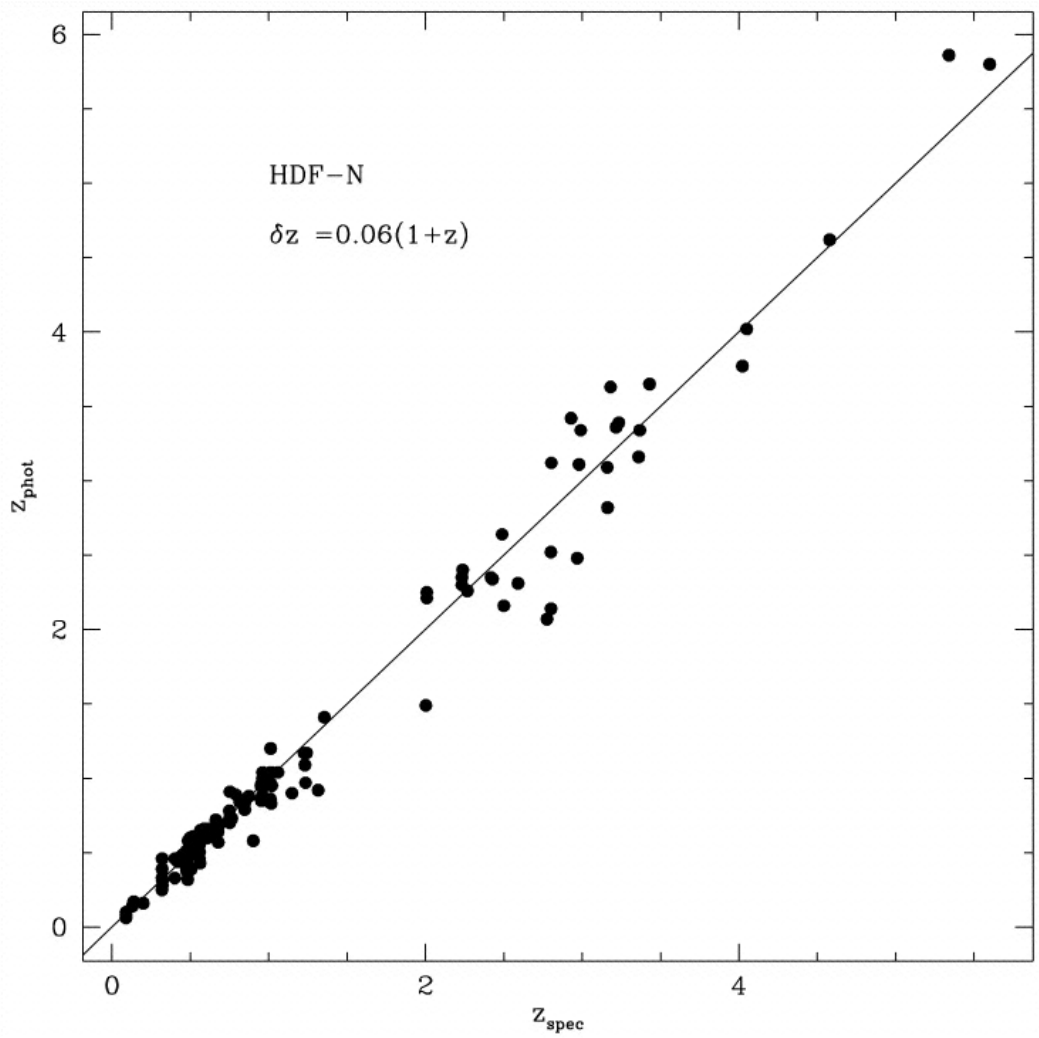


Figure 5.3: BPZ Photometric Redshift Estimation [86]

$22.5 < I < 24.5$. The mean redshift of the survey data is 0.56 for $I < 22.5AB$ and 0.86 for $I < 24.5AB$. [62]

Weak-lensing signals are concentrated towards the magnitude limits of surveys, and yet it is here that photometric measurement errors make photometric redshift estimation the most unreliable. These errors lead not only to a broad distribution of redshift uncertainty, but also to increasing severity of the degeneracy between galaxy type and redshift. Furthermore, the effect of measurement errors is to make the sample distribution of estimated redshifts different from the true distribution of redshifts, which inevitably leads to bias in the values of cosmological parameters estimated from the overall sample. [83] adopts a Bayesian approach using a prior calculated from galaxy luminosity function (GLFs) and its evolution, thereby able to both correct for the bias in the sample distribution and to obtain redshift distributions that smoothly converge on the prior distribution at the limit of faint magnitudes. For each galaxy it does not assign a single definite redshift, but rather the entire posterior probability distribution in redshift, in order to avoid bias.

5.3 Bayesian Statistics in R Statistical Software

bnlearn is an R package which includes several algorithms for learning the structure of Bayesian networks with either discrete or continuous variables. Both constraint-based and score-based algorithms are implemented, and can use the functionality provided by the snow package to improve their performance via parallel computing. Several network scores and conditional independence algorithms are available for both the learning algorithms. Also, several functions for parameter estimation, parametric inference, bootstrap, cross-validation and stochastic simulation are available. Advanced plotting options are provided by the Rgraphviz and lattice packages. [180] [181] [182] [125]

Constraint-based algorithms, also known as conditional independence learners, are all optimized derivatives of the Inductive Causation algorithm. These algorithms use conditional independence tests to detect the Markov blankets of the variables, which in turn are used to compute the structure of the Bayesian network. Score-based learning algorithms

are general purpose heuristic optimization algorithms which rank network structures with respect to a goodness-of-fit score. Hybrid algorithms combine aspects of both constraint-based and score-based algorithms, as they use conditional independence tests (usually to reduce the search space) and network scores (to find the optimal network in the reduced space) at the same time.

bnlearn implements the following constraint-based learning algorithms (the respective bnlearn package function names are reported in parenthesis):

- Grow-shrink (gs): Based on the grow-shrink Markov blanket, the simplest Markov blanket detection algorithm ([183]) used in a structure learning algorithm.

- Incremental association (iamb): Based on the incremental association Markov blanket (IAMB) algorithm ([184]), which is based on a two-phase selection scheme (a forward selection followed by an attempt to remove false positives).

- Fast incremental association (fast.iamb): A variant of IAMB which uses speculative stepwise forward selection to reduce the number of conditional independence tests (Yaramakala and Margaritis 2005).

- Interleaved incremental association (inter.iamb): Another variant of IAMB which uses forward stepwise selection ([184]) to avoid false positives in the Markov blanket detection phase.

Note: The computational complexity of these algorithms is polynomial in the number of tests, usually $O(N^2)$ ($O(N^4)$ in the worst case scenario), where N is the number of variables. Execution time scales linearly with the size of the data set.

bnlearn implements the following score-based learning algorithms (the respective bnlearn package function names are reported in parenthesis):

- Hill-Climbing (hc): a hill climbing greedy search on the space of the directed graphs. The optimized implementation uses score caching, score decomposability and score equivalence to reduce the number of duplicated tests.

- Tabu Search (tabu): a modified hill climbing able to escape local optima by selecting a network that minimally decreases the score function.

bnlearn implements the following hybrid learning algorithms (the respective bnlearn package function names are reported in parenthesis):

- Max-Min Hill-Climbing (mmhc): a hybrid algorithm which combines the Max-Min Parents and Children algorithm (to restrict the search space) and the Hill-Climbing algorithm (to find the optimal network structure in the restricted space).

- Restricted Maximization (rsmax2): a more general implementation of the Max-Min Hill-Climbing, which can use any combination of constraint-based and score-based algorithms.

Other (constraint-based) local discovery algorithms that learn the structure of the undirected graph underlying the Bayesian network, which is known as the skeleton of the network or the (partial) correlation graph. Therefore all the arcs are undirected, and no attempt is made to detect their orientation. They are often used in hybrid learning algorithms.

- Max-Min Parents and Children (mmpc): a forward selection technique for neighbourhood detection based on the maximization of the minimum association measure observed with any subset of the nodes selected in the previous iterations.

- Hiton Parents and Children (si.hiton.pc): a fast forward selection technique for neighbourhood detection designed to exclude nodes early based on the marginal association. The implementation follows the Semi-Interleaved variant of the algorithm.

- Chow-Liu (chow.liu): an application of the minimum-weight spanning tree and the information inequality. It learn the tree structure closest to the true one in the probability space.

- ARACNE (aracne): an improved version of the Chow-Liu algorithm that is able to learn polytrees.

Bayesian Network classifiers are aimed at classification, and favour predictive power over the ability to recover the correct network structure. The implementation in bnlearn assumes that all variables, including the classifiers, are discrete.

- Naïve Bayes (naive.bayes): a very simple algorithm assuming that all classifiers are independent and using the posterior probability of the target variable for classification.
- Tree-Augmented Naïve Bayes (tree.bayes): a improvement over Naïve Bayes, this algorithms uses Chow-Liu to approximate the dependence structure of the classifiers.

The conditional independence tests used in constraint-based algorithms for the discrete case (categorical variables) are the following:

- mutual information: an information-theoretic distance measure. It's proportional to the log-likelihood ratio (they differ by a 2n factor) and is related to the deviance of the tested models. The asymptotic chi-square test (mi and mi-adf, with adjusted degrees of freedom), the Monte Carlo permutation test (mc-mi), the sequential Monte Carlo permutation test (smc-mi), and the semiparametric test (sp-mi) are implemented.
- shrinkage estimator for the mutual information (mi-sh): an improved asymptotic chi-square test based on the James-Stein estimator for the mutual information.
- Pearson's χ^2: the classical Pearson's χ^2 test for contingency tables. The asymptotic χ^2 test (x2 and x2-adf, with adjusted degrees of freedom), the Monte Carlo permutation test (mc-x2), the sequential Monte Carlo permutation test (smc-x2) and semiparametric test (sp-x2) are implemented.

The network scores available for the discrete case (categorical variables) are the following:

- the multinomial log-likelihood (loglik) score, which is equivalent to the entropy measure used in Weka.
- the Akaike Information Criterion score (aic).

- the Bayesian Information Criterion score (bic), which is equivalent to the Minimum Description Length (MDL) and is also known as Schwarz Information Criterion.
- the logarithm of the Bayesian Dirichlet equivalent score (bde), a score equivalent Dirichlet posterior density.
- the logarithm of the modified Bayesian Dirichlet equivalent score (mbde) for mixtures of experimental and observational data (not score equivalent).
- the logarithm of the K2 score (k2), a Dirichlet posterior density (not score equivalent).

Other R-based packages which are able to either learn the structure of a Bayesian network or fit and manipulate its parameters are pcalg, which implements the PC algorithm and focuses on the causal interpretation of Bayesian networks; deal , which implements a hill-climbing search for mixed data; and the suite composed by gRbase, gRain and gRc.

5.3.1 Naive Bayes

Naïve Bayes methods are a set of supervised learning algorithms based on applying Bayes theorem with the "naive" assumption of independence between every pair of features. bnlearn package implements it as well. The network generated as part of Naïve Bayes is simplistic and is as shown in Figure 5.4. There is no step involved in formulating the network based on training data. The error analysis when modeled with color and morphology attributes, and all attributes is shown in Table 5.2 and Table 5.3 respectively.

5.4 Bayesian Photomorphic Redshift

Structural learning becomes significantly more difficult when variables are continuous, as the number and type of possible dependence relations and interactions becomes infinite. However, under some assumptions, such as linear relations and conditionally normal distributions, effective algorithms can be worked out. In our study of application of Bayesian

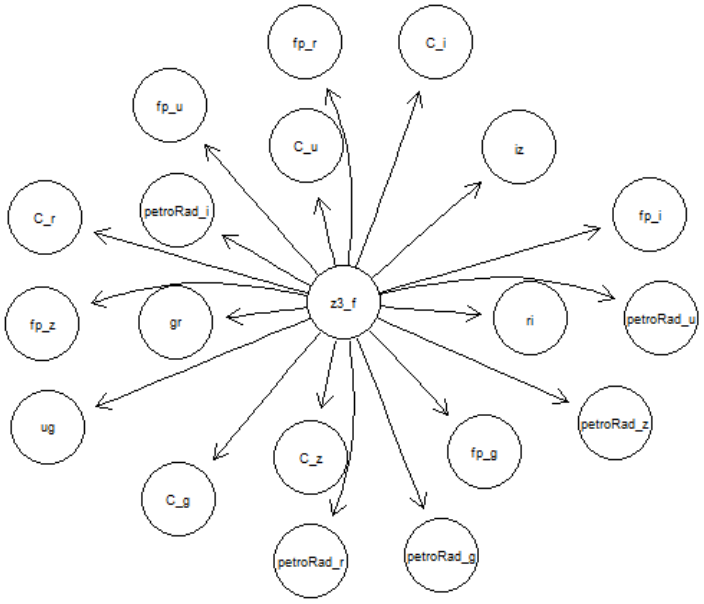


Figure 5.4: Naïve Bayes Model - Color and Morphology Photo-z

Table 5.2: Naïve Bayes Method - Color and Morphology Attributes Photo-z - No. of Quantiles: 5

Range	**Count**	**Mean z-Spec**	**Mean z-Photo**	**Mean Error**	**RMS**	**Catastrophic Error %**
(0-0.1]	29110	0.061	0.066	0.036	0.044	0.95
(0.1-0.2]	25811	0.141	0.128	0.043	0.06	9.84
(0.2-0.3]	4331	0.246	0.241	0.059	0.08	17.55
(0.3-0.4]	3352	0.348	0.328	0.051	0.076	12.77
(0.4-0.5]	1540	0.443	0.372	0.08	0.11	25.91

network to the photometric redshift estimation problem using color, magnitude and morphology data, we will use bnlearn R package primarily with the score-based Hill-Climbing algorithm with different precisions of spectroscopic redshift for training. The other learning algorithm are also attempted to determine their usability and performance.

Note: All the experiments were run using spectroscopic redshift with precision three, for example, spec-z = 0.314. Better performance of these method runs are expected when

Table 5.3: Naïve Bayes Method - All Attributes Photo-z - No. of Quantiles: 5

Range	**Count**	**Mean z-Spec**	**Mean z-Photo**	**Mean Error**	**RMS**	**Catastrophic Error %**
(0-0.1]	30884	0.064	0.063	0.034	0.045	1.8
(0.1-0.2]	28276	0.139	0.133	0.039	0.053	8.14
(0.2-0.3]	4662	0.239	0.237	0.038	0.057	6.65
(0.3-0.4]	3698	0.349	0.369	0.044	0.06	9.82
(0.4-0.5]	1696	0.441	0.428	0.039	0.051	3.18

using spectroscopic redshift with precision one and two since there will be larger amount of data per redshift value. This will also naturally lead to higher certainty for a estimate and lesser catastrophic error rate. This has been attempted with precision two redshift data, for example, spec-z = 0.31.

The data consisting of 700,777 galaxies is split into training and test. Training is ninety percent of the data (630,699 galaxies). Test is the remaining ten percent of the data (70,078 galaxies). The Bayesian network is modeled using one of the learning algorithm against the training data. 50,000 random observations are generated from the fitted model based on the known values for the other attributes (evidence) to formulate the conditional probability table (CPT) of the possible estimated value(s) for each and every test data-point. Refer Table 5.4 for an example of CPT associated with $ug \rightarrow z$ based on an evidence under investigation. Using the model and the associated CPT per node, an estimate of the redshift is predicted for the test data. The estimate includes the potential redshift values as well as the probability of the occurance of each value. A plot showing the probability associated with every possible photometric redshift estimate is shown in Figure 5.5; when selecting maximum only, the estimate will be z=0.078 with certainty 0.01. The value with the maximum probability is returned - in other words, the certainty that the redshift can be that particular value. When combining estimates from different models in an ensemble methods-based application, a likely method could be to take all the values exceeding a particular threshold value. One or more than one estimate and their associated probability can be retrieved from any of the models to be used in the ensemble estimate.

Table 5.4: bnlearn CPT associated with $(ug \rightarrow z)$ based on an evidence under investigation

z	$[-11.2, -5.9]$	$(-5.9, -0.639]$	$(-0.639, 4.62]$	$(4.62, 9.88]$	$(9.88, 15.2]$
0.000	0.11	0.06	0.00	0.01	0.11
0.001	0.03	0.01	0.00	0.00	0.00
0.002	0.00	0.00	0.00	0.00	0.00
...	...	...	...	...	...

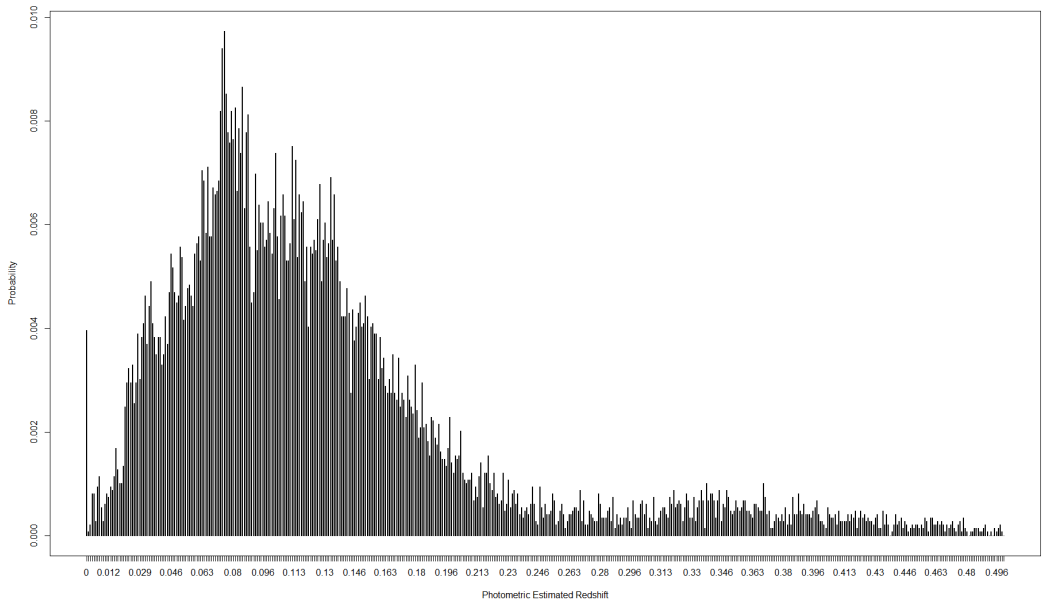
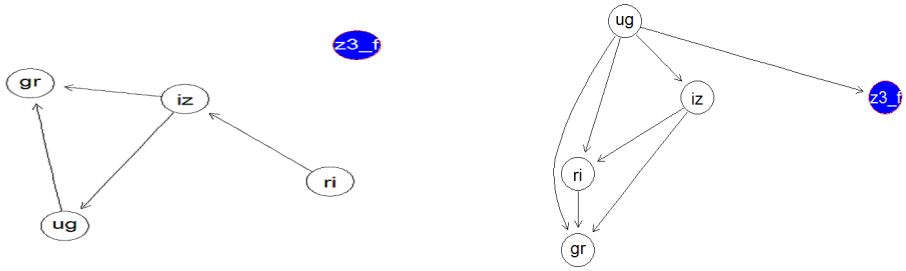


Figure 5.5: Probability of possible photometric redshift estimates based on certain evidence

The number of networks created for each execution of an algorithm was 200. Using the 200 networks created in the model generation phase, the arcs were evaluated and if the arc occured in 85 percent or more of the networks, they were included in the final generated network. Depending on the number of variables involved, a single run can result in a graph leaving out redshift as an separate single node network as shown in Figure 5.6a. The averaged network is as shown in Figure 5.6b. Going forward, the averaged and fitted resulting directed network will be only shown. bnlearn cannot predict based on undirected network, thus, algorithms resulting in undirected network will need to be manipulated to

add arc directions before they can be used. Matching similar networks from other algorithms and domain experience helps in determining the arc directions. For the score-based Hill-Climbing (hc) learning algorithm, the Bayesian Dirichlet equivalent score (bde) score was used with the imaginary sample size as 10. A more exhaustive study of changing the score and imaginary sample size should be undertaken to investigate it's impact on defining the overall network structure.



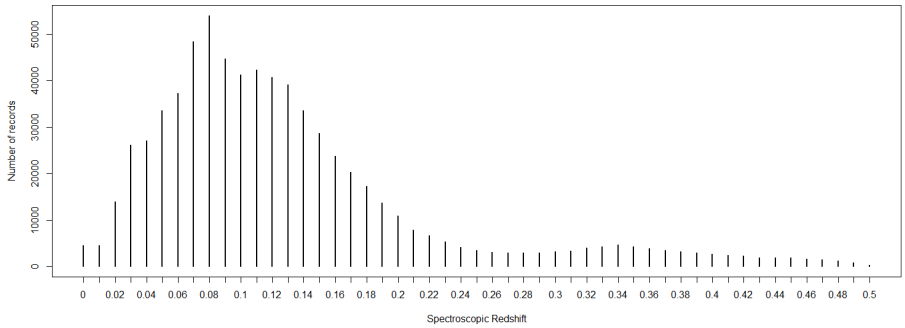
(a) HC Single Run (b) HC Averaged and Fitted

Figure 5.6: bnlearn Hill-Climbing (HC) Method - Color Only Photo-z Error Analysis

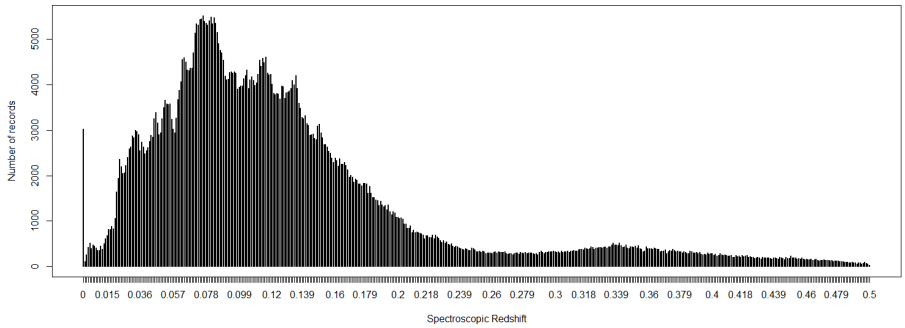
5.4.1 Discretization Method

Since continuous data needs to be converted to discrete categorical data before Bayesian network learner can be applied, the discretization method employed plays a significant role. If the discretization method introduces a bias in the distribution of data, it will influence the estimator to predict more often in the heavily populated bin (also termed level). This includes behaviors such as majority of the data in certain bin or a wide range for an attribute though most of the data is within a narrow range. The distribution of data across the different redshift values in two-digit and three-digit precision is as shown in Figure 5.7a and Figure 5.7b respectively.

Interval versus quantile discretization was used in the model generation. Interval was used for the analysis using color data only. Since data is concentrated in $z = 0 - 0.25$ range, the interval discretization resulted in low predictive accuracy with strong correlation to



(a) Two-digit Precision



(b) Three-digit Precision

Figure 5.7: SDSS Galaxy Data - Record count per redshift

$u-g$ color. It resulted in near 100% catastrophic error rate in all the five redshift ranges of $z = 0-0.5$ with intervals of 0.1. Based on this, the interval method was not analyzed for the remaining combination of attributes (ColorMag, ColorMorphology, Six highest correlated variables and All). In contrast, however, the quantile method with five levels resulted in 0.3% catastrophic error rate in the most populated $z = 0 - 0.1$ range (as shown in Table 5.5). The models for color and magnitude only subset of the data continued to model the strong correlation between z and $u - g$ color when using interval versus using quantile disretization. Refer Figure 5.8 for the interval discretization-based model and Figure 5.9 for the quantile discretization-based model for color and magnitude attributes only subset of the data. Improved accuracy is displayed when the redshift attribute is used with two-digit

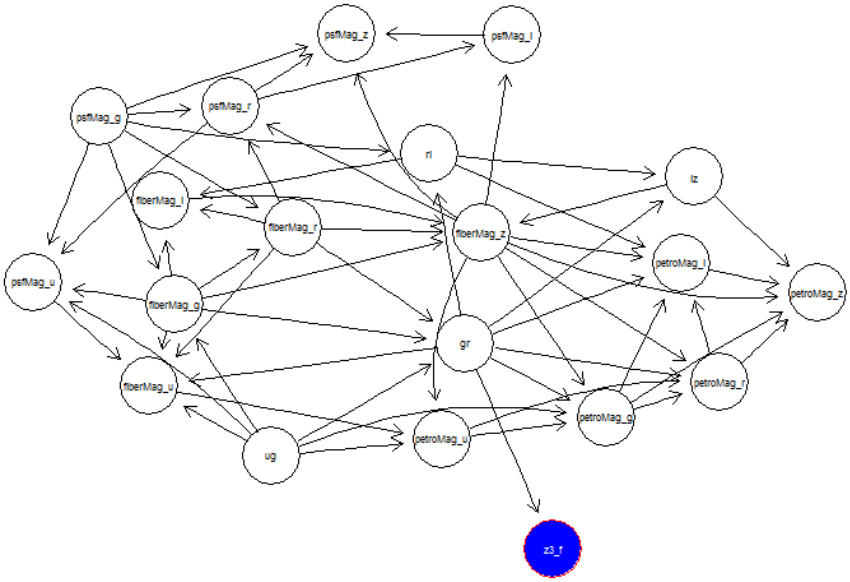


Figure 5.8: bnlearn HC Method Model - Color and Magnitude Photo-z - No. of Intervals: 5

precision since there is more data contributing to any given value instead of being split among multiple values of higher precision. Two-digit precision redshift data also requires lesser memory consumption and faster computation due to the the reduced size of the CPT maintained per node.The precision of the remaining attributes do not matter as they are split into quantile breaks prior to their use.

Table 5.5: bnlearn HC Method - Color Only Photo-z - Number of Quantiles = 5

Range	**Count**	**Mean z-Spec**	**Mean z-Photo**	**Mean Error**	**RMS**	**Catastrophic Error %**
(0-0.1]	31037	0.064	0.064	0.023	0.031	0.33
(0.1-0.2]	27951	0.14	0.114	0.038	0.054	6.61
(0.2-0.3]	4535	0.239	0.151	0.116	0.147	41.48
(0.3-0.4]	3738	0.348	0.156	0.202	0.242	71.64
(0.4-0.5]	1741	0.441	0.175	0.267	0.313	75.76

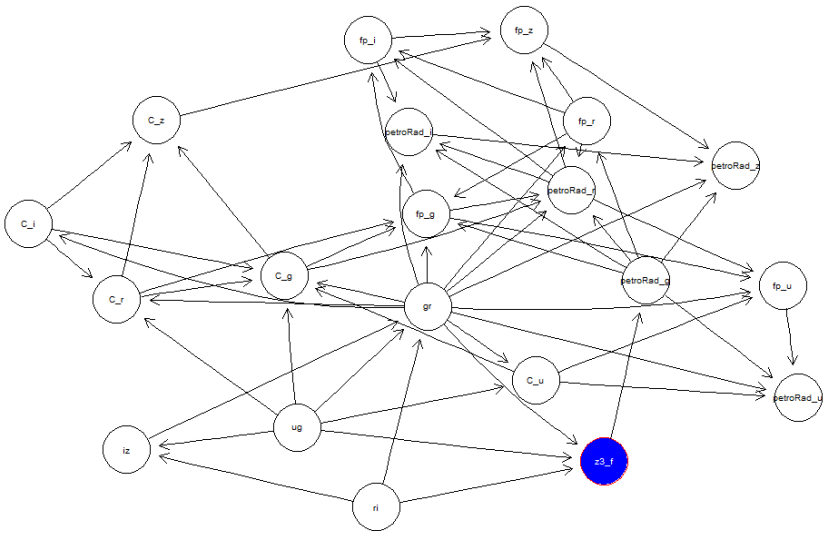


Figure 5.9: bnlearn HC Method Model - Color and Magnitude Photo-z - No. of Quantiles: 5

Other discretization methods should be investigated to study their impact on the predictive and computational performance. Potential methods include the Hartemink method pairwise mutual information method and Bayesian blocks aiming for same uncertainty in every bin.

5.4.2 Photometric Attributes

The entire dataset was subset to color and magnitude attributes only and the three-digit precision of redshift. This subset was studied using multiple algorithms. The constraint-based algorithms did not include redshift in the network and thus, the estimates from the their models are inaccurate and/or often not calculated due to missing probability values. The constraint based algorithms applied are Grow-Shrink(gs), IAMB and Fast IAMB. The model generated by GS and IAMB are as shown in Figure 5.10 and Figure 5.11 respectively.

With three-digit precision and five quantiles, the performance is accuracte only in the redshift range of $z = 0-0.1$ [12] for all the applied algorithms and their relative performance is

[12]Decreasing precision and increasing the number of quantile breaks increases the accuracy in the other ranges as well, as shown later in this section.

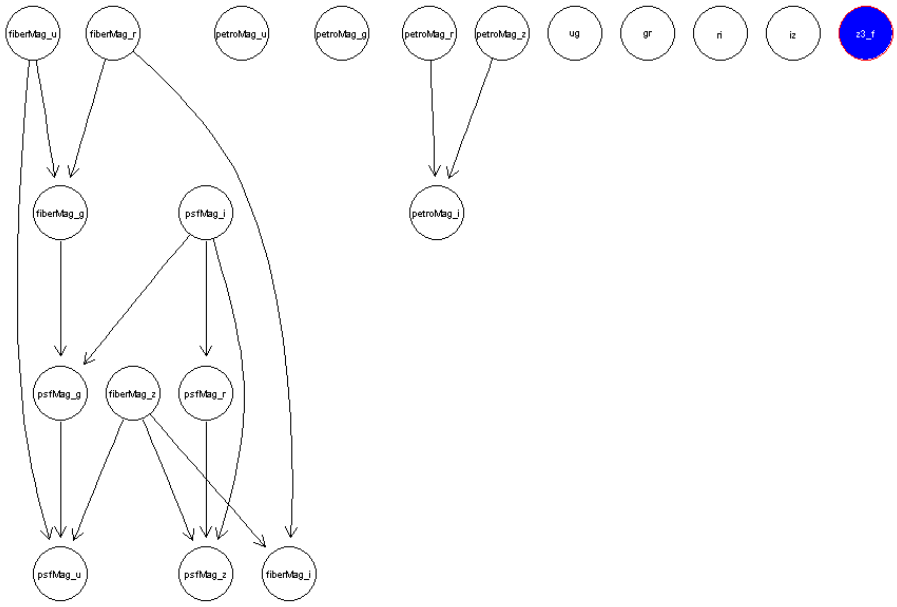


Figure 5.10: bnlearn GS Method Model - Color and Magnitude Photo-z - No. of Quantiles: 5

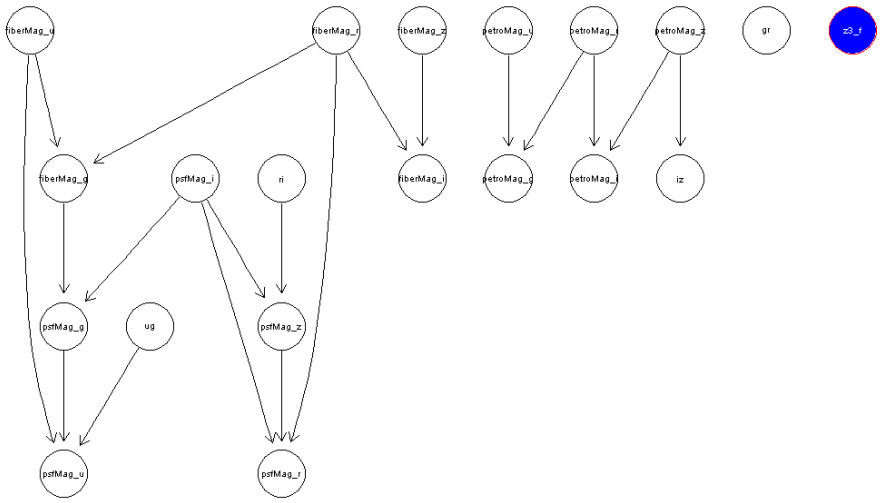


Figure 5.11: bnlearn IAMB Method Model - Color and Magnitude Photo-z - No. of Quantiles: 5

as shown in Table 5.7. Score-based algorithms Tabu (model in Figure 5.12) and HC (model in Figure 5.9) generate the best estimators for this attribute subset. The color attribute $g - r$ is selected by both HC and Tabu as the attribute with connection to redshift. HC additionally identifies the colors $u - g$ and $r - i$. Significant catastrophic error ($> 10\%$) was observed in the ranges other than $z = 0 - 0.2$ as shown in Table 5.6; it corresponds to the error analysis of the Tabu algorithm execution. The Tabu method resulted in a number of high certainty estimates and the spread of the estimates is also limited in any given region, as displayed in Figure 5.14 and Figure 5.13. The green points in Figure 5.14 correspond to the estimates with certainty greater than 0.5 (highest being 1.0).

Table 5.6: bnlearn Tabu Method - Color and Magnitude Photo-z - No. of Quantiles: 5

Range	**Count**	**Mean z-Spec**	**Mean z-Photo**	**Mean Error**	**RMS**	**Catastrophic Error %**
(0-0.1]	15161	0.062	0.073	0.031	0.041	2.28
(0.1-0.2]	8374	0.134	0.126	0.049	0.069	10.58
(0.2-0.3]	1587	0.254	0.234	0.09	0.107	37.93
(0.3-0.4]	3491	0.35	0.185	0.172	0.189	85.56
(0.4-0.5]	1671	0.442	0.176	0.267	0.279	95.45

Table 5.7: bnlearn Methods - Color and Magnitude Photo-z - No. of Quantiles: 5 - Relative Performance of Algorithms $z < 0.1$

Method	**Count**	**Mean z-Spec**	**Mean z-Photo**	**Mean Error**	**RMS**	**Catastrophic Error %**
HC	17943	0.062	0.072	0.031	0.041	2.16
Tabu	15161	0.062	0.073	0.031	0.041	2.28
GS	26(note)	0.060	0.12	0.049	0.081	34.62

5.4.3 Photomorphic Attributes

The entire dataset was subset to color and morphology attributes only and the three-digit precision of redshift. Multiple Bayesian learning algorithms were applied to the data subset.

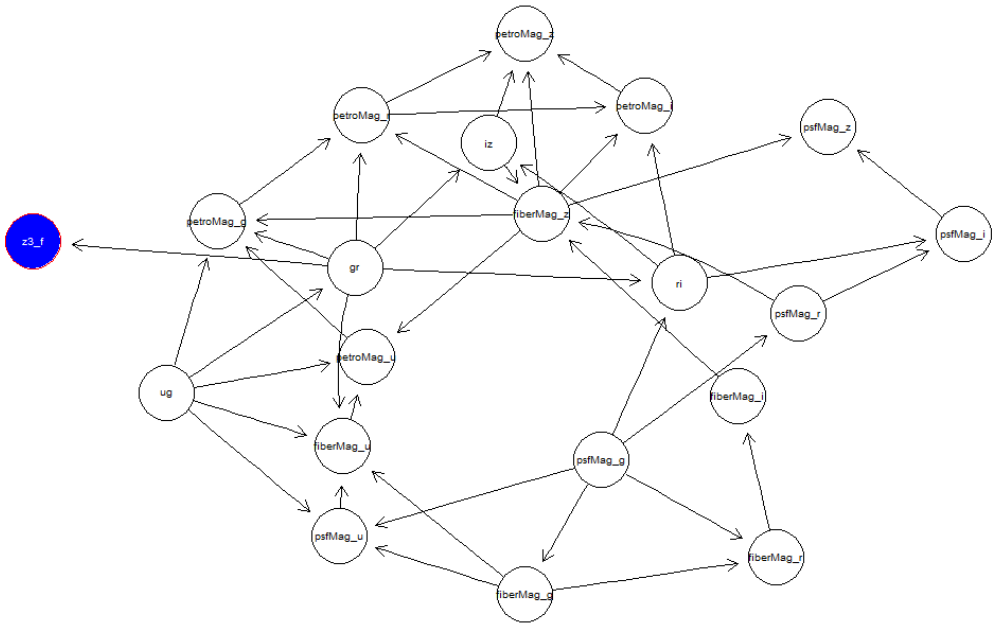


Figure 5.12: bnlearn Tabu Method Model - Color and Magnitude Photo-z - No. of Quantiles: 5

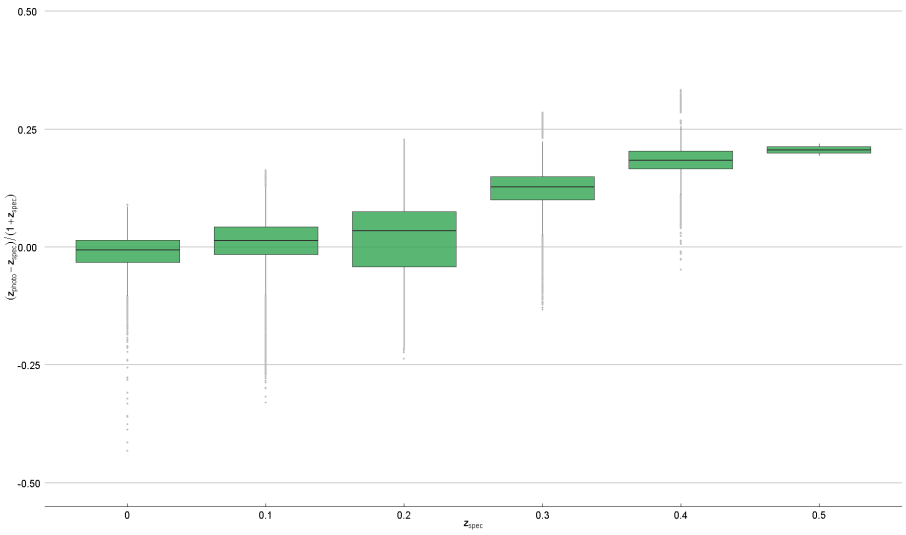


Figure 5.13: bnlearn Tabu Method Test Result across different z-Ranges - Color and Magnitude Photo-z - No. of Quantiles: 5

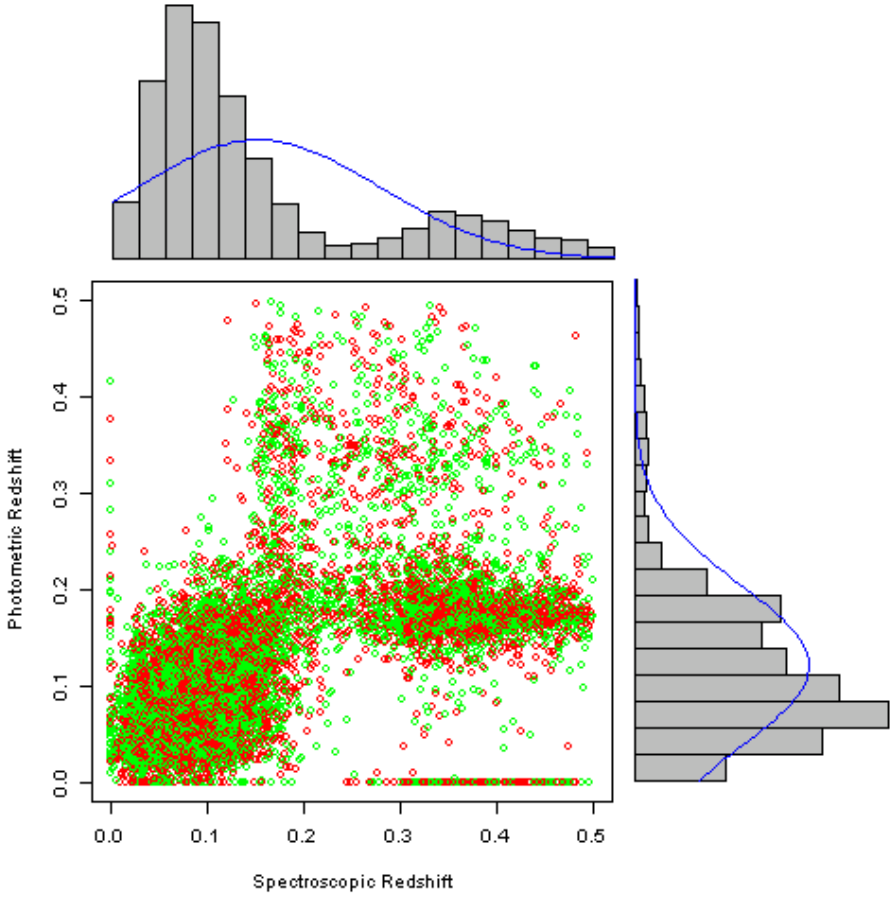


Figure 5.14: bnlearn Tabu Method Test of Model - Color and Magnitude Photo-z - No. of Quantiles: 5 - Spec-z vs. Photo-z

The constraint-based (GS, IAMB, Fast IAMB) and hybrid algorithms (MMHC and Semi-Interleaved HITON-PC) did not include redshift in the network and thus, the estimates from the their models are inaccurate and/or often not calculated due to missing probability values. The models generated using GS, IAMB, Fast IAMB, MMHC and Semi-Interleaved HITON-PC are as shown in Figure5.15, Figure5.16, Figure5.17, Figure5.18 and Figure5.19 respectively.

Score-based algorithms HC (model in Figure 5.20), Tabu (model in Figure 5.21) and Aracne (model in Figure 5.22) generate the best estimators for this attribute subset. The Aracne-based model needed manual updates in providing the arc directions of $fp\_r \rightarrow fp\_i$,

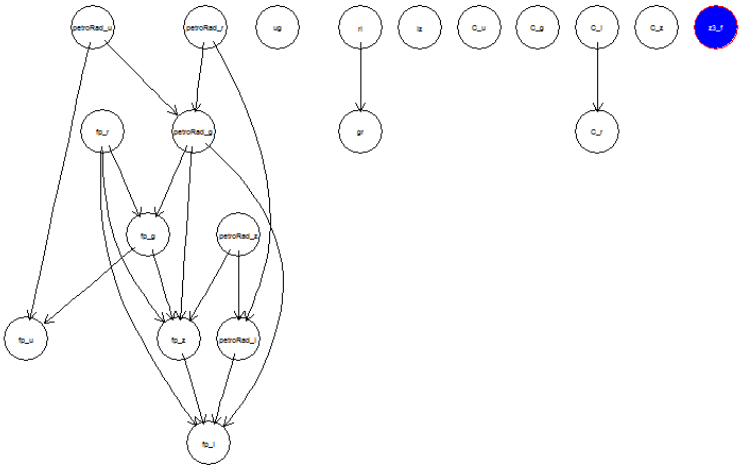


Figure 5.15: bnlearn GS Method Model - Color and Morphology - No. of Quantiles: 5

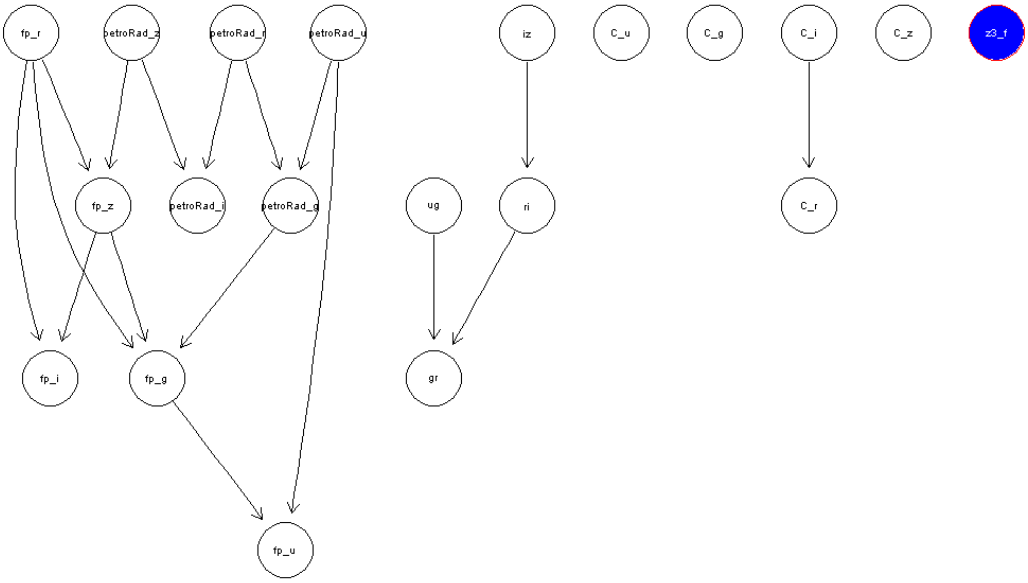


Figure 5.16: bnlearn IAMB Method Model - Color and Morphology Photo-z - No. of Quantiles: 5

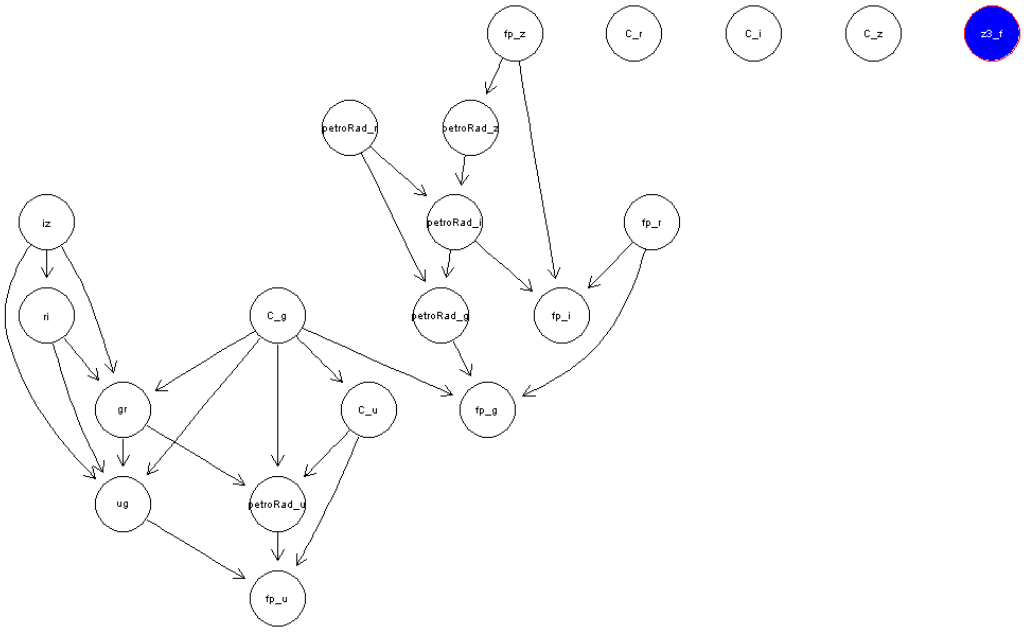


Figure 5.17: bnlearn Fast IAMB Method Model - Color and Morphology Photo-z - No. of Quantiles: 5

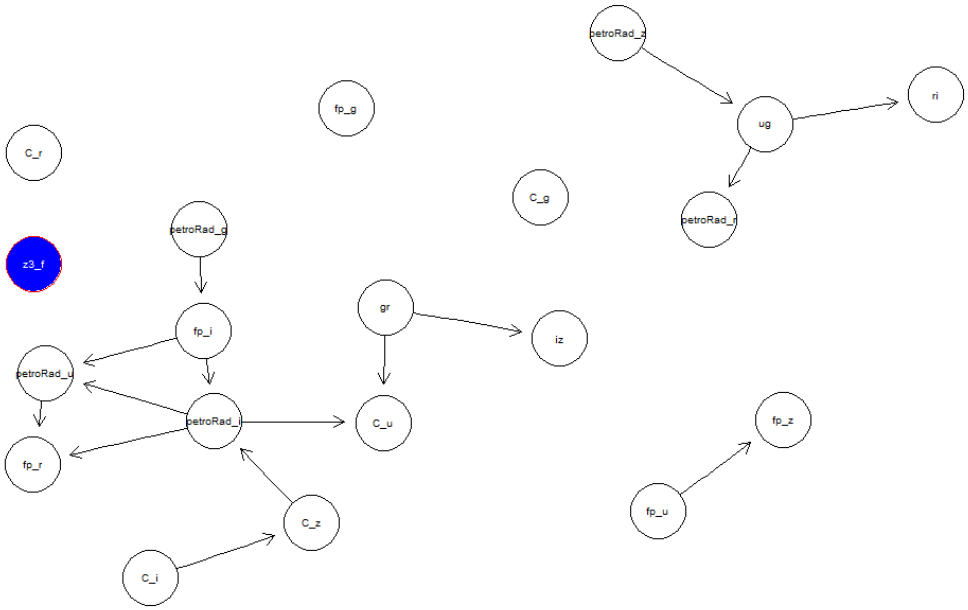


Figure 5.18: bnlearn MMHC Method Model - Color and Morphology Photo-z - No. of Quantiles: 5

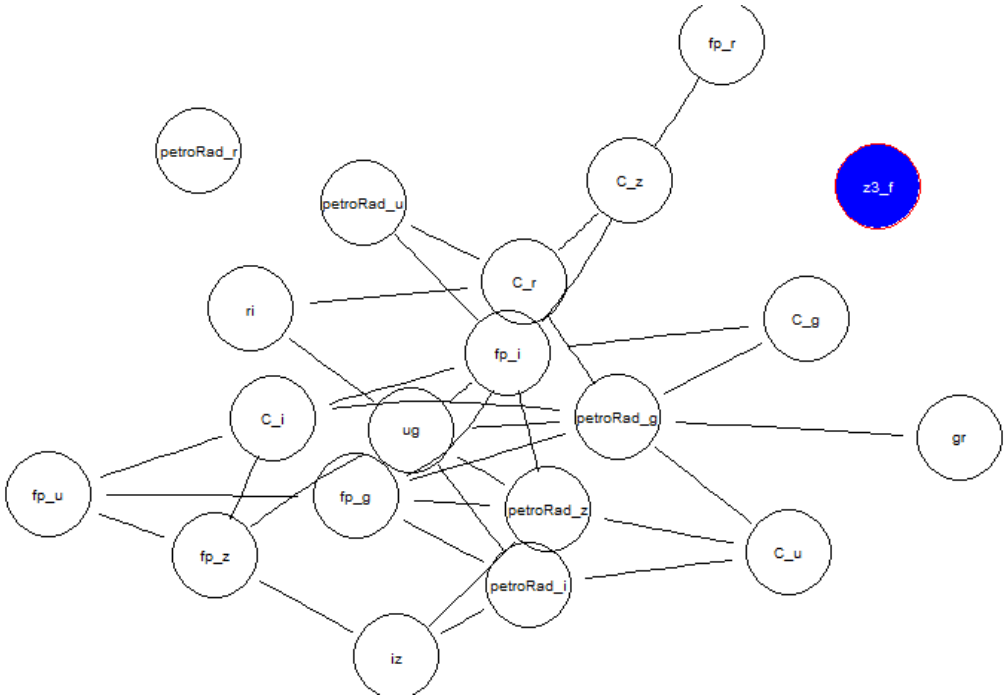


Figure 5.19: bnlearn Semi-Interleaved HITON-PC Method Model - Color and Morphology Photo-z - No. of Quantiles: 5

$petroRad\_i \rightarrow fp_i$, $petroRad\_r \rightarrow petroRad\_i$ and $fp\_r \rightarrow petroRad\_r$. The direction was determined from the models generated by earlier executed methods. Typically, domain expertise should be used to determine the direction. Partially directed networks cannot be used for estimation, thus, the need for the mentioned manual update. Aracne is the first method to show redshift dependence on an attribute other than color $g - r$; it shows dependence on magnitude attribute petroMag_g.

Redshift has dependence on the color attribute $u - g$ and $g - r$ in both the HC and Tabu models. HC additionally includes dependence on the color $r - i$. Significant catastrophic error ($> 10\%$) was observed in the ranges other than $z = 0 - 0.2$. Table 5.8 shows the details of the error analysis of the Tabu algorithm execution. Similarly, Table 5.9 and Table 5.10 give the details of the error analysis of Aracne and HC algorithm respectively. The performance of the three algorithms are pretty close in the $z < 0.1$ range with the largest amount of data; refer Table 5.11.

Both the Tabu and Aracne method resulted in a number of high certainty estimates and the spread of the estimates is also limited in any given region, as displayed in Figure 5.24 and Figure 5.23 for Tabu and Figure 5.26 and Figure 5.25 for Aracne. The green points in these figures correspond to the estimates with certainty greater than 0.5 (highest being 1.0) and the larger the number of high certainty estimates, the more responsive the algorithm in modeling the data as well as the adequacy of the training data.

Table 5.8: bnlearn Tabu Method - Color and Morphology Photo-z - No. of Quantiles: 5

Range	**Count**	**Mean z-Spec**	**Mean z-Photo**	**Mean Error**	**RMS**	**Catastrophic Error %**
(0-0.1]	11042	0.057	0.06	0.029	0.037	1.31
(0.1-0.2]	3643	0.13	0.109	0.047	0.065	8.48
(0.2-0.3]	211	0.241	0.244	0.09	0.109	36.97
(0.3-0.4]	159	0.35	0.302	0.091	0.12	33.96
(0.4-0.5]	59	0.438	0.308	0.14	0.188	40.68

Table 5.9: bnlearn Aracne Method - Color and Morphology Photo-z - No. of Quantiles: 5

Range	**Count**	**Mean z-Spec**	**Mean z-Photo**	**Mean Error**	**RMS**	**Catastrophic Error %**
(0-0.1]	7752	0.054	0.055	0.03	0.038	1.16
(0.1-0.2]	2017	0.13	0.114	0.057	0.078	13.09
(0.2-0.3]	141	0.241	0.28	0.096	0.113	41.13
(0.3-0.4]	112	0.346	0.323	0.076	0.102	27.68
(0.4-0.5]	25	0.429	0.345	0.097	0.114	44

Table 5.10: bnlearn HC Method - Color and Morphology Photo-z - No. of Quantiles: 5

Range	**Count**	**Mean z-Spec**	**Mean z-Photo**	**Mean Error**	**RMS**	**Catastrophic Error %**
(0-0.1]	10712	0.058	0.061	0.028	0.036	0.9
(0.1-0.2]	3762	0.13	0.108	0.043	0.057	6.7
(0.2-0.3]	182	0.236	0.251	0.079	0.101	34.62
(0.3-0.4]	136	0.344	0.328	0.083	0.106	33.82
(0.4-0.5]	42	0.434	0.331	0.114	0.146	45.24

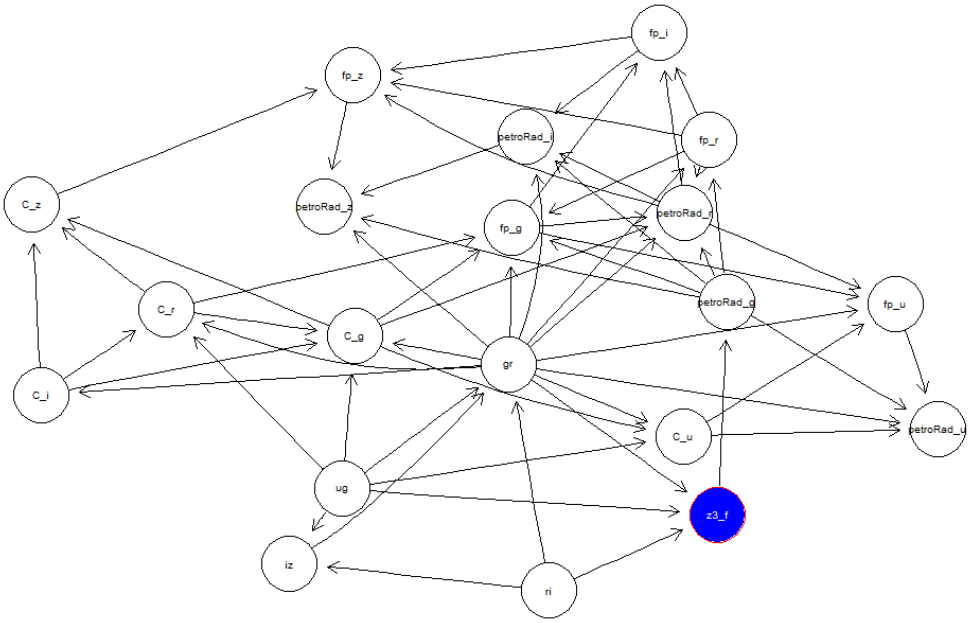


Figure 5.20: bnlearn HC Method - Color and Morphology Photo-z - No. of Quantiles: 5

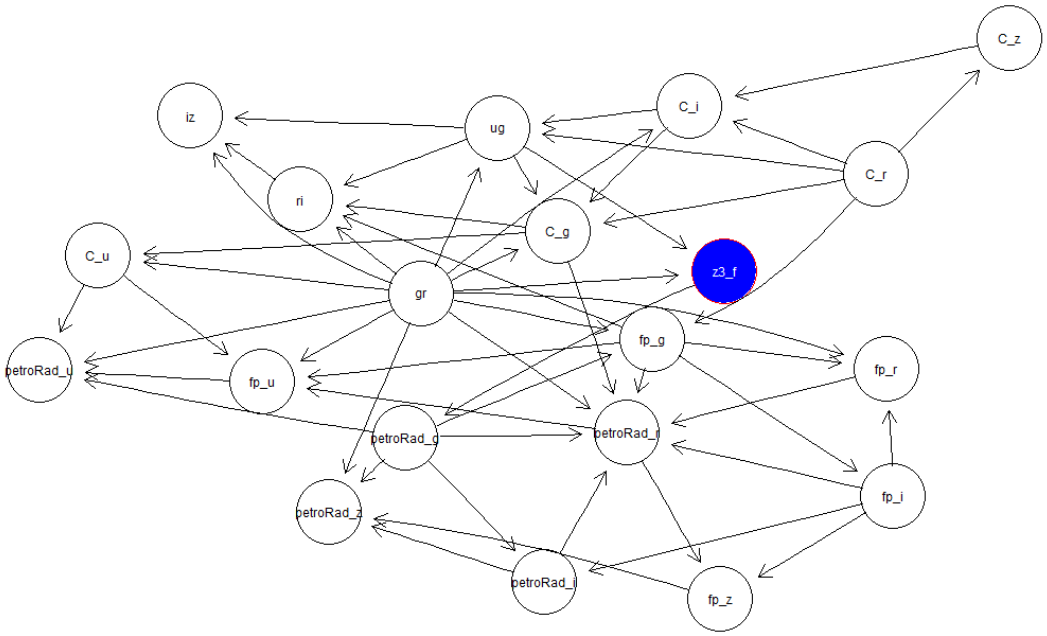


Figure 5.21: bnlearn Tabu Method - Color and Morphology Photo-z - No. of Quantiles: 5

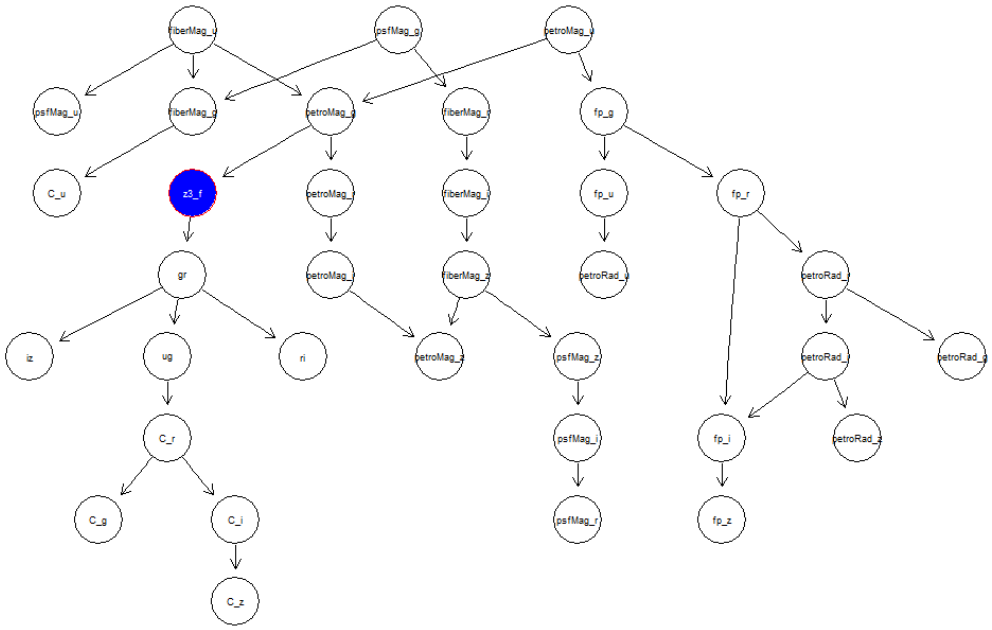


Figure 5.22: bnlearn Arcane Method - Color and Morphology Photo-z - No. of Quantiles: 5

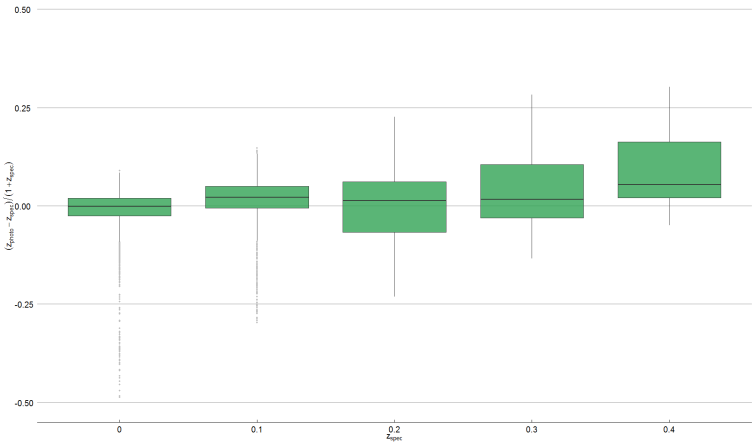


Figure 5.23: bnlearn Tabu Method Test Result across different z-Ranges - Color and Morphology Photo-z - No. of Quantiles: 5

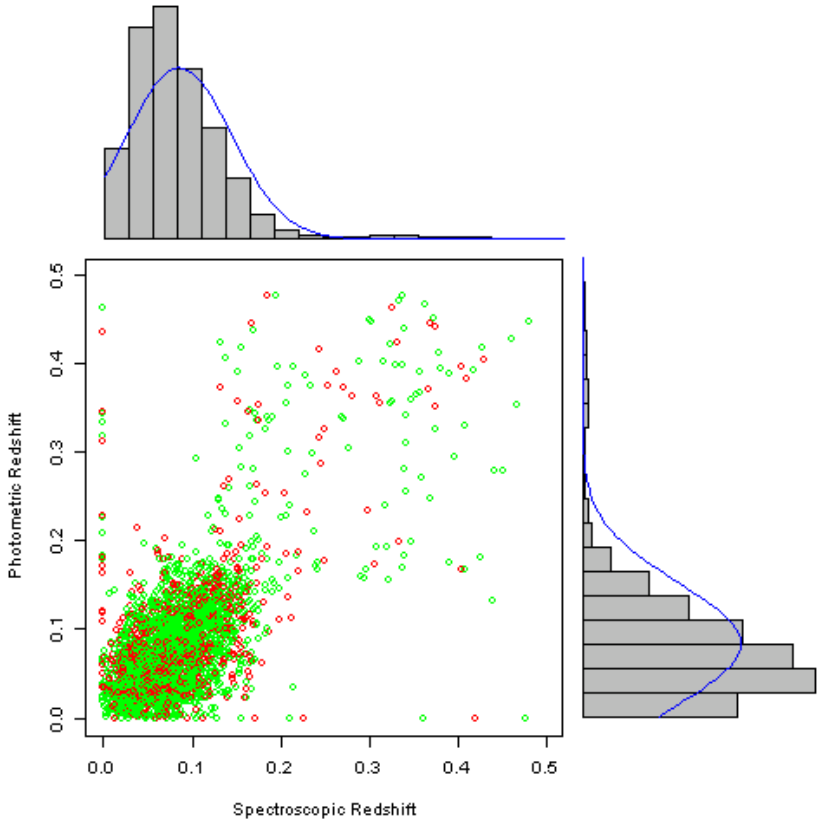


Figure 5.24: bnlearn Tabu Method Test of Model - Color and Morphology Photo-z - No. of Quantiles: 5 - Spec-z vs. Photo-z

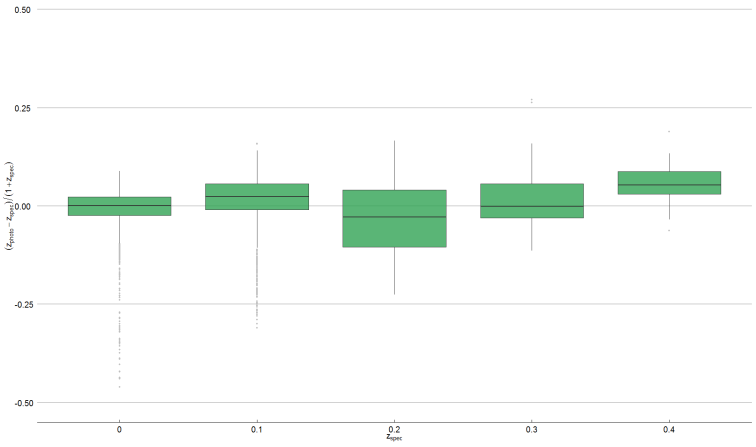


Figure 5.25: bnlearn Aracne Method Test Result across different z-Ranges - Color and Morphology Photo-z - No. of Quantiles: 5

Table 5.11: bnlearn Methods - Color and Morphology Photo-z - No. of Quantiles: 5 - Relative Performance of Algorithms $z < 0.1$

Method	Count	Mean z-Spec	Mean z-Photo	Mean Error	RMS	Catastrophic Error %
HC	10712	0.058	0.061	0.028	0.036	0.90
Tabu	11012	0.057	0.060	0.029	0.037	1.12
Arcane	7752	0.054	0.055	0.030	0.038	1.16

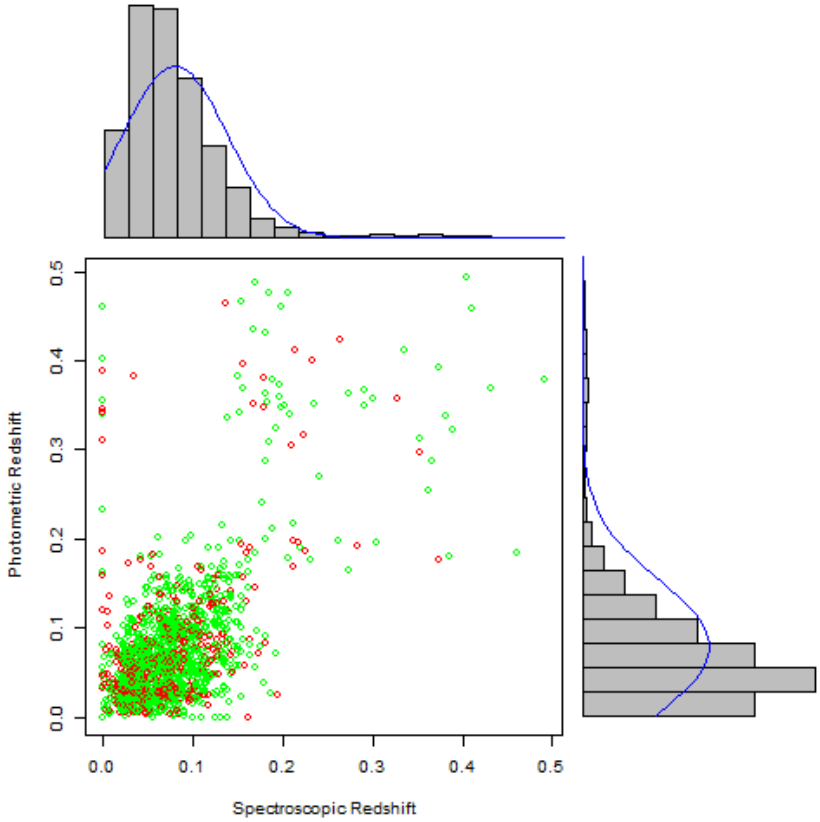


Figure 5.26: bnlearn Aracne Method Test of Model - Color and Morphology Photo-z - No. of Quantiles: 5 - Spec-z vs. Photo-z

Note: Tree-Augmented Naïve Bayes method was applied but it displayed high catastrophic error rates ($> 60\%$). Hence, it was not investigated in further detail.

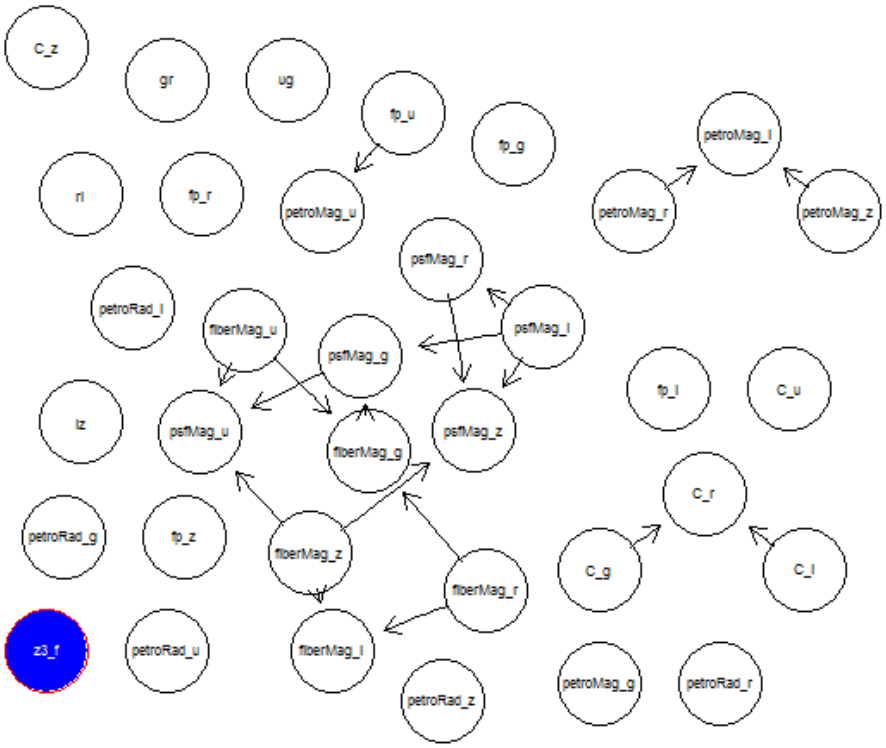


Figure 5.27: bnlearn GS Method - All Attributes Photo-z - No. of Quantiles: 5

5.4.4 All Attributes

Trying to apply the Bayesian learning methods on all the attributes -color, magnitude and morphology - did not return any promising result. The constraint-based methods - GS, IAMB and Fast IAMB - did not include redshift in the network, similar to the previous attempts in the different subsets of attributes. Refer to Figure 5.27, Figure 5.28 and Figure 5.29 for the network generated via application of GS, IAMB and Fast IAMB method respectively.The network generated using HC could not estimate for most of the test data points. Model generated via Tabu performed better than HC with high certainty for the estimated values but similar to HC it could not estimate for most of the test data. Figure 5.30 shows the predictive performance of Tabu method.

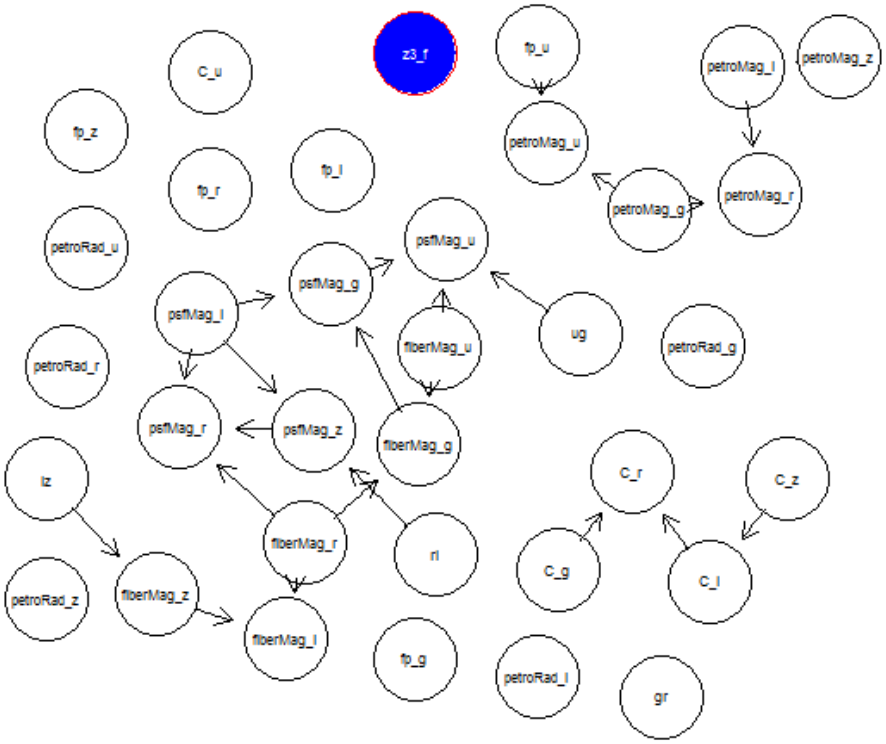


Figure 5.28: bnlearn IAMB Method - All Attributes Photo-z - No. of Quantiles: 5

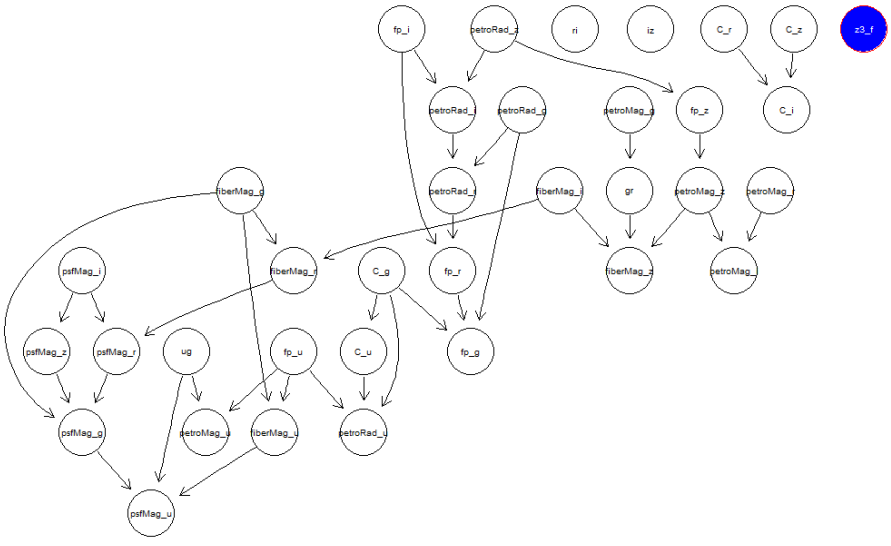


Figure 5.29: bnlearn GS Method - All Attributes Photo-z - No. of Quantiles: 5

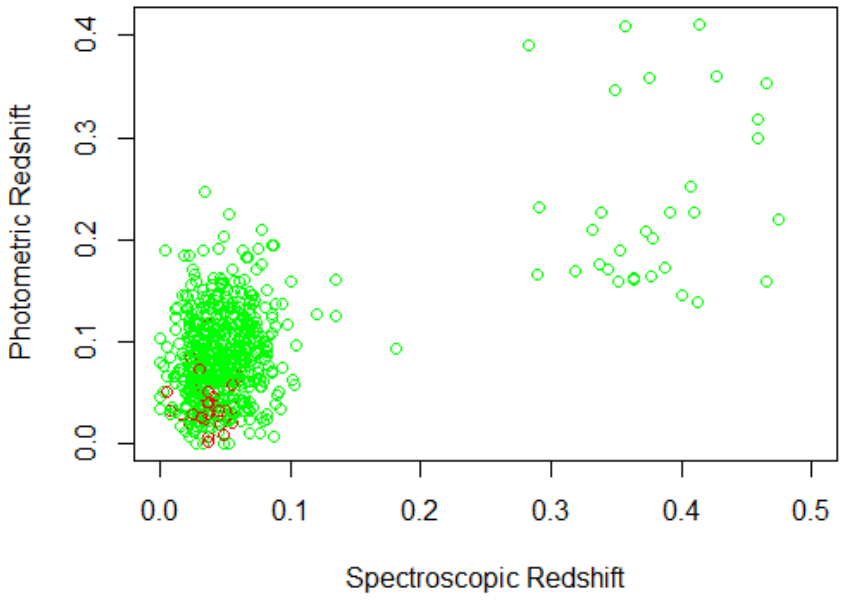


Figure 5.30: bnlearn GS Method - All Attributes Photo-z - No. of Quantiles: 5 - Spec-z vs. Photo-z

5.4.5 Highest Correlation Attribute Subset

Based on Pearson correlation of the attributes, following is the correlation with respect to redshift in increasing order: fp_g(-0.398), fp_r, fp_z, fp_i, petroR50_r, petroR50_g, petroR50_i, petroR50_z, petroR90_g, petroRad_r, petroR90_r, petroR90_i, petroR90_z, petroRad_i, petroRad_g, petroR90_u, fp_u, C_u, petroRad_z petroRad_u(-0.007), petroR50_u(0.077), C_g, iz, C_z, C_r, C_i, psfMag_z, psfMag_i, fiberMag_z, ug, psfMag_r, ri, fiberMag_i, petroMag_z, fiberMag_r, petroMag_i, petroMag_r, petroMag_u, psfMag_g, gr, psfMag_u, fiberMag_g, petroMag_g, fiberMag_u, rdshft(1.0)

The Minimum Spanning Tree (MST) tree based on correlation using Bayesia results in the tree as shown in Figure 5.31. Analyzing the tree leads to six attributes that redshift depends on and they are: fp_r, fp_u, petroMag_u. petroMag_g, fiberMag_g and fiberMag_u.

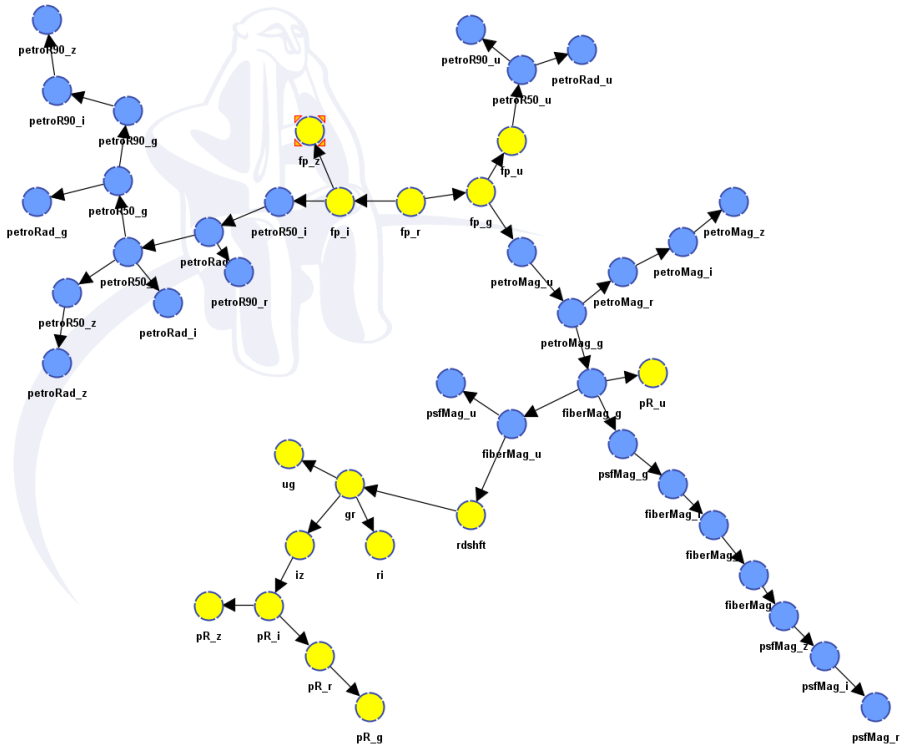


Figure 5.31: Bayesian MST

The catastrophic error rate of the Bayesian method using only these six attributes needs to be investigated further.

Depending on the cut of the data being used for training, the six high can result in ideal photo-z estimators. One instance using the training:test split of 90%:10% for two-digit z precision for training is as shown in Table 5.12.

Table 5.12: bnlearn HC Method - Color and Morphology Photo-z - No. of Quantiles: 5

Range	**Count**	**Mean z-Spec**	**Mean z-Photo**	**Mean Error**	**RMS**	**Catastrophic Error %**
(0-0.1]	26994	0.062	0.062	0	0	0
(0.1-0.2]	23591	0.14	0.14	0	0	0
(0.2-0.3]	3476	0.247	0.247	0	0	0
(0.3-0.4]	3260	0.348	0.348	0	0	0
(0.4-0.5]	1462	0.442	0.442	0	0	0

5.4.6 Discretization - Number of levels

In addition to the discretization method, the number of levels or breaks used to convert the continuous data to categorical data is crucial to the methods overall predictive performance. It was investigated how changing the number of breaks affects the catastrophic error percent of the estimated values. Using only the color attributes and spectroscopic redshift, the continuous-values were converted to categorical values using quantile breaks. It was found that for HC method, increasing the number of quantile breaks from five through twenty-five reduced the catastrophic error percent for all the ranges of z. Quantile break equal to ten appears to be the most optimal. Anything beyond ten quantile breaks, the performance improves for some of the regions but not as significant and immediately apparent as the improvement from five quantile breaks to ten quantile breaks, as shown in Figure 5.32. As the number of breaks increased, the proportion of test data with estimates certainty greater than 0.5 increased as shown in the spec-z vs. photo-z plots for quantile breaks = 6, 8, 10, 15, 20, 25. Refer Figure 5.34a to see the difference visually. Another way to look at this data as has been shown before is to look at the spread of the estimates in the different ranges. Figure 5.35a displays this desired information as boxplots of predictive performance across the different z ranges for quantile breaks = 6, 8, 10, 15, 20, 25. The networks generated for the different quantile breaks is as shown in Figure 5.36a.

On the contrary, the GS method deteriorates even further with increasing the quantile breaks as shown in Figure 5.33.

Similar analysis needs to be performed for these subsets of attributes: Color-Magnitude, Color-Morphology and six highest correlated attributes.

5.4.7 Redshift Precision

When two-digit precision of redshift was used in the training phase, it resulted in model with lesser catastrophic error percent as there was more elements to train each value of redshift in comparision to three-digit precision. For example, the number of records to train for z = 0.005, 0.006, ..., 0.014 in three-digit precision training is now all being used to

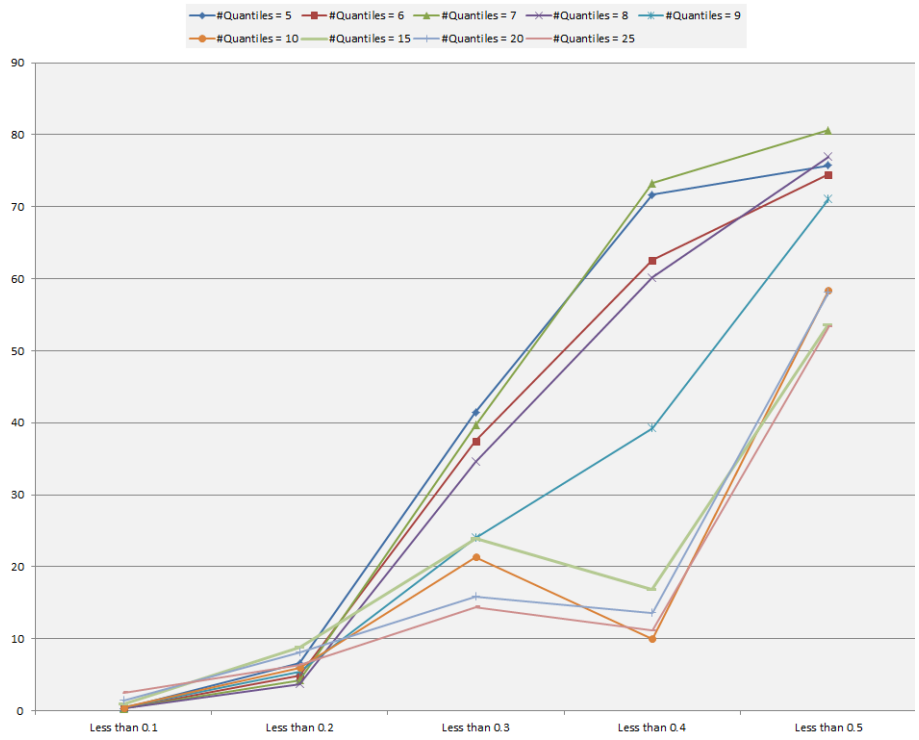


Figure 5.32: bnlearn HC Method - Color Only - Redshift Vs Catastrophic Error Percent By No. of Quantile Breaks

train z = 0.01. The comparative performance of BN learning using only color data is shown in Table 5.13 (same as Table 5.5) when three-digit precision is used versus Table 5.14 when two-digit precision is used. The performance for all the ranges is improved when two-digit precision redshift is used for training the model. With two-digit precision, if the number of quantiles is increased further, the performance improves even further as shown in Table 5.15.

5.4.8 Uniform Sampling

Due to the concentration of data in $z = 0 - 0.25$, the generated models using the different Bayesian learning methods result in the catastrophic error rate being high in the remaining sparsely populated ranges. It was investigated if uniform sampling from the different ranges will address this, if not completely, atleast to a certain extent. Alternative sophisticated

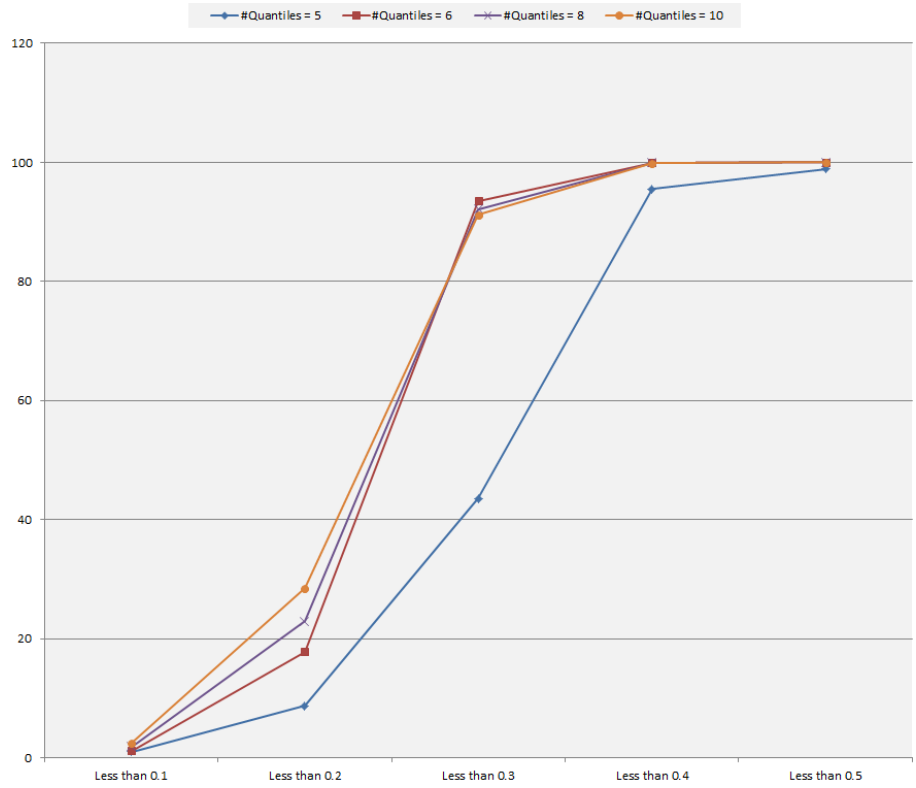
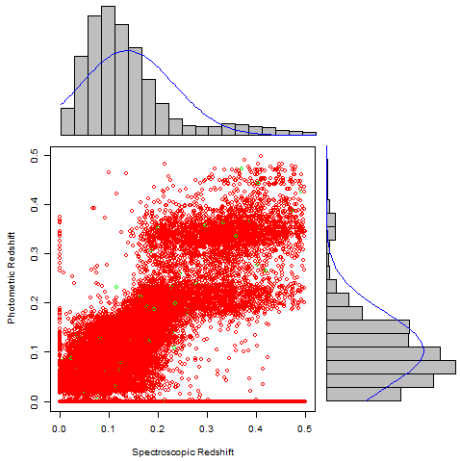


Figure 5.33: bnlearn GS Method - Color Only - Redshift Vs Catastrophic Error Percent By No. of Quantile Breaks

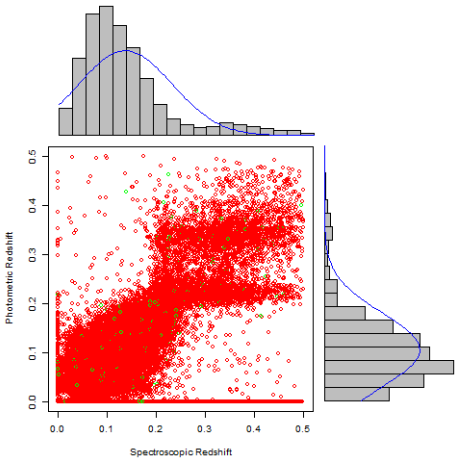
Table 5.13: bnlearn HC Method - Color Only Photo-z - Number of Quantiles = 5, z Precision = 3

Range	**Count**	**Mean z-Spec**	**Mean z-Photo**	**Mean Error**	**RMS**	**Catastrophic Error %**
(0-0.1]	31037	0.064	0.064	0.023	0.031	0.33
(0.1-0.2]	27951	0.14	0.114	0.038	0.054	6.61
(0.2-0.3]	4535	0.239	0.151	0.116	0.147	41.48
(0.3-0.4]	3738	0.348	0.156	0.202	0.242	71.64
(0.4-0.5]	1741	0.441	0.175	0.267	0.313	75.76

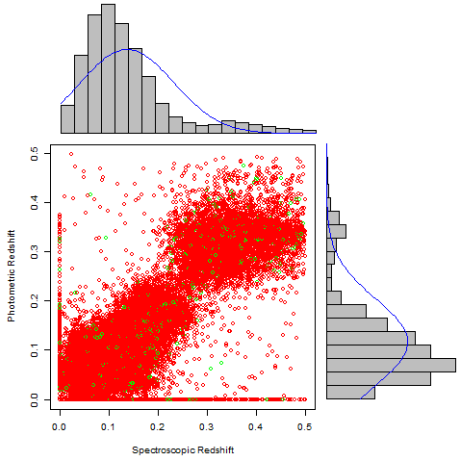
methods of sampling should be investigated to further improve the model generation process that is not biased towards regions with high data concentration while providing good predictive accuracy over all ranges.



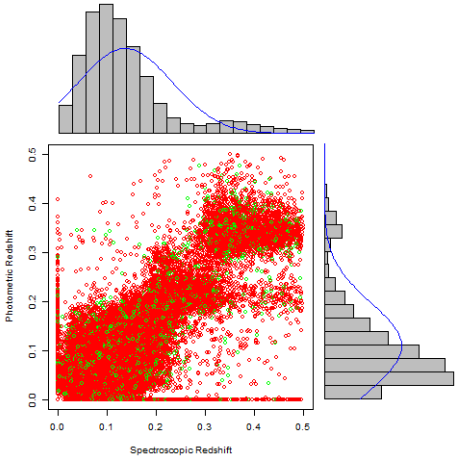
(a) No. of quantile breaks = 6



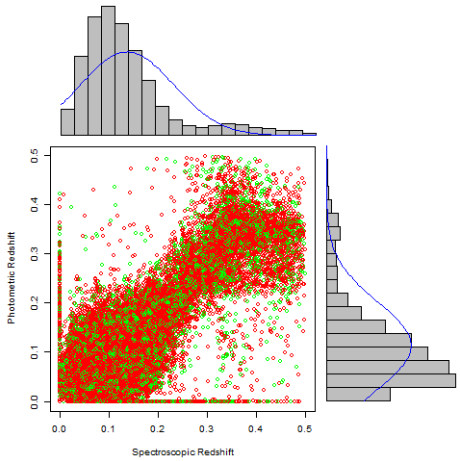
(b) No. of quantile breaks = 8



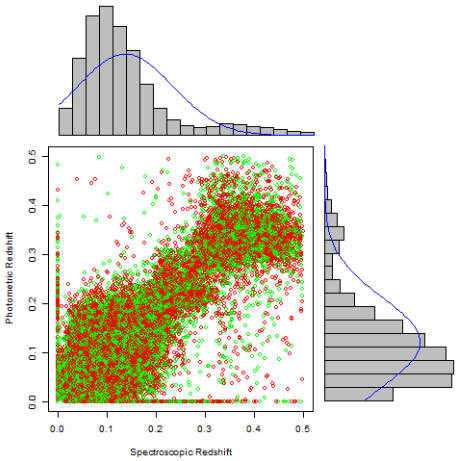
(c) No. of quantile breaks = 10



(d) No. of quantile breaks = 15

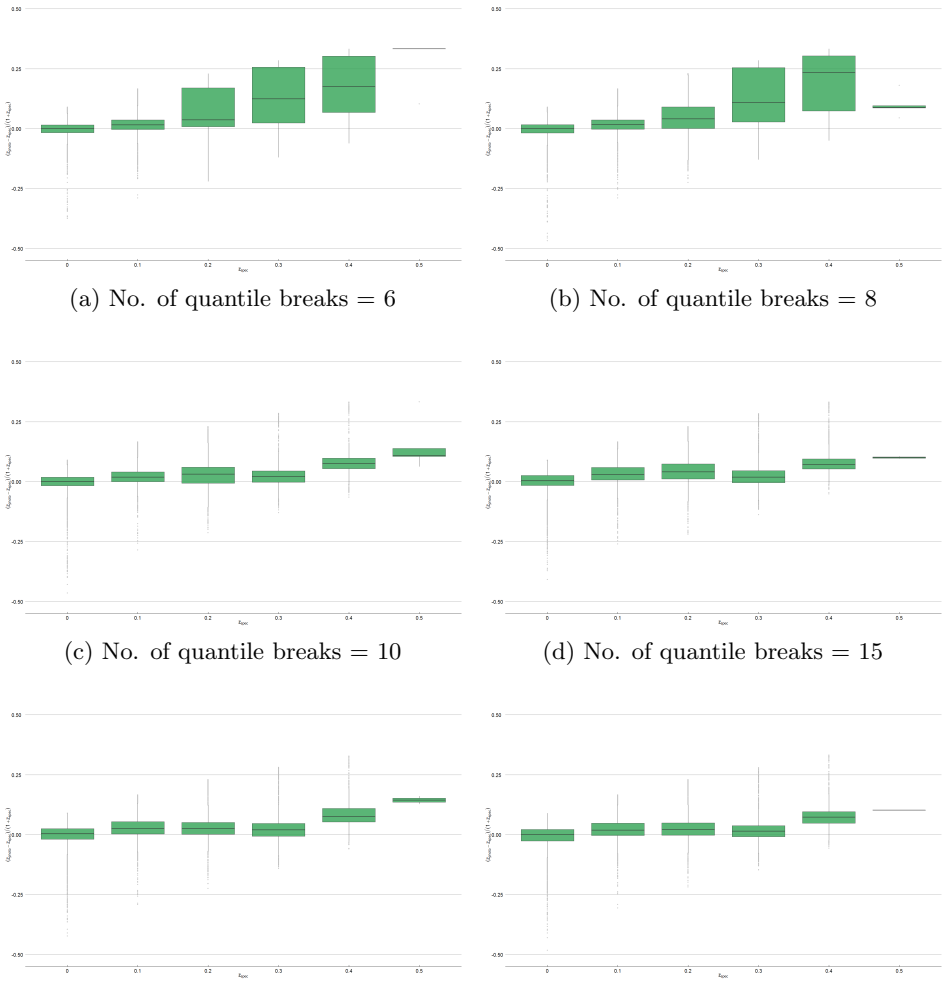


(e) No. of quantile breaks = 20



(f) No. of quantile breaks = 25

Figure 5.34: bnlearn HC Method - Color Only Photo-z - Impact of No. of Quantile Breaks



(a) No. of quantile breaks = 6

(b) No. of quantile breaks = 8

(c) No. of quantile breaks = 10

(d) No. of quantile breaks = 15

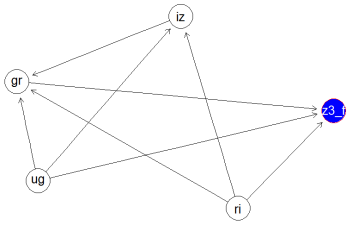
(e) No. of quantile breaks = 20

(f) No. of quantile breaks = 25

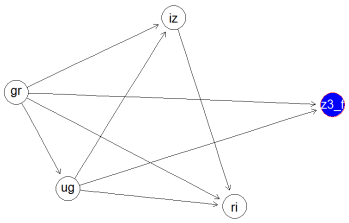
Figure 5.35: bnlearn HC Method - Color Only Photo-z - Impact of No. of Quantile Breaks on spread of estimates in the different z ranges

Color and Magnitude with 2000 points per possible z-value has performance as shown in Table 5.16.

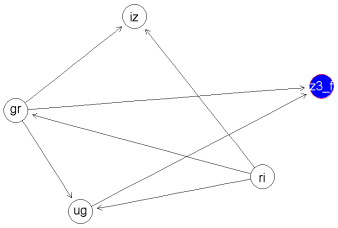
The six best attributes set from the MST analysis can result in good photo-z estimators. One instance using the training of 2000 points for each two-digit precision z value (total 10200 points) and test of 100 points for each two-digit precision z value (total 5100 points) is as shown in Table 5.17.



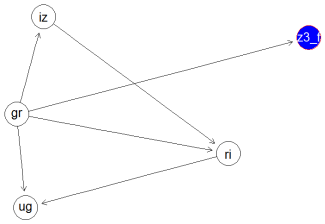
(a) No. of quantile breaks = 6



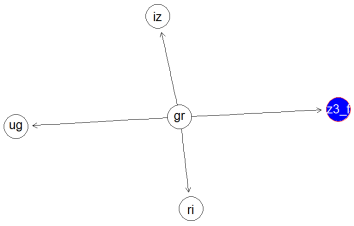
(b) No. of quantile breaks = 8, 9



(c) No. of quantile breaks = 10



(d) No. of quantile breaks = 15, 20



(e) No. of quantile breaks = 25

Figure 5.36: bnlearn HC Method - Color Only Photo-z - Generated Network for the different number of Quantile Breaks

Table 5.14: bnlearn HC Method - Color Only Photo-z - Number of Quantiles = 5, z Precision = 2

Range	**Count**	**Mean z-Spec**	**Mean z-Photo**	**Mean Error**	**RMS**	**Catastrophic Error %**
(0-0.1]	28663	0.061	0.063	0.021	0.03	0.31
(0.1-0.2]	26199	0.141	0.13	0.034	0.051	4.89
(0.2-0.3]	4332	0.246	0.237	0.06	0.073	15.35
(0.3-0.4]	3303	0.348	0.298	0.07	0.093	31.97
(0.4-0.5]	1549	0.443	0.336	0.11	0.131	43.45

Table 5.15: bnlearn HC Method - Color Only Photo-z - Number of Quantiles = 10, z Precision = 2

Range	**Count**	**Mean z-Spec**	**Mean z-Photo**	**Mean Error**	**RMS**	**Catastrophic Error %**
(0-0.1]	28594	0.062	0.066	0.02	0.028	0.38
(0.1-0.2]	25791	0.141	0.126	0.028	0.038	1.71
(0.2-0.3]	4198	0.246	0.23	0.049	0.066	7.65
(0.3-0.4]	3421	0.349	0.338	0.033	0.046	2.75

Table 5.16: bnlearn HC Method - Color and Magnitude Photo-z - Number of Quantiles = 10, z Precision = 2

Range	**Count**	**Mean z-Spec**	**Mean z-Photo**	**Mean Error**	**RMS**	**Catastrophic Error %**
(0-0.1]	532	0.046	0.047	0.026	0.036	0.94
(0.1-0.2]	274	0.152	0.149	0.037	0.05	5.11
(0.2-0.3]	731	0.261	0.273	0.043	0.063	8.62
(0.3-0.4]	887	0.35	0.383	0.064	0.079	11.53
(0.4-0.5]	902	0.45	0.443	0.045	0.057	6.87

Table 5.17: bnlearn HC Method - Color and Morphology Photo-z - No. of Quantiles: 5, 2000 points per value

Range	**Count**	**Mean z-Spec**	**Mean z-Photo**	**Mean Error**	**RMS**	**Catastrophic Error %**
(0-0.1]	716	0.049	0.066	0.04	0.058	6.84
(0.1-0.2]	781	0.151	0.142	0.047	0.06	7.3
(0.2-0.3]	967	0.256	0.263	0.069	0.092	24.3
(0.3-0.4]	899	0.35	0.385	0.062	0.076	20.13
(0.4-0.5]	900	0.45	0.385	0.078	0.093	32.89

Chapter 6: Future Work

6.1 Ensemble Methods

One major issue with most learning algorithms is that good performance on the training data does not necessarily lead to good generalization performance. This has lead to stacking (Netflix community terminology blending; runner-up algorithm Ensemble is the term used more often these days) which is the process of building a variety of different models and using a meta-learning model to combine the multiple model outputs. Nearest Neighbor, SVM, Restricted Boltzmann Machines are most commonly used models. This can be considered akin to the "audience vote" in the "Who Wants to be a Millionaire?" game show - an ensemble of decision makers.

Stacked generalization works by deducing the biases of the generalizer(s) with respect to a provided learning set. This deduction proceeds by generalizing in a second space whose inputs are (for example) the guesses of the original generalizers when taught with part of the learning set and trying to guess the rest of it, and whose output is (for example) the correct guess. When used with multiple generalizers, stacked generalization can be seen as a more sophisticated version of cross-validation, exploiting a strategy more sophisticated than cross-validations crude winner-takes-all for combining the individual generalizers. [185] Ensemble methods can be categorized as shown below:

- *Classifier Selection* Using different subset of training data to get multiple learners from a single learning method - a divide and conquer approach.
  - Resample the training data or divide the dataspace into partitions based on certain condition (e.g., bagging, boosting, cross-validation): The instability of the base classifier, i.e., the property that small changes in the training set will

cause large changes in the learned classifier, is usually required to expect some diversity in the classifiers.

  - Use intersecting or mutually exclusive subset of the features for the different classifiers: If the features from different groups are not too correlated, the combined classifiers can be expected to have high diversity.

- *Classifier Fusion* Using different learning methods trained over the entire feature space

- *Injecting Randomness* Using different parameters for a given learning method (e.g., For example, in neural networks, the initial configuration of weights is chosen at random. If the algorithm is applied with the same training data but different initial weights, the resulting classifiers can be quite different.)

A key observation with ensemble methods is that it is not optimal to minimize the RMSE of the individual predictors. Only the RMSE of the ensemble counts. Thus the predictors which achieve the best blending results are the ones, which have the right balance between being uncorrelated to the rest of the ensemble and achieving a low RMSE individually. An ideal solution would be to train all models in parallel and treat the ensemble as one big model. The big problem is that training 100+ models in parallel and tuning all parameters simultaneously is computationally not feasible. [186]

Ensemble methods leads to increased computation but less overall risk of making a poor decision and thereby, increased confidence on the result. Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. For this ensemble of classification trees, an upper bound of the ensemble error depends on the average pairwise correlation between members of the ensemble [187] There is experimental evidence that ensembles could be more accurate than individual classifiers when the predictions of their members share a low level of dependence, or at least reflect some level of diversity. This concept of diversity is generally thought as the ability of the classifiers to make different errors on new data points. [188] Effectively however, neither the combined performance nor

its improvement against mean classifier performance seem to be measurable in a consistent and well defined manner. Thus, the most successful diversity measure, barely regarded as measuring diversity, is the classification accuracy of an ensemble. It is both cheaper and more accurate to the other diversity measures. [189]

6.2 Deep Learning

The key requirement of machine learning is that the representations are learned and not manually entered. Neural networks had a wave of excitement in the 1980s as they could create their own internal representations. It fell out of fashion in the 1990s primarily due to the following three reasons [190]:

- Not enough labeled data
- Not fast enough computers
- Weights of the back-propagation were not initialized correctly

They made a comeback in 2006 with advances in the computational power of processors and in the learning techniques for training neural networks with many layers of representation. This "deep learning" (aka large number of layers of neural networks) improved the state-of-the-art in speech recognition and object recognition.

As soon as you end up extracting lots of layers of features and combining them in a way that is non-linear, then you're already into something like backpropagation through a deep net. One of the most important lessons is that if you can find a good objective function and if you can compute its gradient efficiently, you can get a long way. Bigger data sets and much faster computers have taken us into the regime where neural nets can really win.[191] Many layered feed-forward neural network pre-trained one layer at a time, treating each layer as a unsupervised restricted Boltzmann machine and fine-tuned using supervised back-propagation.

Deep learning techniques are machine learning methods that involve at least three, adaptive nonlinear processing steps from the input to the output. Deep models that learn many

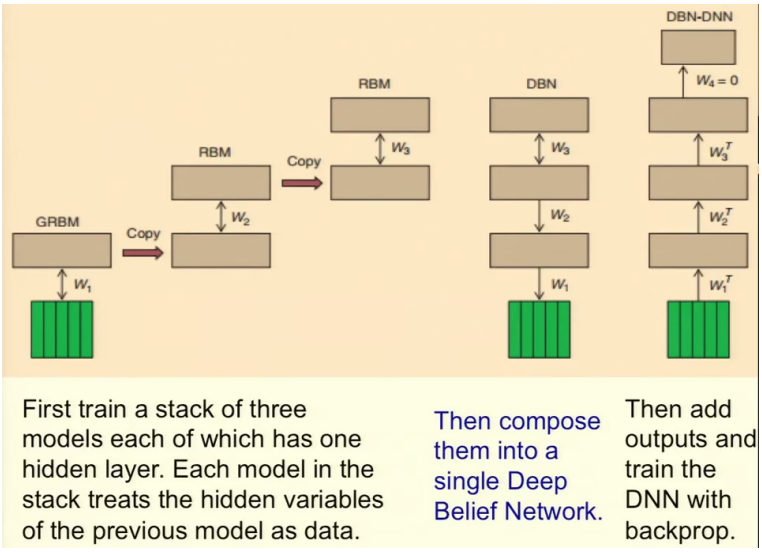


Figure 6.1: Deep Learning Network [194]

layers of features can potentially extract much better information from the input signal for the task in hand (e.g., classification or synthesis), through many layers of nonlinear evidence combination; as shown in Figure 6.1. Each layer helps achieve a new and better lower bound on the log probability of the training data. [190] [192] In other words, Deep Learning addresses the problem of learning hierarchical representations with a single algorithm. It is inspired from how the mammalian visual cortex works - simple cells detect local features while complex cells "pool" the outputs of simple cells within a retinotopic neighborhood. [193]

The first big success for deep neural nets was the improvement seen in a problem in speech recognition domain. It involved looking at multiple frames of coefficients and predicting the states of Hidden Markov Models (HMMs) that model phonemes (e.g., /k/, which occurs in words such as cat, kit, school, skill). This performed better than the Gaussian mixture "shallow" models (GMMs) used for the previous 30 years. GMMs have only one layer of latent variables; thus, lack multiple layers of adaptive nonlinear features. [1] The

[1]Research was focused on finding better ways of estimating the GMM parameters so that error rates are decreased or the margin between different classes is increased. Similarly, in the field of natural language processing (NLP), maximum entropy (MaxEnt) models and conditional random fields (CRFs) were popular

next big success was object recognition. This is being used by Google and Microsoft for image and voice search.

Averaging many models is the idea behind ensemble methods used most often in solutions for machine learning competitions. Averaging many decision trees is called random forests. Individual trees can be trained differently using different training sets. Avergaing multiple deep neural nets is hard since each net takes a long time to learn and it is not efficient at test time. We can combine models by taking the arithmetic mean or geometric mean (nth root of the product of n numbers; renormalize since the sum of geometric means is not equal to one) of their output probabilities. Dropout is an efficient way to perform geometric mean in deep neural nets. [197] Dropout requires that each time a training example is presented, each hidden unit of the hidden layer is omitted with probability 0.5. Thus, all the architectures share equal weight via this random sampling from 2^H different architectures. It helps avoid overfitting.

Thus, deep learning consists of essentially two main steps executed in an iterative manner in order to generate the successive layers of learning:

- Unsupervised Pre-training to determine the structure Neural Network
- Supervised Back-Propagation to classify labeled data

6.2.1 Deep Belief Network (DBN)

Boltzmann machines are one of the first examples of a neural network capable of learning internal representations, and are able to represent and (given sufficient time) solve difficult combinatoric problems. They are theoretically intriguing because of the locality and Hebbian [2] nature of their training algorithm, and because of their parallelism and the

for the last decade. Hierarchical (or stacked) HMMs or CRFs and multi-level detection-based systems are deep models.[195] However, they are statistically inefficient for modeling data that lie on or near a nonlinear manifold in the data space. For example, modeling the set of points that lie very close to the surface of a sphere only requires a few parameters using an appropriate model class, but it requires a very large number of diagonal Gaussians or a fairly large number of full-covariance Gaussians. [196]

[2]Hebbian theory is a theory in neuroscience that proposes an explanation for the adaptation of neurons in the brain during the learning process.

resemblance of their dynamics to simple physical processes. Due to a number of issues discussed below, Boltzmann machines with unconstrained connectivity have not proven useful for practical problems in machine learning or inference, but if the connectivity is properly constrained, the learning can be made efficient enough to be useful for practical problems. Although learning is impractical in general Boltzmann machines, it can be made quite efficient in an architecture called the Restricted Boltzmann Machine (RBM).

RBM does not allow intralayer connections between hidden units. After training one RBM, the activities of its hidden units can be treated as data for training a higher-level RBM. This method of stacking RBM's makes it possible to train many layers of hidden units efficiently and is one of the most common deep learning strategies. As a new layer is added to improve the prior on the previous layer, the overall generative model gets better. Deep Belief Network (DBN) is a greedy, layer-by-layer unsupervised learning algorithm that consists of learning a stack of RBMs one layer at a time. The whole stack can be viewed as a single probabilistic model. The top two layers form a restricted Boltzmann machine, but the lower layers form a directed sigmoid belief network (*aka* Bayesian network).

Deep Boltzmann machines (DBM) are interesting for several reasons. First, like DBN, DBMs have the ability to learn internal representations that capture very complex statistical structure in the higher layers. As has already been demonstrated for DBNs, this is a promising way of solving object and speech recognition problems. High-level representations can be built from a large supply of unlabeled data, and a much smaller supply of labeled data can then be used to fine-tune the model for a specific discrimination task. Second, again like DBNs, if DBMs are learned in the right way, there is a very fast way to initialize the states of the units in all layers by simply doing a single bottom-up pass using twice the weights to compensate for the initial lack of top-down feedback. Third, unlike DBNs and many other models with deep architectures, the approximate inference procedure, after the initial bottom-up pass, can incorporate topdown feedback. This allows DBMs to use higher-level knowledge to resolve uncertainty about intermediate-level features, thus creating better data-dependent representations and better data-dependent statistics for learning. [198]

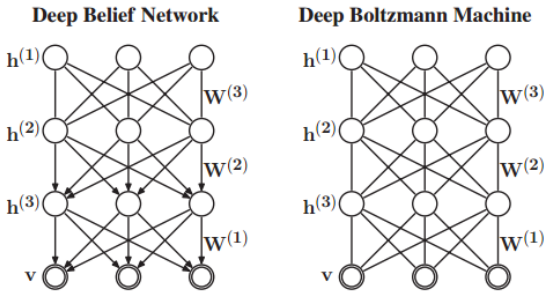


Figure 2: (Left) Deep belief network (DBN). The top two layers form an undirected graph, and the remaining layers form a belief net with directed, top-down connections (Right) Deep Boltzmann machine (DBM), with both visible-to-hidden and hidden-to-hidden connections but no within-layer connections. All the connections in a DBM are undirected.

Figure 6.2: Deep Belief Network (DBN) vs. Deep Boltzmann Machine (DBM) [198]

Refer Figure 6.2 to note the essential difference between DBN and DBM.

[198] present a new learning algorithm for Boltzmann machines that contain many layers of hidden variables. Data-dependent statistics are estimated using a variational approximation that tends to focus on a single mode, and data-independent statistics are estimated using persistent Markov chains. The use of two quite different techniques for estimating the two types of statistic that enter into the gradient of the log likelihood makes it practical to learn Boltzmann machines with multiple hidden layers and millions of parameters. The learning can be made more efficient by using a layer-by-layer pretraining phase that initializes the weights sensibly. The pretraining also allows the variational inference to be initialized sensibly with a single bottom-up pass.

[199] addresses deep neural networks in which the output of each node is a quadratic function of its inputs. Similar to other deep architectures, these networks can compactly represent any function on a finite training set. Basis Learner is an efficient layer-by-layer algorithm for training such networks. It is compared against kernel learning.

The redshift estimation should be attempted as estimation using different layers. For example, the first layer could be based on color attributes. The next layer could use the photometric redshift estimate from the color attributes layer, combine with magnitude or morphology attributes. Analysis needs to be performed to determine the order of the

attributes in the different layers. The metrics need to be designed to better understand the characteristics of an attribute that can best position it among the different layers.

6.3 Calibrating Photo-z in absence of Spectro-z

Faint objects cannot be spectroscopically confirmed for their redshift. Thus, alternative approaches need to be designed that can serve the equivalent of spectroscopic measurement of redshift. Galaxy merging [3] is a process that lasts several billion years rather than a short-lived event. A galaxy merger is a pair of galaxies which are gravitationally bound and whose orbits will dynamically decay such that their nuclei will merge within x billion years, where x is typically 13 Gyr for major mergers with mass ratios greater than 1:3. [200] [201] talks about using tidal pairs (close galaxy pairs with merging features) and using the difference of the photometric redshift of the two galaxies as a measure of redshift precision. A study using 69 isolated tidal pairs is shown in Figure 6.3.

(Kovac et al., 2010) [202] studies how the photo-z probability density is modified by using the local density as a constraint, reducing the photo-z errors to within the scale of the smoothing kernel used to probe the density field ($\delta z \sim 0.05$).

Gamma-Ray Bursts (GRB) are the most energetic events in the Universe, and provide a complementary probe of dark energy by allowing the measurement of cosmic expansion history that extends to $z > 6$. The current GRB data is summarized by a set of model-independent distance measurements, with negligible loss of information in (Wang, 2008) [203]. This formulates five calibration relations for GRBs that relate its luminosity or the total burst energy in the gamma rays to observables of the light curves and/or spectra: time lag, variability, peak of the spectrum, minimum rise time.

(Wang, 2007) [204] derives a simple empirical photometric redshift estimator for Type Ia supernovae (SNe Ia) using a training set of SNe Ia with multiband (griz) light-curves and

[3] A key obstacle to understanding the galaxy merger rate and its role in galaxy evolution is the difficulty in constraining the merger properties and time-scales from instantaneous snapshots of the real Universe. The most common way to identify galaxy mergers is by morphology.

Redshift Calibration with Tidal Pairs

How can we test LSST photometric redshifts without spectroscopy?

Close galaxy pairs with strong tidal features and no other companions have a high likelihood of physical association;
⇒ Δz_{phot} of each pair is a measure of redshift precision (Quadri+ 2009)

+ Biased towards challenging z_{phot} galaxies (star-forming, blue, dusty)
+ Tails detectable in full-depth LSST images out to z~1
+ Hundreds of pairs per LSST pointing

− Careful selection required to exclude false pairs
− Blending of very close pairs may give bad photometry, z_{phot}

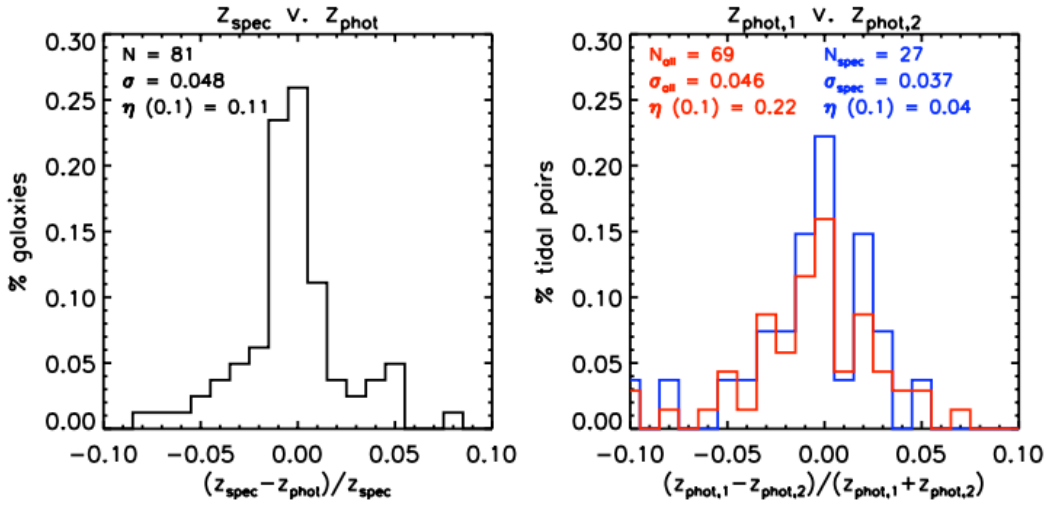


Fig. 6. 69 isolated tidal pairs (138 galaxies) at z<1 were selected from CFHTLS* Deep *i* images (Gwyn 2009) and matched to public z_{phot} (Ilbert+ 2006, 2009) and z_{spec} (Davis+ 2003, Lilly+2007) catalogs. ***Left:* Z_s v. Z_p for Tidal Galaxies**. 81/138 have spectroscopic redshifts; $\sigma[(z_p-z_s)/z_s)] = 0.048$, and catastrophic error rate $\eta = 11\%$ (where$|(z_p-z_s)/z_s| \geq 0.1$). ***Right:* Z_{p1} v. Z_{p2} for Tidal Pairs**. The distribution of $(z_{p1}-z_{p2})/(z_{p1}+z_{p2})$ for all 69 tidal pairs (red), and for 27 spectroscopically-confirmed tidal pairs (blue). $\sigma[(z_{p1}-z_{p2})/(z_{p1}+z_{p2})] \sim \sigma[(z_p-z_s)/z_s)) \sim 0.05$, and is similar for both spectroscopically-confirmed pairs and full sample at 0.1 and -0.1. However, η is higher (22%) for full sample due to false pair contamination.

* Based on observations obtained with MegaPrime/MegaCam, a joint project of CFHT and CEA/DAPNIA, at the Canada-France-Hawaii Telescope (CFHT) which is operated by the National Research Council (NRC) of Canada, the Institut National des Science de l'Univers of the Centre National de la Recherche Scientifique (CNRS) of France, and the University of Hawaii. This work is based in part on data products produced at the Canadian Astronomy Data Centre as part of the Canada-France-Hawaii Telescope Legacy Survey, a collaborative project of NRC and CNRS.

Figure 6.3: Redshift Calibration using Tidal Pairs

spectroscopic redshifts obtained by the Supernova Legacy Survey (SNLS). This estimator is analytical and model-independent regression ($z = c_1 + c_2(U-B) + c_3(B-V) + C_4(V-I)$, where c_1, c_2, c_3, c_4 are the estimated co-efficients); it does not use spectral templates.

Clustering redshift is an alternative approach to photometric redshift. It utilizes the positional information, not flux information, of objects. The idea is simple - cross-correlation between two galaxy samples yields a signal where they overlap in redshift and the clustering signal can thus be used as a redshift inference. One powerful application of this technique is to use a sample of spectroscopic redshifts, in which the redshift distribution is precisely known, as a reference. If one makes a narrow redshift bin with a spectroscopic sample and cross-correlate it with a photometric sample, the clustering signal is proportional to the number of photometric galaxies in that redshift bin. By shifting the spectroscopic redshift bin, one can in principle reconstruct a redshift distribution of the input photometric sample. However, there is one uncertainty here; the clustering signal is also proportional to the bias of the photometric sample. This can be a serious issue when the photometric sample covers a wide range of redshift or has multiple peaks in the redshift distribution, which may often be the case in real analysis. Clustering information can be used to estimate the accuracy with which photometric redshifts can be inferred and in particular characterize the fraction of catastrophic outliers. This direction of research has stayed at the level of a theoretical idea and has not led to the promised advances in redshift estimation. [88] [205]

6.4 Leveraging Big Data Technologies

Big data consists of expansive collections of data (large volumes) that are updated quickly and frequently (high velocity) and that exhibit a huge range of different formats and content (wide variety). There are challenges not just in Volume, but also in Variety and Velocity. Variety refers to heterogeneity of data types, representation, and semantic interpretation. Velocity denotes both the rate at which data arrive and the time frame in which they must be acted upon. [206] The analysis of Big Data is an iterative process, each with its

own challenges, that involves many distinct phases. As expected, the issues that ail data mining such as heterogeneity of data, inconsistency and incompleteness, timeliness, privacy, visualization and collaboration are also applicable for Big Data.

One challenge is to define the "on-line" filters in such a way they do not discard useful information, since the raw data is often too voluminous to even allow the option of storing it all. Data volume is increasing faster than CPU speeds and other compute resources. Due to power constraints, clock speeds have largely stalled and processors are being built with increasing numbers of cores. In short, one has to deal with parallelism within a single node. Unfortunately, parallel data processing techniques that were applied in the past for processing data across nodes do not directly apply for intranode parallelism, since the architecture looks very different. For example, there are many more hardware resources such as processor caches and processor memory channels that are shared across cores in a single node. [1].

Another dramatic shift under way is the move toward cloud computing, which now aggregates multiple disparate workloads with varying performance goals into very large clusters. This level of sharing of resources on expensive and large clusters stresses grid and cluster computing techniques from the past, and requires new ways of determining how to run and execute data processing jobs so we can meet the goals of each workload cost-effectively, and to deal with system failures, which occur more frequently as we operate on larger and larger systems.[1].

Governments deal not only with general issues of big-data integration from multiple sources and in different formats and cost but also with some special challenges. The biggest is collecting data; governments have difficulty, as the data not only comes from multiple channels (such as social networks, the Web, and crowdsourcing) but from different sources (such as countries, institutions, agencies, and departments). Sharing data and information between countries is a special challenge. sharing information across national boundaries involves language translation and interpretation of text semantics (meaning of content) and sentiment (emotional content) so the true meaning is not lost. Decision making in

government usually takes much longer and is conducted through consultation and mutual consent of a large number of diverse actors, including officials, interest groups, and ordinary citizens. Many well-defined steps are therefore required to reduce risk and increase the efficiency and effectiveness of government decision making. Data sharing within a country among different government departments and agencies is another challenge. The "tower of Babel" in which each system keeps its data isolated from other systems complicates trying to integrate complementary data among government agencies and departments. [6]

Computing in large-scale systems is shifting away from the traditional compute-centric model successfully used for many decades into one that is much more data-centric. This transition is driven by the evolving nature of what computing comprises, no longer dominated by the execution of arithmetic and logic calculations but instead becoming dominated by large data volume and the cost of moving data to the locations where computations are performed. Data movement impacts performance, power efficiency and reliability, three fundamental components of a system. These trends are leading to changes in the computing paradigm, in particular the notion of moving computation to the data in a so-called Near-Data Processing approach, which seeks to perform computations in the most appropriate location depending on where data resides and what needs to be extracted from that data. Examples already exist in systems that perform some computations closer to disk storage, leveraging the data streaming coming from the disks, filtering the data so that only useful items are transferred for processing in other parts of the system. Conceptually, the same principle can be applied throughout a system, by placing computing resources close to where data is located, and decomposing applications so that they can leverage such a distributed and potentially heterogeneous computing infrastructure.

While Volume, Variety and Velocity are important, additional important requirements such as Veracity, Privacy and Usability still remain.[1]

Citizen science is growing: the Cornell Lab for Ornithology's eBird project and Galaxy Zoo in astronomy and are but two examples, each involving tens of thousands if not hundreds of thousands of people who have never been socialized into research work. Such people

may make unconventional demands if they feel they are not properly compensated for their important efforts. Such power conflicts do not arise from open networked environments, per se, but from the new opportunities enabled by such environments in circumstances of constrained resources. During much of the 20th century the U.S. research enterprise capitalized on the benefits of scientific agriculture, advanced the industrial revolution, improved human health, and helped achieve victory in conflicts such as World War II and the Cold War. Knowledge discovery was a public good, and more was better. Now, politicians and policymakers acknowledge the value of scientific knowledge discovery, but at the same time ask how much is needed, at what price, paid for by whom, and benefiting whom? Scientific knowledge discovery has become important. Important things become political. [207]

Beyond technological innovations that make it possible to accumulate and process massive amounts of data ever more cost-effectively, the other key concept here is a competitive mandate that businesses continuously improve their decision-making capabilities in order to survive. The consistent, systematic analysis of complex data for decision making enables a company to operate more intelligently at all levels. In particular, the emphasis upon strategic business analytics in recent years has elevated executive expectations and helped to transform the business analytics ideal into a significant competitive force. The application of business analytics methods leads to improvement in an organization's overall decision-making capacity, which enhances its ability to conduct its business intelligently. So, the desire (and accelerating need) to achieve a higher level of organizational intelligence is a prime driver for implementing business analytics. [206]

Making sense of big data requires more, and with our increasing inundation with data comes new and creative opportunities to build unique interfaces. [208] Intelligent Service Machine (ISM) then refers to an intelligent design of the service machine featuring the embodied cognition of co-production in terms of modeling and automating the cognitive process and knowledge representations as required. The causal technical and social elements that shape the interactions of all relevant factors and actors and influence the trajectory of technological and social outcomes, and these salient elements include goals, problem solving

strategy, solution requirements, theories, tacit knowledge, and design methods [209]

The Big Data phenomenon presents opportunities and perils. On the optimistic side of the coin, massive data may amplify the inferential power of algorithms that have been shown to be successful on modest-sized data sets. The challenge is to develop the theoretical principles needed to scale inference and learning algorithms to massive, even arbitrary, scale. On the pessimistic side of the coin, massive data may amplify the error rates that are part and parcel of any inferential algorithm. The challenge is to control such errors even in the face of the heterogeneity and uncontrolled sampling processes underlying many massive data sets. Another major issue is that Big Data problems often come with time constraints, where a high-quality answer that is obtained slowly can be less useful than a medium-quality answer that is obtained quickly. Overall we have a problem in which the classical resources of the theory of computatione.g., time, space and energytrade off in complex ways with the data resource. [210]

There is an increased and emerging need for robust and scalable algorithms and tools to analyze and mine these tera- and peta-scale data to determine patterns and trends. If the algorithm is distributed in nature, it will not involve the vast communication and data transfer needs that are usually associated with analyzing these types of large data collections. Everyone is being overwhelmed by data, and the promise of simplification becomes really attractive. This simplification today primarily comes from three advances in technology: the fact that storage of data via the cloud, GPU-driven calculations and software tools like Apache Hadoop that have simplified the processing of large-scale datasets (aka "Big Data") [208] using MapReduce algorithm [211]

Big Data is "big" in two different senses - the quantity and variety of data to be processed as well as the scale of analysis (termed analytics) that can be applied to those data to make inferences and draw conclusions. Data fusion occurs when data from different sources are brought into contact and new facts emerge. Individually, each data source may have a specific, limited purpose. Their combination, however, may uncover new meanings. More broadly, data analytics discovers patterns and correlations in large corpuses of data, using

increasingly powerful statistical algorithms. If those data include personal data, privacy is a concern. We can safely ignore this concern in our present study based on astronomy data. Data mining, sometimes loosely equated to analytics but actually only a subset of it, refers to a computational process that discovers patterns in large data sets. It is a convergence of many fields of academic research in both applied mathematics and computer science, including statistics, databases, artificial intelligence, and machine learning. [2] Thus, the studies using the data from the multiple sky surveys are Big Data in nature and require novel data manipulation and analysis techniques for maximum leverage and extraction of the yet-to-be-understood knowledge in it.

6.5 DTW and SAX application

While many symbolic representations of time series have been introduced over the past decades, they all suffer from two fatal flaws. Firstly, the dimensionality of the symbolic representation is the same as the original data, and virtually all data mining algorithms scale poorly with dimensionality. Secondly, although distance measures can be defined on the symbolic approaches, these distance measures have little correlation with distance measures defined on the original time series. SAX is a symbolic representation of time series. The utility of SAX representation on various data mining tasks of clustering, classification, query by content, anomaly detection, motif discovery, and visualization is mentioned in [212] [213]. SAX builds on piecewise constant modeling technique, Piecewise Aggregate Approximation (PAA), and symbolizes the PAA representation into a discrete string.

The way PAA works is to reduce the time series from n dimensions to w dimensions, the data is divided into w equal sized frames. Prior to this, normalize each time series to have a mean of zero and a standard deviation of one before converting it to the PAA. The mean value of the data falling within a frame is calculated and a vector of these values becomes the data-reduced representation. The PAA dimensionality reduction is intuitive and simple, yet has been shown to rival more sophisticated dimensionality reduction techniques like Fourier transforms and wavelets.

Given that the normalized time series have highly Gaussian distribution, we can simply determine the breakpoints that will produce a equal-sized areas under Gaussian curve. Breakpoints are a sorted list of numbers $B = \beta_1, \cdots, \beta_{a-1}$ such that the area under a N(0,1) Gaussian curve from β_i to $\beta_{i+1} = \frac{1}{a}$ (β_0 and β_a are defined as $-\infty$ and ∞, respectively). These breakpoints may be determined by looking them up in a statistical table. For example, when a = 3, β_1 = qnorm(1/3) = -0.43, β_2 = qnorm(2/3) = 0.43. Similarly, when a = 4, β_1 = qnorm(1/4) = -0.67, β_2 = qnorm(2/4) = 0, β_3 = qnorm(3/4) = 0.67.

All PAA coefficients that are below the smallest breakpoint are mapped to the symbol a, all coefficients greater than or equal to the smallest breakpoint and less than the second smallest breakpoint are mapped to the symbol b, etc. The concatenation of symbols that represent a subsequence is called a word and is termed a SAX representation.

Distance measure of the words is defined as a MINDIST function that returns the minimum distance between the original time series of two symbolic representation $\hat{Q} = \hat{q_1} \cdots \hat{q_w}, \hat{C} = \hat{c_1} \cdots \hat{c_w}$ where w is the number of PAA segments representing time series, is as shown in Equation 6.1. The dist() function can be implemented using a table lookup where the value in cell (r,c) for any lookup table can be calculated by the Equation 6.2.

$$Mindist(\hat{Q}, \hat{C}) = \sqrt{\frac{n}{w}} \sqrt{\sum_{i=1}^{w} (dist(\hat{q}, \hat{c}))^2} \tag{6.1}$$

$$\begin{cases} 0, if |r - c| \leq 1 \\ \beta_{max(r,c)-1} - \beta_{min(r,c)}, if otherwise \end{cases} \tag{6.2}$$

Its discrete nature enables emerging tasks such as anomaly detection and motif discovery. It may be possible to create a lower bounding approximation of Dynamic Time Warping, by slightly modifying the classic string edit distance. Finally, there may be utility in extending our work to multidimensional time series.

The Jaccard index, also known as the Jaccard similarity coefficient (originally coined

coefficient de communaut by Paul Jaccard), is a statistic used for comparing the similarity and diversity of sample sets.

6.6 Quantum Machine Learning

A quantum bit or qubit is a two-state quantum-mechanical system where the two states are vertical polarization and horizontal polarization. In a classical system, a bit would have to be in one state or the other. A qubit is a unit of quantum information.

The two states in which a qubit may be measured are known as basis states (or basis vectors). As is the tradition with any sort of quantum states, they are represented by Diracor "braket"notation. This means that the two computational basis states are conventionally written as $|0\rangle$ and $|1\rangle$ (pronounced "ket 0" and "ket 1").

A pure qubit state is a linear superposition of the basis states. This means that the qubit can be represented as a linear combination of $|0\rangle$ and $|1\rangle$: $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ where α and β are probability amplitudes of classical states $|0\rangle$ and $|1\rangle$ and can in general both be complex numbers.

When we measure this qubit in the standard basis, the probability of outcome $|0\rangle$ is $|\alpha|^2$ and the probability of outcome $|1\rangle$ is $|\beta|^2$. Because the absolute squares of the amplitudes equate to probabilities, it follows that α and β must be constrained by the equation $|\alpha|^2 + |\beta|^2 = 1$.

A true quantum computer could encode information in so-called qubits that can be 0 and 1 at the same time. Doing so could reduce the time required to solve a difficult problem that would otherwise take several years of computation to mere seconds. However, such a device would be highly sensitive to outside interference. Quantum computers will be best suited to very specific tasks, most notably to simulate quantum mechanical systems or to factor large numbers to break codes in classical cryptography. Yet there is one way that quantum computing might be able to assist big data: by searching very large, unsorted data sets for matching patterns.

Quantum RAM (Q-RAM) has been prototyped with an accompanying program Q-App (pronounced quapp) targeted to machine learning. The data is not accessed, only the common features are. Quantum computing is expected to work well for powerhouse machine-learning algorithms capable of identifying patterns in huge data sets. [214]

Quantum Machines and Machine Learning both involve manipulation of vectors and vector spaces. That implies that certain problems involving Machine Learning can take advantage of Quantum Machines-related processing. [215] Quantum Support Vector Machines, Quantum Clustering and k-nearest neighbor methods and Quantum Neural Networks are some of the proposed and ongoing work on machine learning algorithm in the quantum computing domain.

Chapter 7: Summary Conclusions

Machine learning techniques require a representative training sample, which in practice means they do not go fainter than the spectroscopic limits. Additionally, another focus of the solution should be to take advantage of the likelihood associated with the different estimators and leverage it to provide a more robust estimate (in contrast to the loss of uncertainty information when combining point estimates from the individual estimators to provide a final estimate).

$g-r$ color consistently shows up a predictor of redshift. Figure 7.1 shows the correlation of redshift with $g-r$ in the different ranges and the reason for it being selected by multiple methods as the attribute most closely related with redshift.

Cosmology has undergone rapid changes over the past centuries as our understanding of the surrounding universe has evolved. [216] provides a brief history of this growth. A major factor in the recent growth of this understanding by leaps and bounds are the wide-deep surveys being undertaken by the powerful telescopes. Wide-angle surveys that sample a specific section of the sky are best for topological studies of the three-dimensional structure. The ability to deeply and rapidly image much of the sky - billions of galaxies and stars - has great impact in astrophysics. Large aperture aperture and wide field survey telescopes together with powerful data processing computational systems open up the whole universe for exploration. Theorists, observers, and computational scientists need to work together to to develop algorithms that maximize the scientific returns of such programs. This requires us to address the inherent technical challenges in data management and automated discovery.

In cosmology, baryon acoustic oscillations (BAO) refers to regular, periodic fluctuations in the density of the visible baryonic matter of the Universe. Nearly all matter that may be encountered or experienced in everyday life is baryonic matter, which includes atoms of any sort, and provides those with the quality of mass. Non-baryonic matter, as implied by the

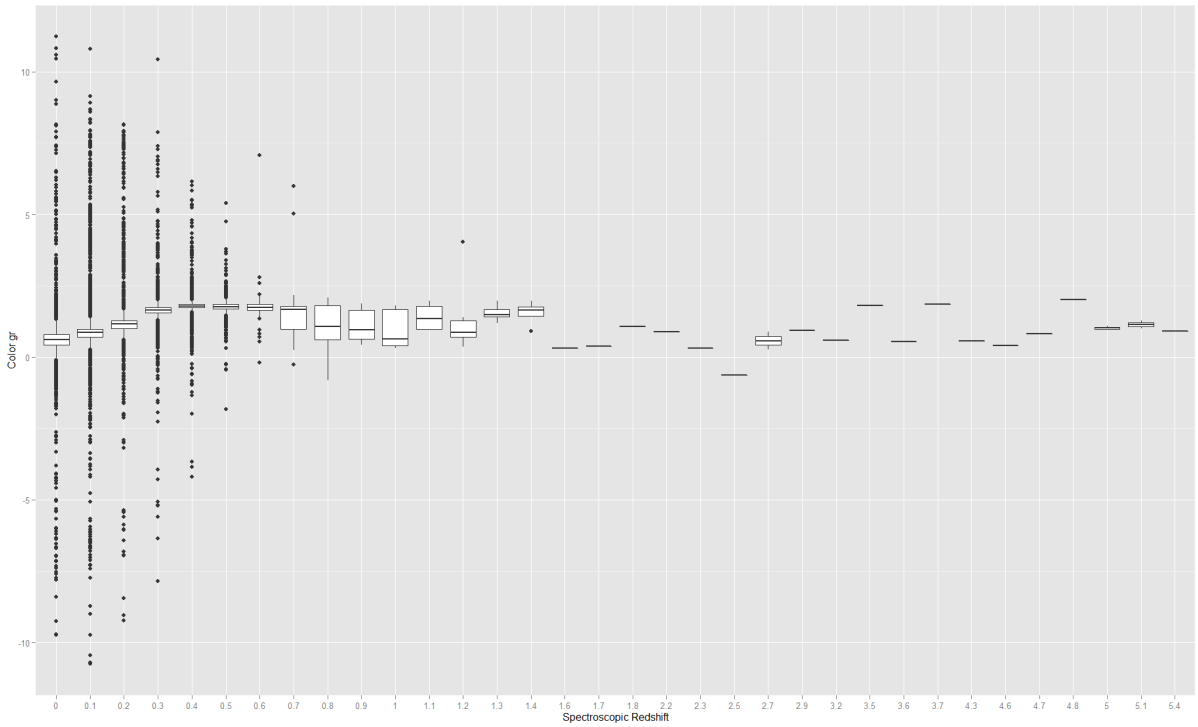


Figure 7.1: Color $g - r$ distribution with redshift

name, is any sort of matter that is not composed primarily of baryons. Those might include neutrinos or free electrons dark matter, such as supersymmetric particles, axions, or black holes. BAO measurements help cosmologists understand more about the nature of dark energy (which causes the apparent slight acceleration of the expansion of the Universe). The cosmic microwave background (CMB) [1] radiation is light emitted after electrons and protons in the plasma could combine to form neutral hydrogen atoms which is only now reaching our telescopes. Wilkinson Microwave Anisotropy Probe (WMAP) [217] data shows an image of the Universe when it was only 379,000 years old. [218]

[1] About 400,000 years after the big bang, the continued expansion and cooling of the universe had dropped the temperature to about 3,000 degrees, which was cool enough for the first hydrogen atoms to form. This is the epoch of recombination. A fundamental change in the universe occurred at that time when the cosmos went from being filled with a plasma that was opaque to light to being filled with an atomic gas through which light could freely pass. It is this freely streaming radiation that we observe at radio wavelengths as the faint glow known as the CMB. The near uniformity of the CMB observed across the sky and the nature of the minute brightness fluctuations we measure in the CMB are just what is expected if inflation occurred. The CMB is therefore a fantastic signal telling us about the early universe. [14]

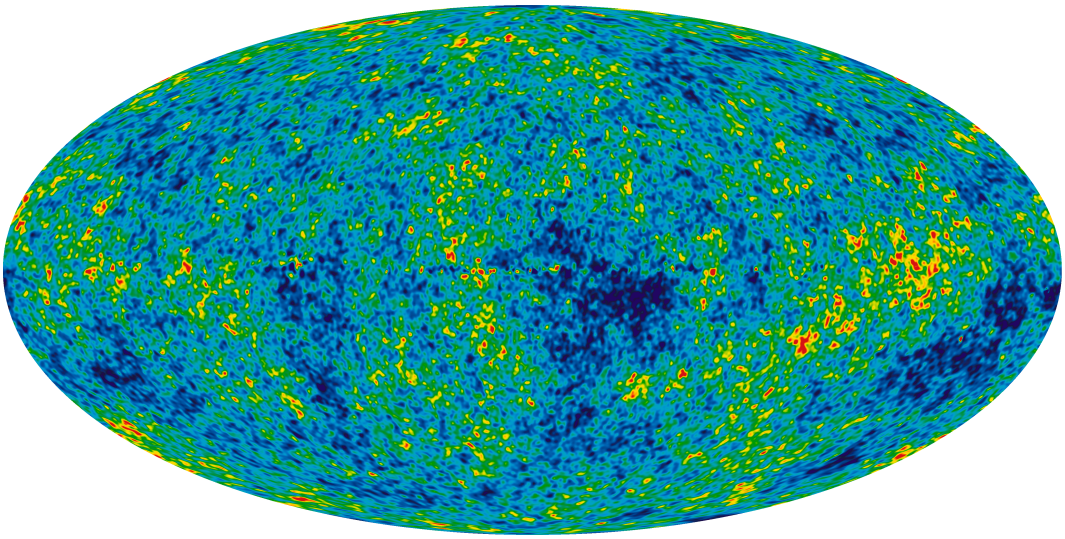


Figure 7.2: Temperature anisotropies of the CMB based on the nine year WMAP data (2012) [219]

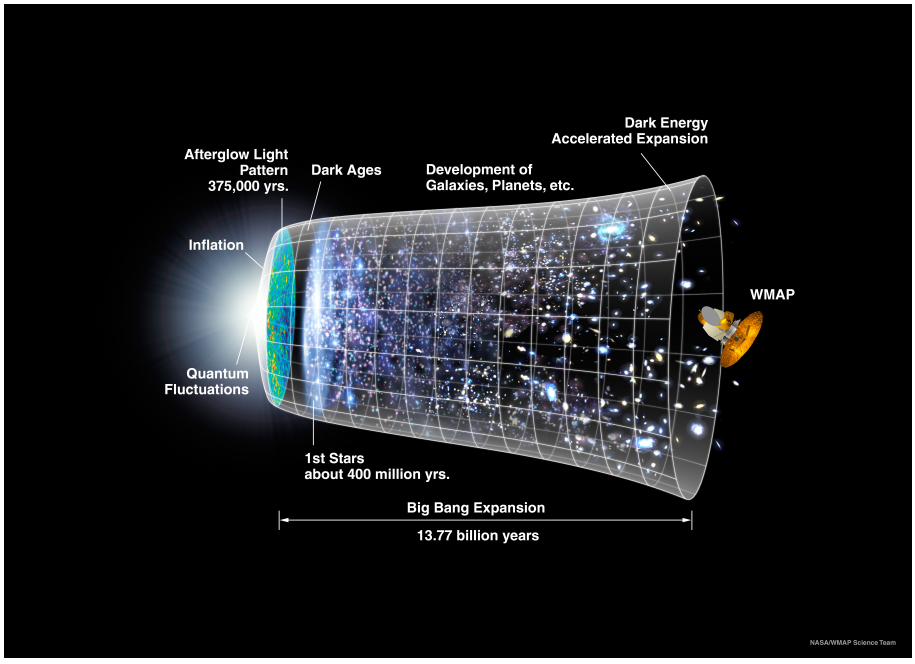


Figure 7.3: Universe Timeline WMAP data (2012) [220]

WMAP indicates (refer Figure 7.2 [219] - 13.77 billion year old temperature fluctuations (shown as color differences) that correspond to the seeds that grew to become the galaxies. This image shows a temperature range of 200 microKelvin.) a smooth, homogeneous universe with density anisotropies of 10 parts per million.[221] However, when we observe the Universe today we find large structure and density fluctuations. Galaxies, for instance, are a million times more dense than the Universe's mean density. The current belief is that the Universe was built in a bottom-up fashion, meaning that the small anisotropies of the early universe acted as gravitational seeds for the structure we see today. Overdense regions attract more matter, while underdense regions attract less, and thus these small anisotropies we see in the CMB become the large scale structures we observe in the Universe today.

A representation of the evolution of the universe over 13.77 billion years is shown in 7.3. The far left depicts the earliest moment we can now probe, when a period of "inflation" produced a burst of exponential growth in the universe. (Size is depicted by the vertical extent of the grid in this graphic.) For the next several billion years, the expansion of the universe gradually slowed down as the matter in the universe pulled on itself via gravity. More recently, the expansion has begun to speed up again as the repulsive effects of dark energy have come to dominate the expansion of the universe. The afterglow light seen by WMAP was emitted about 375,000 years after inflation and has traversed the universe largely unimpeded since then. The conditions of earlier times are imprinted on this light; it also forms a backlight for later developments of the universe.

A three dimensional map of the universe can be formulated using the valuable likelihood information for each possible value that is associated with each Bayesian photomorphic redshift estimate. The map will represent the likelihood mass distribution of the universe and can be validated and improved upon as the ongoing deep survey collect more data related to dark energy. The map based on this study can be the proof of concept to be validated. Other potential problem statements where this formalism of estimating using Bayesian Network modeling need to be considered in the astronomy domain as well as other domains.

Preselection by photometric redshifts may be required to select the rare high-redshift luminous galaxies from the more numerous lower redshift galaxies. For the redshift range between 1.2 and 2.0, this may require near-infrared imaging data as the 4000 Åbreak shifts out beyond $1\mu m$. If one abandons spectroscopy in favor of photometric redshifts, near-infrared data is likely invaluable for the redshift range between 1.2 and 2.5 if one is to reach the required 4% uncertainty goal. A ground-based survey such as LSST would have the depth, but it is not clear that photometric redshift accuracy would be sufficient over the full redshift range. The acoustic oscillation method works better at $z > 1$ and can carry distance measurements out to $z \sim 3$ or higher. [221] A comparison study needs to be performed to validate this.

Association rule learners have been found to be very useful in wide range of domains such as predicting traffic by autonomous agents within a vehicle route planning system [222], items frequently sold together at a retail store (the famous diaper-beer example), ¡add examples¿. Based on these applications, it is worthwhile to investigate if rule-learners will be equally effective in analysis of astronomy objects. If the existing algorithms such as AQ21, Ripper, C4.5 among others are not effective in their present form, analysis and implementation of potential changes that can improve the performance will improve their utility as a tool. It will shed additional light on how the intricacies of a particular domain can be harnessed to fine-tune and accessorize rule-learners. Neural networks have been successfully used for this problem and they have shown promising results. This makes it even more important to check out if rule learners can be harnessed for this interesting problem of distance. The expressive power of the rule learners and the ease of comprehensibility that comes with it makes it an even more effective science tool.

Generalized Linear Model and Naïve Bayes have been used to successfully estimate in the range $z = 0 - 0.25$ where there is concentrated amount of data. Bayesian network learning models such as HC and Tabu (score-based methods) and Aracne (hybrid method) perform well as well. The performance can improve further with more sophisticated data manipulation techniques involving the discretization method, the number of levels or breaks

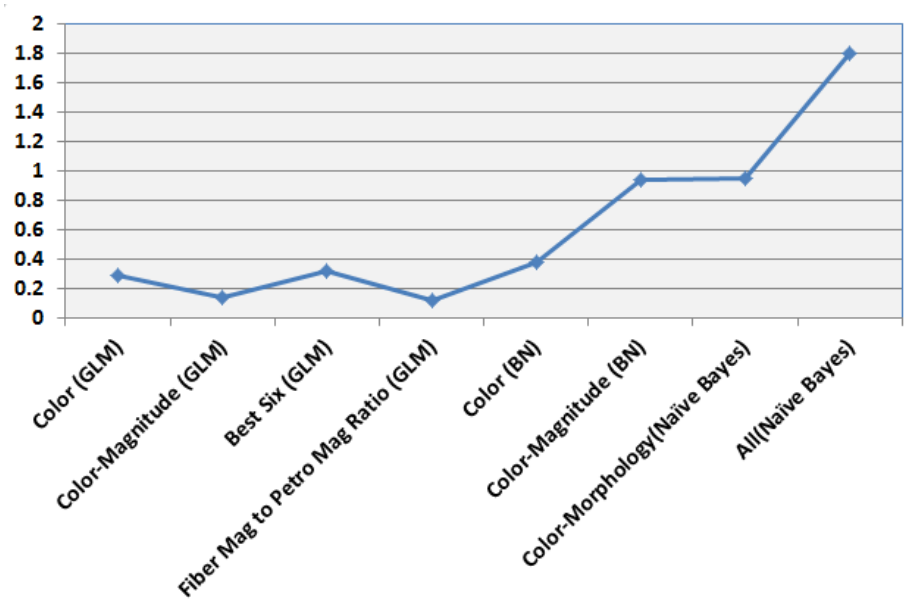


Figure 7.4:]
Performance Comparison of Method and Attribute set vs. Catastrophic Error% for spec-z = (0-0.1]

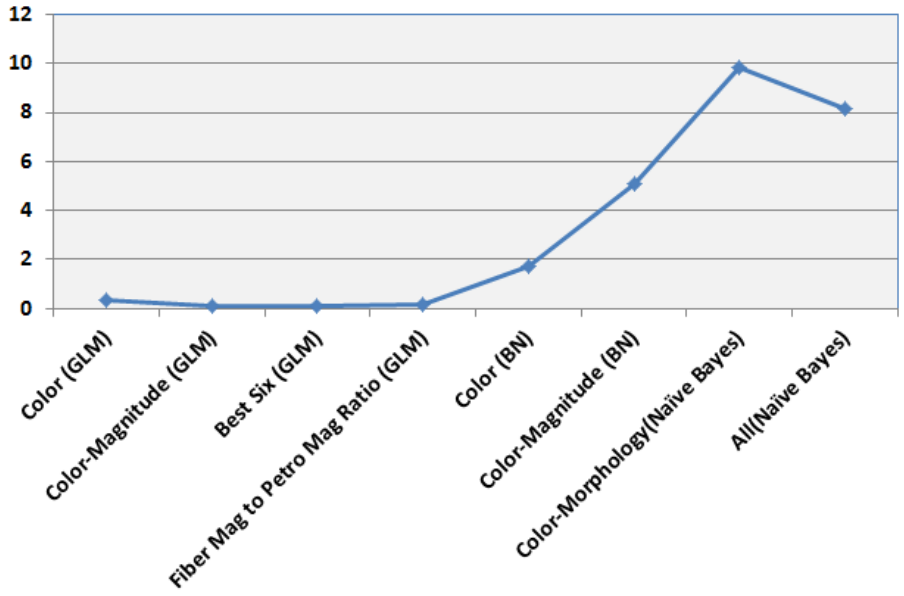


Figure 7.5:]
Performance Comparison of Method and Attribute set vs. Catastrophic Error% for spec-z = (0.1-0.2]

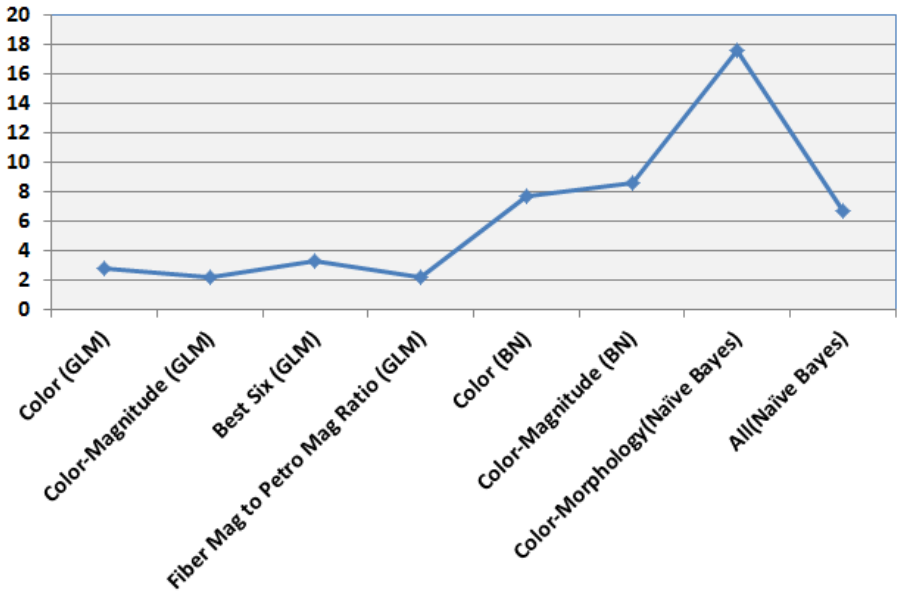


Figure 7.6:]
Performance Comparison of Method and Attribute set vs. Catastrophic Error% for spec-z = (0.2-0.3]

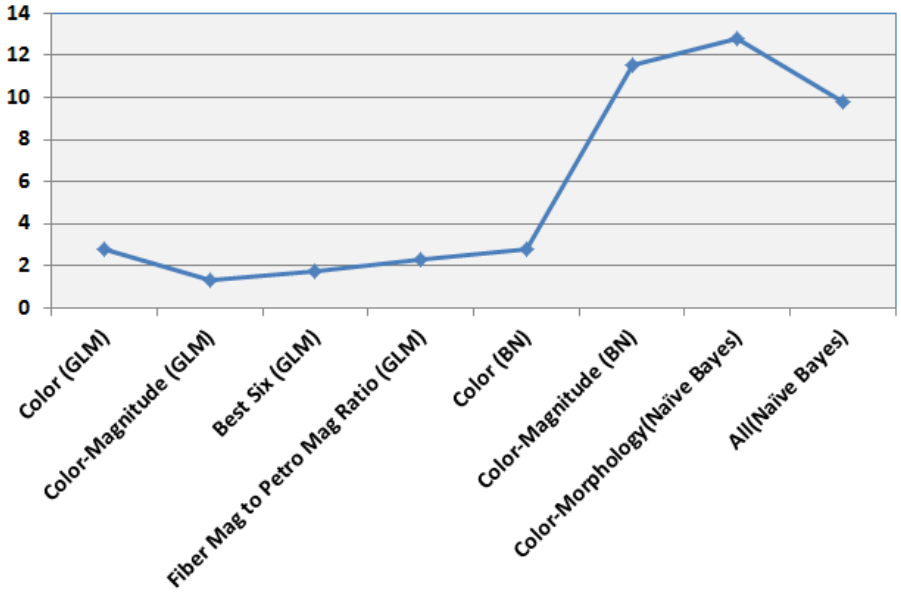


Figure 7.7:]
Performance Comparison of Method and Attribute set vs. Catastrophic Error% for spec-z = (0.3-0.4]

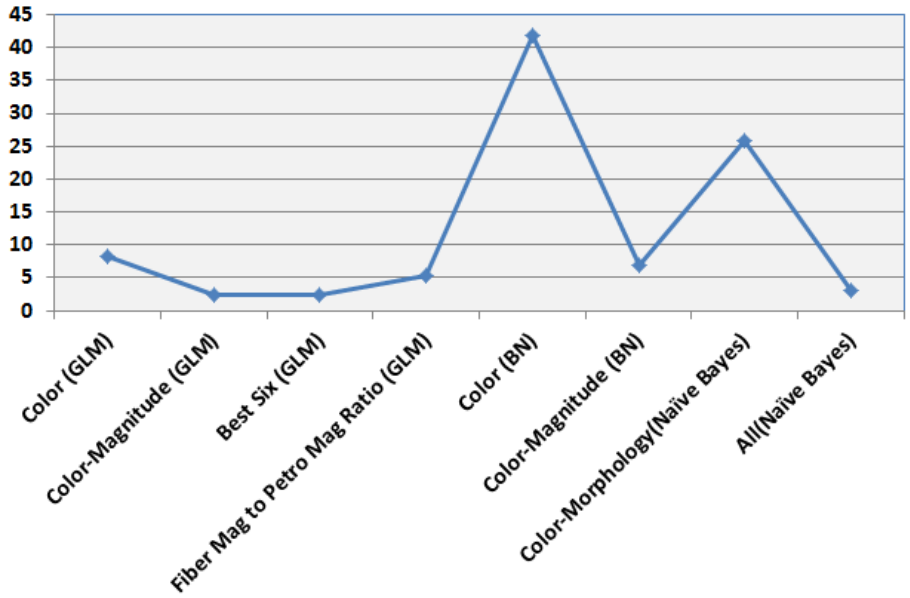


Figure 7.8:]
Performance Comparison of Method and Attribute set vs. Catastrophic Error% for spec-z = (0.4-0.5]

and aggregation of methods in an ensemble. Uniform sampling from the different ranges shows promise as a technique. Additionally, data imputation techniques should be surveyed to generate data for higher redshift ranges and include them in the training data inorder to increase the coverage as well as accuracy of the estimation methods.

GLM beats all the other methods in these ranges (0-0.1],(0.1-0.2],(0.2-0.3] and (0.3-0.4]. Nave Bayes using all the attributes performs in par with the GLM methods in the range (0.4-0.5]. Even though Bayesian Network perform worse than GLM, the likelihood distribution of the estimate provided by my Bayesian Network models are an important contribution to grow our understanding of the data. It is useful in ensemble methods in merging results from different methods of varying configuration, representation and complexity. The likelihood distribution is useful as the kernel for mass distribution in the map of the universe. Photomorphic redshift performs comparable to photometric redshift in all the ranges. It is better in the ranges where there is more data available. More data implies that the data scatter is constrained.

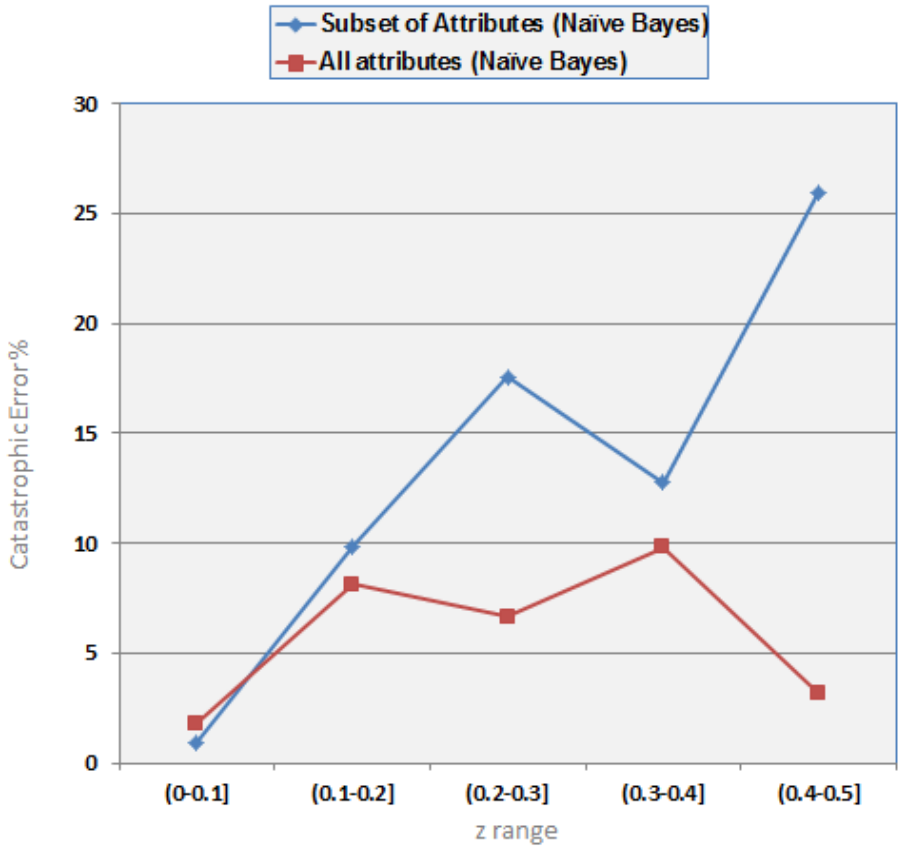


Figure 7.9: Naïve Bayes Performance Summary

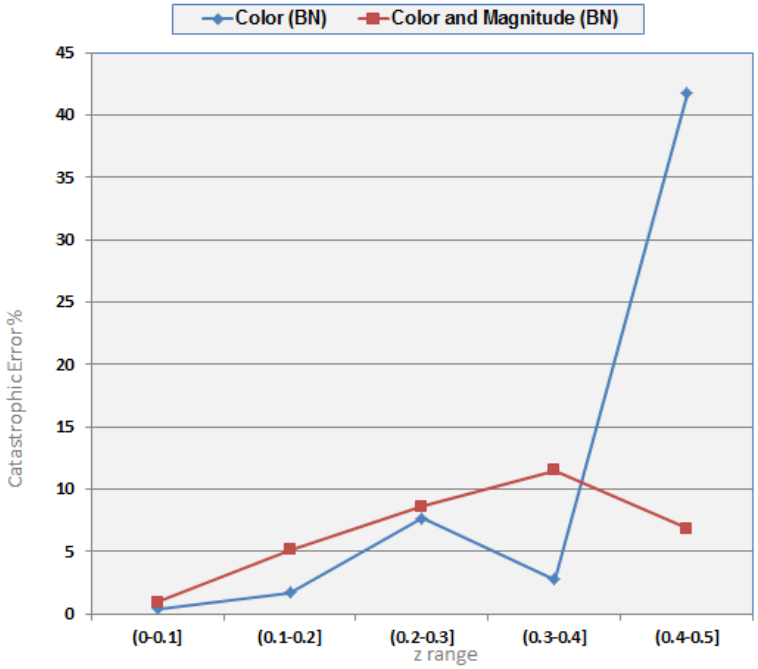


Figure 7.10: Bayesian Network Performance Summary

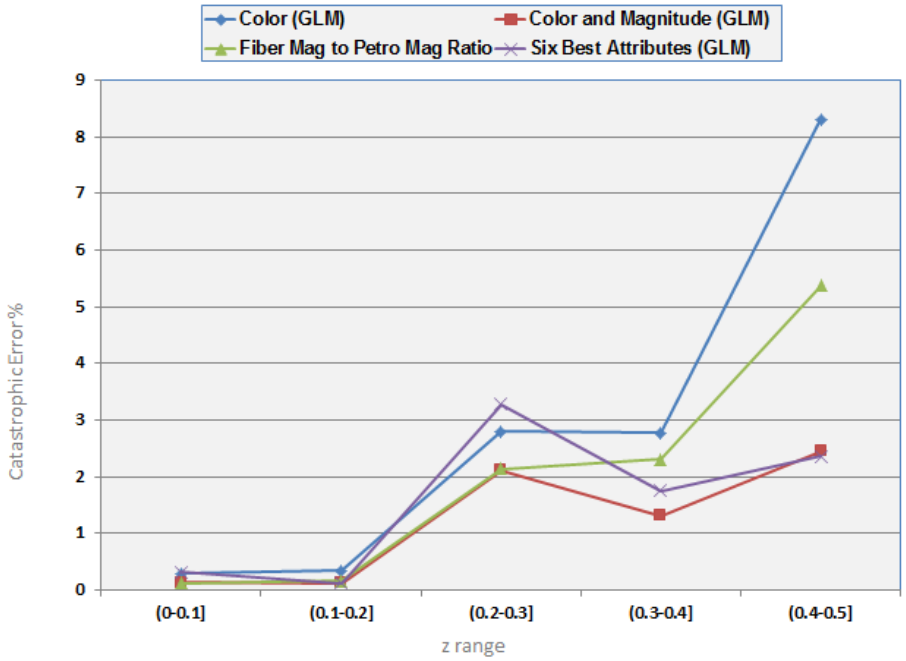


Figure 7.11: GLM Performance Summary

Appendix A: SDSS CAS Server SQL Query

```
select
    spct.z rdshft,                                                    -- Redshift
    psfMag_u, psfMag_g, psfMag_r, psfMag_i, psfMag_z,                 -- PSF Flux
    fiberMag_u, fiberMag_g, fiberMag_r, fiberMag_i, fiberMag_z, -- Flux in 3 arcsec↩
        diameter fiber radius
    petroMag_u, petroMag_g, petroMag_r, petroMag_i, petroMag_z, -- Petrosian flux
    -- Ratio of Fiber Magnitude/Petrosian Magnitude
    fiberMag_u/petroMag_u fp_u, fiberMag_g/petroMag_g fp_g, fiberMag_r/petroMag_r ↩
        fp_r, fiberMag_i/petroMag_i fp_i, fiberMag_z/petroMag_z fp_z,
    petroRad_u, petroRad_g, petroRad_r, petroRad_i, petroRad_z, -- Petrosian radius
    petroR50_u, petroR50_g, petroR50_r, petroR50_i, petroR50_z, -- Radius ↩
        containing 50 percent of Petrosian flux
    petroR90_u, petroR90_g, petroR90_r, petroR90_i, petroR90_z, -- Radius ↩
        containing 90 percent of Petrosian flux
    -- Ratio of Radius containing 50 percent Petrosian flux to Radius containing 90↩
        percent Petrosian flux
    -- Angular Size that determines concentration and spread of galaxy. Directly ↩
        proportional to the density of the galaxy.
    petroR50_u/petroR90_u pR_u, petroR50_g/petroR90_g pR_g, petroR50_r/petroR90_r ↩
        pR_r, petroR50_i/petroR90_i pR_i, petroR50_z/petroR90_z pR_z,
    u-g ug, g-r gr, r-i ri, i-glxy.z iz                               -- Model magnitude
into MyDB.PhotoZLTE05
from DR10.SpecObj spct, DR10.Galaxy glxy
where spct.specObjid = glxy.specObjID and
  psfMag_u != -9999 and  psfMag_g != -9999 and  psfMag_r != -9999 and  psfMag_i != ↩
      -9999 and  psfMag_z != -9999 and
  petroR50_u != -9999 and   petroR50_g != -9999 and   petroR50_r != -9999 and   ↩
      petroR50_i != -9999 and  petroR50_z != -9999 and
  zWarning = dbo.fSpecZWarning('OK') and sourceType = 'GALAXY' and
  spct.z <= 0.5;
```

Appendix B: Petrosian Quantities

The Petrosian (1976) magnitude is based on the flux within an aperture defined by the ratio of the local surface brightness to the mean interior surface brightness. The size of this aperture depends on the shape of the galaxy's radial surface brightness profile but not its amplitude.

Let $I(\theta)$ be the azimuthally averaged surface brightness profile of a galaxy, as a function of angular distance from its center, θ. *Petrosian ratio* is the ratio of the surface brightness in an annulus $0.8\theta 1.25\theta$ to the mean surface brightness within θ,

$$\Re(\theta) = \frac{2\pi \int_{0.8\theta}^{1.25\theta} I(\theta')\theta' d\theta' / (\pi[(1.25\theta)^2 - (0.8\theta)^2])}{2\pi \int_0^{\theta} I(\theta')\theta' d\theta' / (\pi\theta^2)} \tag{2.1}$$

The use of a fairly thick annulus reduces the sensitivity of $I(\theta)$ to noise and to small-scale fluctuations in $I(\theta)$. *Petrosian radius* θ_P is the radius that satisfies the condition $\Re(\theta_P) = 0.2$.

The *Petrosian flux* F_P is defined as the flux within a circular aperture of radius twice the Petrosian radius. *Petrosian Aperture* is another name for twice the Petrosian radius.

$$\textit{Petrosian flux}, F_P = 2\pi \int_0^{2\theta_P} I(\theta')\theta' d\theta' \tag{2.2}$$

$$\textit{Total flux}, F_{tot} = 2\pi \int_0^{\infty} I(\theta')\theta' d\theta' \tag{2.3}$$

Petrosian half-light radius θ_{50} is the radius within which is enclosed half the Petrosian flux,

$$\int_0^{\theta_{50}} I(\theta')\theta' d\theta' = 0.5 \int_0^{2\theta_P} I(\theta')\theta' d\theta' \tag{2.4}$$

Note: Because the flux within $2\theta_P$ is insensitive to small errors in θ_P, the θ_{50} is robustly measured.

Appendix C: Flowchart for GLM regression

A brief summary of GLM analysis and model diagnostics is shown below [99]. It comprises:

- Acquire the dataset.
- Choose the response variable to be modelled.
- Choose predictor variables.
- Choose GLM family, e.g. Gaussian, Poisson, binomial.
- Choose either a maximum-likelihood or a Bayesian approach.
- Choose link function.
- Estimating coefficients by means of a GLM or Bayesian GLM analysis
- Classification and diagnostic tests:
  - ROC curve-probability threshold.
  - Confusion Matrix and assigned class memberships.

Bibliography

[1] H. V. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, and C. Shahabi, "Big data and its technical challenges," *Commun. ACM*, vol. 57, no. 7, pp. 86–94, Jul. 2014. [Online]. Available: http://doi.acm.org/10.1145/2611567

[2] Presidents Council of Advisors on Science and Technology, "Report to the President - Big Data and Privacy: A Technological Perspective," 2014. [Online]. Available: http://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf

[3] G. Anthes, "Data brokers are watching you," *Commun. ACM*, vol. 58, no. 1, pp. 28–30, Dec. 2014. [Online]. Available: http://doi.acm.org/10.1145/2686740

[4] M. Fertik, "The rich see a different internet than the poor," *Scientific American*, vol. 308, no. 2, Feb 2013. [Online]. Available: http://www.scientificamerican.com/article/rich-see-different-internet-than-the-poor/

[5] A. Wright, "Big data meets big science," *Commun. ACM*, vol. 57, no. 7, pp. 13–15, Jul. 2014. [Online]. Available: http://doi.acm.org/10.1145/2617660

[6] G.-H. Kim, S. Trimi, and J.-H. Chung, "Big data Applications in the Government Sector," *Commun. ACM*, vol. 57, no. 3, pp. 78–85, Mar. 2014. [Online]. Available: http://doi.acm.org/10.1145/2500873

[7] "Are you prepared to meet the challenges of the data act and open the door wider on government spending?" AGA Web Conference, November 5 2014.

[8] M. Y. Vardi, "Science has only two legs," *Commun. ACM*, vol. 53, no. 9, pp. 5–5, Sep. 2010. [Online]. Available: http://doi.acm.org/10.1145/1810891.1810892

[9] L. Wynholds, D. Fearon, C. L. Borgman, and S. Traweek, "Awash in stardust: Data practices in astronomy," in *Proceedings of the 2011 iConference*, ser. iConference '11. New York, NY, USA: ACM, 2011, pp. 802–804. [Online]. Available: http://doi.acm.org/10.1145/1940761.1940912

[10] G. B. Berriman and S. L. Groom, "How will astronomy archives survive the data tsunami?" *Commun. ACM*, vol. 54, no. 12, pp. 52–56, Dec. 2011. [Online]. Available: http://doi.acm.org/10.1145/2043174.2043190

[11] K. Borne, A. Accomazzi, J. Bloom, R. Brunner, D. Burke, N. Butler, D. F. Chernoff, B. Connolly, A. Connolly, A. Connors, C. Cutler, S. Desai, G. Djorgovski, E. Feigelson, L. S. Finn, P. Freeman, M. Graham, N. Gray, C. Graziani, E. F. Guinan, J. Hakkila,

S. Jacoby, W. Jefferys, Kashyap, B. Kelly, K. Knuth, D. Q. Lamb, H. Lee, T. Loredo, A. Mahabal, M. Mateo, B. McCollum, A. Muench, M. Pesenson, V. Petrosian, F. Primini, P. Protopapas, A. Ptak, J. Quashnock, M. J. Raddick, G. Rocha, N. Ross, L. Rottler, J. Scargle, A. Siemiginowska, I. Song, A. Szalay, J. A. Tyson, T. Vestrand, J. Wallin, B. Wandelt, I. M. Wasserman, M. Way, M. Weinberg, A. Zezas, E. Anderes, J. Babu, J. Becla, J. Berger, P. J. Bickel, M. Clyde, I. Davidson, D. van Dyk, T. Eastman, B. Efron, C. Genovese, A. Gray, W. Jang, E. D. Kolaczyk, J. Kubica, J. M. Loh, X.-L. Meng, A. Moore, R. Morris, T. Park, R. Pike, J. Rice, J. Richards, D. Ruppert, N. Saito, C. Schafer, P. B. Stark, M. Stein, J. Sun, D. Wang, Z. Wang, L. Wasserman, E. J. Wegman, R. Willett, R. Wolpert, and M. Woodroofe, "Astroinformatics: A 21st Century Approach to Astronomy," in *astro2010: The Astronomy and Astrophysics Decadal Survey*, ser. Astronomy, vol. 2010, 2009, p. 6P.

[12] K. Borne, J. Becla, I. Davidson, A. Szalay, and J. A. Tyson, "The LSST Data Mining Research Agenda," in *American Institute of Physics Conference Series*, ser. American Institute of Physics Conference Series, C. A. L. Bailer-Jones, Ed., vol. 1082, Dec. 2008, pp. 347–351.

[13] L. Wynholds, D. S. Fearon, Jr., C. L. Borgman, and S. Traweek, "When use cases are not useful: Data practices, astronomy, and digital libraries," in *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, ser. JCDL '11. New York, NY, USA: ACM, 2011, pp. 383–386. [Online]. Available: http://doi.acm.org/10.1145/1998076.1998146

[14] N. R. Council, *New Worlds, New Horizons in Astronomy and Astrophysics*. Washington, DC: The National Academies Press, 2010. [Online]. Available: http://www.nap.edu/catalog/12951/new-worlds-new-horizons-in-astronomy-and-astrophysics

[15] D. Kirk, O. Lahav, S. Bridle, S. Jouvel, F. B. Abdalla, and J. A. Frieman, "Optimising Spectroscopic and Photometric Galaxy Surveys: Same-sky Benefits for Dark Energy and Modified Gravity," *ArXiv e-prints*, Jul. 2013.

[16] V. P. Reshetnikov, "Sky surveys and deep fields of ground-based and space telescopes," *Physics-Uspekhi*, vol. 48, no. 11, p. 1109, 2005. [Online]. Available: http://stacks.iop.org/1063-7869/48/i=11/a=R02

[17] Image Credits: X-ray: NASA/CXC/CfA/E. O'Sullivan Optical: Canada-France-Hawaii-Telescope/Coelum, "Stephan's Quintet-A Galaxy Collision in Action," 2009. [Online]. Available: http://www.nasa.gov/multimedia/imagegallery/image_feature_1408.html

[18] Space.com, "Gallery: 65 All-Time Great Galaxy Hits," 2011. [Online]. Available: http://www.space.com/13262-65-great-galaxy-photos-space-images.html/

[19] SDSS, "Characteristics of Galaxies," 2013. [Online]. Available: http://skyserver.sdss.org/dr1/en/proj/advanced/galaxies/characteristics.asp

[20] N. R. Tanvir, D. B. Fox, A. J. Levan, E. Berger, K. Wiersema, J. P. U. Fynbo, A. Cucchiara, T. Krühler, N. Gehrels, J. S. Bloom, J. Greiner, P. A. Evans, E. Rol,

F. Olivares, J. Hjorth, P. Jakobsson, J. Farihi, R. Willingale, R. L. C. Starling, S. B. Cenko, D. Perley, J. R. Maund, J. Duke, R. A. M. J. Wijers, A. J. Adamson, A. Allan, M. N. Bremer, D. N. Burrows, A. J. Castro-Tirado, B. Cavanagh, A. de Ugarte Postigo, M. A. Dopita, T. A. Fatkhullin, A. S. Fruchter, R. J. Foley, J. Gorosabel, J. Kennea, T. Kerr, S. Klose, H. A. Krimm, V. N. Komarova, S. R. Kulkarni, A. S. Moskvitin, C. G. Mundell, T. Naylor, K. Page, B. E. Penprase, M. Perri, P. Podsiadlowski, K. Roth, R. E. Rutledge, T. Sakamoto, P. Schady, B. P. Schmidt, A. M. Soderberg, J. Sollerman, A. W. Stephens, G. Stratta, T. N. Ukwatta, D. Watson, E. Westra, T. Wold, and C. Wolf, "A γ-ray burst at a redshift of $z \sim 8.2$," *Nature*, vol. 461, pp. 1254–1257, Oct. 2009.

[21] S. L. Finkelstein, C. Papovich, M. Dickinson, M. Song, V. Tilvi, A. M. Koekemoer, K. D. Finkelstein, B. Mobasher, H. C. Ferguson, M. Giavalisco, N. Reddy, M. L. N. Ashby, A. Dekel, G. G. Fazio, A. Fontana, N. A. Grogin, J.-S. Huang, D. Kocevski, M. Rafelski, B. J. Weiner, and S. P. Willner, "A galaxy rapidly forming stars 700 million years after the Big Bang at redshift 7.51," *Nature*, vol. 502, pp. 524–527, Oct. 2013.

[22] M. Kilbinger, L. Fu, C. Heymans, F. Simpson, J. Benjamin, T. Erben, J. Harnois-Déraps, H. Hoekstra, H. Hildebrandt, T. D. Kitching, Y. Mellier, L. Miller, L. Van Waerbeke, K. Benabed, C. Bonnett, J. Coupon, M. J. Hudson, K. Kuijken, B. Rowe, T. Schrabback, E. Semboloni, S. Vafaei, and M. Velander, "CFHTLenS: combined probe cosmological model comparison using 2D weak gravitational lensing," *mnras*, vol. 430, pp. 2200–2220, Apr. 2013.

[23] M. J. Jee, J. A. Tyson, M. D. Schneider, D. Wittman, S. Schmidt, and S. Hilbert, "Cosmic shear results from the deep lens survey. i. joint constraints on m and 8 with a two-dimensional analysis," *The Astrophysical Journal*, vol. 765, no. 1, p. 74, 2013. [Online]. Available: http://stacks.iop.org/0004-637X/765/i=1/a=74

[24] M. J. Z. Shichao Zhang, "Mining multiple data sources: Local pattern analysis," in *Data Mining and Knowledge Discovery*, S. S. . B. M. Inc., Ed., vol. 12, April 2006, pp. 121–125.

[25] J. A. Newman, M. C. Cooper, M. Davis, S. M. Faber, A. L. Coil, P. Guhathakurta, D. C. Koo, A. C. Phillips, C. Conroy, A. A. Dutton, D. P. Finkbeiner, B. F. Gerke, D. J. Rosario, B. J. Weiner, C. N. A. Willmer, R. Yan, J. J. Harker, S. A. Kassin, N. P. Konidaris, K. Lai, D. S. Madgwick, K. G. Noeske, G. D. Wirth, A. J. Connolly, N. Kaiser, E. N. Kirby, B. C. Lemaux, L. Lin, J. M. Lotz, G. A. Luppino, C. Marinoni, D. J. Matthews, A. Metevier, and R. P. Schiavon, "The DEEP2 Galaxy Redshift Survey: Design, Observations, Data Reduction, and Redshifts," *apjl*, vol. 208, p. 5, Sep. 2013.

[26] UCI Machine Learning Repository, "Low Resolution Spectrometer Data Set," 1988. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Low+Resolution+Spectrometer

[27] T. Nagata and S. Sato, "Near infrared sky survey: its history and future." *Astronomical Herald*, vol. 90, pp. 568–574, 1997.

[28] I. N. Reid, "Photographic Sky Surveys," in *New Horizons from Multi-Wavelength Sky Surveys*, ser. IAU Symposium, B. J. McLean, D. A. Golombek, J. J. E. Hayes, and H. E. Payne, Eds., vol. 179, 1998, pp. 41–+.

[29] T. degree-Field (2dF) Redshift Survey, "The two-degree-field (2df) redshift survey summary statistics," 2002, [Online;The 2dF is designed to allow the acqusition of up to 400 simultaneous spectra of objects anywhere within a two degree field on the sky. The 2dFGRS data base is available on the World Wide Web at http://www.mso.anu.edu.au/2dFGRS.]. [Online]. Available: http://www2.aao.gov.au/~TDFgg/Public/Survey/statusfinal.html

[30] 2MASS, "2MASS Website," 2009, 2MASS. [Online]. Available: http://www.ipac.caltech.edu/2mass/overview/about2mass.html

[31] SDSS, "Sloan Digital Sky Survey website."

[32] Sloan Digital Sky Survey III Data Release 10, "Understanding SDSS spectroscopic data," 2013. [Online]. Available: https://www.sdss3.org/dr10/spectro/spectro_basics.php

[33] Sloan Digital Sky Survey III Data Release 10, "Galaxy Properties," 2013. [Online]. Available: https://www.sdss3.org/dr10/spectro/galaxy.php

[34] Daniel J. Matthews and Jeffrey A. Newman and Alison L. Coil and Michael C. Cooper and Stephen D. J. Gwyn, "Extended Photometry for the DEEP2 Galaxy Redshift Survey: A Testbed for Photometric Redshift Experiments," *apjs*, vol. 204, no. 2, p. 21, 2013. [Online]. Available: http://stacks.iop.org/0067-0049/204/i=2/a=21

[35] R. G. Abraham, K. Glazebrook, P. J. McCarthy, D. Crampton, R. Murowinski, I. Jrgensen, K. Roth, I. M. Hook, S. Savaglio, H.-W. Chen, R. O. Marzke, and R. G. Carlberg, "The Gemini Deep Deep Survey. I. Introduction to the Survey, Catalogs, and Composite Spectra," *The Astronomical Journal*, vol. 127, no. 5, p. 2455, 2004. [Online]. Available: http://stacks.iop.org/1538-3881/127/i=5/a=2455

[36] C. C. Steidel, A. E. Shapley, M. Pettini, K. L. Adelberger, D. K. Erb, N. A. Reddy, and M. P. Hunt, "A survey of star-forming galaxies in the 1.4 z 2.5 redshift desert: Overview," *apjl*, vol. 604, no. 2, p. 534, 2004. [Online]. Available: http://stacks.iop.org/0004-637X/604/i=2/a=534

[37] The Dark Energy Survey, "The Dark Energy Survey," 2015. [Online]. Available: http://www.darkenergysurvey.org/

[38] F. Abdalla, J. Annis, D. Bacon, S. Bridle, F. Castander, M. Colless, D. DePoy, H. T. Diehl, M. Eriksen, B. Flaugher, J. Frieman, E. Gaztanaga, C. Hogan, S. Jouvel, S. Kent, D. Kirk, R. Kron, S. Kuhlmann, O. Lahav, J. Lawrence, H. Lin, J. Marriner, J. Marshall, J. Mohr, R. C. Nichol, M. Sako, W. Saunders, M. Soares-Santos, D. Thomas, R. Wechsler, A. West, and H. Wu, "The Dark Energy Spectrometer (DESpec): A Multi-Fiber Spectroscopic Upgrade of the Dark Energy Camera and Survey for the Blanco Telescope," *ArXiv e-prints*, Sep. 2012.

[39] A. Connolly, "What's the next window into our universe?" TED Talk, 2014. [Online]. Available: https://www.ted.com/talks/andrew_connolly_what_s_the_next_window_into_our_universe

[40] IVOA, "International Virtual Observatory Alliance," 2009, international Virtual Observatory Alliance. [Online]. Available: http://www.ivoa.net/

[41] US-NVO, "US National Virtual Observatory," 2009, uS National Virtual Observatory. [Online]. Available: http://www.us-vo.org

[42] J. B. Hutchings and L. Bianchi, "QSOs in the Combined SDSS/GALEX Database," *pasp*, vol. 120, pp. 275–280, Feb. 2008.

[43] CDS, "CDS website," 2009. [Online]. Available: http://cdsarc.u-strasbg.fr/

[44] IVOA-Arch, "International Virtual Observatory Alliance architecture," 2009, international Virtual Observatory Alliance Architecture. [Online]. Available: http://www.ivoa.net/Documents/latest/IVOArch.html

[45] Center for Astrophysics (CfA) Redshift Survey, "The CfA Redshift Survey," 2014. [Online]. Available: https://www.cfa.harvard.edu/~dfabricant/huchra/zcat/

[46] M. Colless, G. Dalton, S. Maddox, W. Sutherland, P. Norberg, S. Cole, J. Bland-Hawthorn, T. Bridges, R. Cannon, C. Collins, W. Couch, N. Cross, K. Deeley, R. De Propris, S. P. Driver, G. Efstathiou, R. S. Ellis, C. S. Frenk, K. Glazebrook, C. Jackson, O. Lahav, I. Lewis, S. Lumsden, D. Madgwick, J. A. Peacock, B. A. Peterson, I. Price, M. Seaborne, K. Taylor, and (the 2dFGRS team), "The 2df galaxy redshift survey: spectra and redshifts," *Monthly Notices of the Royal Astronomical Society*, vol. 328, no. 4, pp. 1039–1063, 2001. [Online]. Available: http://dx.doi.org/10.1046/j.1365-8711.2001.04902.x

[47] S. I. D. R. 12, "Sdss scope," 2014, [Online; Data Release 12 represents the culmination of the third phase of the Sloan Digital Sky Survey. It includes all SDSS data taken through 14 July 2014, and encompasses more than one-third of the entire celestial sphere.]. [Online]. Available: http://www.sdss.org/dr12/scope/

[48] N. C. Hambly, H. T. MacGillivray, M. A. Read, S. B. Tritton, E. B. Thomson, B. D. Kelly, D. H. Morgan, R. E. Smith, S. P. Driver, J. Williamson, Q. A. Parker, M. R. S. Hawkins, P. M. Williams, and A. Lawrence, "The SuperCOSMOS Sky Survey - I. Introduction and description," *Monthly Notices of the Royal Astronomical Society*, vol. 326, pp. 1279–1294, Oct. 2001.

[49] DEEP2 Redshift Survey, "DEEP2 Redshift Survey," 2013, dEEP2. [Online]. Available: http://deep.ps.uci.edu/

[50] J. Kanipe, "The universe in your computer," *Commun. ACM*, vol. 52, no. 1, pp. 12–14, Jan. 2009. [Online]. Available: http://doi.acm.org/10.1145/1435417.1435424

[51] C. Moskowitz, "Einstein's 'Biggest Blunder' Turns Out to Be Right," 2010. [Online]. Available: http://www.space.com/9593-einstein-biggest-blunder-turns.html

[52] NASA, "Expanding universe," 2009. [Online]. Available: http://archive.ncsa.uiuc.edu/Cyberia/Cosmos/ExpandUni.html

[53] A. S. D. Center, "What wavelength goes with a color?" 2009. [Online]. Available: http://eosweb.larc.nasa.gov/EDDOCS/Wavelengths_for_Colors.html

[54] SDSS, "What is color?" 2009, what is Color? [Online]. Available: http://skyserver.sdss.org/dr1/en/proj/advanced/color/whatis.asp

[55] R. Giovanelli and M. P. Haynes, "Redshift surveys of galaxies," *Annual Review of Astronomy Astrophysics*, vol. 29, pp. 499–541, 1991.

[56] A. Finkbeiner, "Astronomy: Laser focus," *Nature*, vol. 517(7535), pp. 430–, Jan. 2015. [Online]. Available: http://adsabs.harvard.edu/abs/2013Natur.502..524F

[57] A. M. Manuel, D. W. Phillion, S. S. Olivier, K. L. Baker, and B. Cannon, "Curvature wavefront sensing performance evaluation for active correction of the large synoptic survey telescope (lsst)," *Opt. Express*, vol. 18, no. 2, pp. 1528–1552, Jan 2010. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-18-2-1528

[58] B. Ratra and M. S. Vogeley, "The Beginning and Evolution of the Universe," *pasp*, vol. 120, pp. 235–265, Mar. 2008.

[59] P. Hickson and M. K. Mulrooney, "University of british columbia-nasa multi-narrowband survey. i. description and photometric properties of the survey," *The Astrophysical Journal Supplement Series*, vol. 115, no. 1, p. 35, 1998. [Online]. Available: http://stacks.iop.org/0067-0049/115/i=1/a=35

[60] S. Phleps, S. Drepper, K. Meisenheimer, and B. Fuchs, "Galactic structure from the Calar Alto Deep Imaging Survey (CADIS)," *aap*, vol. 443, pp. 929–943, Dec. 2005.

[61] C. Wolf, K. Meisenheimer, H.-W. Rix, A. Borch, S. Dye, and M. Kleinheinrich, "The combo-17 survey: Evolution of the galaxy luminosity function from 25000 galaxies with $0.2 < z < 1.2$," *aap*, vol. 401, no. 1, pp. 73–98, 2003.

[62] A. Molino, N. Benítez, M. Moles, A. Fernández-Soto, D. Cristóbal-Hornillos, B. Ascaso, Y. Jiménez-Teja, W. Schoenell, P. Arnalte-Mur, M. Pović, D. Coe, C. López-Sanjuan, L. A. Díaz-García, J. Varela, I. Matute, J. Masegosa, I. Márquez, J. Perea, A. Del Olmo, C. Husillos, E. Alfaro, T. Aparicio-Villegas, M. Cerviño, M. Huertas-Company, A. L. Aguerri, T. Broadhurst, J. Cabrera-Caño, J. Cepa, R. M. González Delgado, L. Infante, V. J. Martínez, F. Prada, and J. M. Quintana, "The ALHAMBRA Survey: Bayesian Photometric Redshifts with 23 bands for 3 squared degrees," *ArXiv e-prints*, Jun. 2013.

[63] Pan-STARRS, "Pan-STARRS Website," 2005. [Online]. Available: http://pan-starrs.ifa.hawaii.edu/public/

[64] N. Benítez, M. Moles, J. A. L. Aguerri, E. Alfaro, T. Broadhurst, J. Cabrera-Caño, F. J. Castander, J. Cepa, M. Cerviño, D. Cristóbal-Hornillos, A. Fernández-Soto, R. M. González Delgado, L. Infante, I. Márquez, V. J. Martínez, J. Masegosa, A. Del Olmo, J. Perea, F. Prada, J. M. Quintana, and S. F. Sánchez, "Optimal Filter Systems for Photometric Redshift Estimation," *apjl*, vol. 692, pp. L5–L8, Feb. 2009.

[65] J. Elliott, R. S. de Souza, A. Krone-Martins, E. Cameron, E. E. O. Ishida, and J. Hilbe, "The overlooked potential of Generalized Linear Models in astronomy-II: Gamma regression and photometric redshifts," *Astronomy and Computing*, vol. 10, pp. 61–72, Apr. 2015.

[66] R. Cowen, "Galaxy Hunters: The Search for Cosmic Dawn," *National Geographic Magazine*, pp. 40–65, Feb. 2003.

[67] scikits learn Web Page, "Regression: Photometric Redshifts of Galaxies," 2011. [Online]. Available: http://www.astroml.org/sklearn_tutorial/regression.html

[68] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[69] SDSS, "SDSS Redshift," 2009. [Online]. Available: http://cas.sdss.org/dr6/en/proj/advanced/hubble/conclusion.asp

[70] A. J. Connolly, I. Csabai, A. S. Szalay, D. C. Koo, R. G. Kron, and J. A. Munn, "Slicing Through Multicolor Space: Galaxy Redshifts from Broadband Photometry," *aj*, vol. 110, p. 2655, Dec. 1995.

[71] W. A. Baum, "Photoelectric Magnitudes and Red-Shifts," in *Problems of Extra-Galactic Research*, ser. IAU Symposium, G. C. McVittie, Ed., vol. 15, 1962, p. 390.

[72] D. C. Koo, "Overview - Photometric Redshifts: A Perspective from an Old-Timer[!] on their Past, Present, and Potential," in *Photometric Redshifts and the Detection of High Redshift Galaxies*, ser. Astronomical Society of the Pacific Conference Series, R. Weymann, L. Storrie-Lombardi, M. Sawicki, and R. Brunner, Eds., vol. 191, 1999, p. 3.

[73] J. J. Puschell, F. N. Owen, and R. A. Laing, "Near-infrared photometry of distant radio galaxies - Spectral flux distributions and redshift estimates," *apjl*, vol. 257, pp. L57–L61, Jun. 1982.

[74] S. A. Butchins, "Predicted redshifts of galaxies by broadband photometry," *aap*, vol. 97, pp. 407–409, Apr. 1981.

[75] S. A. Butchins, "Two-colour photometry of a sample of faint galaxies," *mnras*, vol. 203, pp. 1239–1252, Jun. 1983.

[76] D. C. Koo, "Multicolor photometry of the red cluster 0016+16 at Z = 0.54," *apjl*, vol. 251, pp. L75–L79, Dec. 1981.

[77] D. C. Koo, "Optical multicolors - A poor person's Z machine for galaxies," *aj*, vol. 90, pp. 418–440, Mar. 1985.

[78] D. C. Koo, "Multicolor photometry of field galaxies to B = 24," *apjl*, vol. 311, pp. 651–679, Dec. 1986.

[79] E. D. Loh and E. J. Spillar, "Photometric redshifts of galaxies," *apjl*, vol. 303, pp. 154–161, Apr. 1986.

[80] R. G. Kron, "Photometry of a complete sample of faint galaxies," *apjs*, vol. 43, pp. 305–325, Jun. 1980.

[81] R. J. Brunner, A. J. Connolly, A. S. Szalay, and M. A. Bershady, "Toward more precise photometric redshifts: Calibration via ccd photometry," *aj*, vol. 482, no. 1, p. L21, 1997. [Online]. Available: http://stacks.iop.org/1538-4357/482/i=1/a=L21

[82] M. Bolzonella, J.-M. Miralles, and R. Pelló, "Photometric redshifts based on standard SED fitting procedures," *aap*, vol. 363, pp. 476–492, Nov. 2000.

[83] E. M. Edmondson, L. Miller, and C. Wolf, "Bayesian photometric redshifts for weak-lensing applications," *mnras*, vol. 371, pp. 1693–1704, Oct. 2006.

[84] F. B. Abdalla, M. Banerji, O. Lahav, and V. Rashkov, "A comparison of six photometric redshift methods applied to 1.5 million luminous red galaxies," *Monthly Notices of the Royal Astronomical Society*, vol. 417, no. 3, pp. 1891–1903, 2011. [Online]. Available: http://mnras.oxfordjournals.org/content/417/3/1891.abstract

[85] T. Kodama, E. F. Bell, and R. G. Bower, "A bayesian classifier for photometric redshifts: identification of high-redshift clusters," *Monthly Notices of the Royal Astronomical Society*, vol. 302, no. 1, pp. 152–166, 1999. [Online]. Available: http://mnras.oxfordjournals.org/content/302/1/152.abstract

[86] N. Benítez, "Bayesian Photometric Redshift Estimation," *apjl*, vol. 536, pp. 571–583, Jun. 2000.

[87] Y. Wang, N. Bahcall, and E. L. Turner, "A Catalog of Color-based Redshift Estimates for $Z \leq 4$ Galaxies in the Hubble Deep Field," *aj*, vol. 116, pp. 2081–2085, Nov. 1998.

[88] M. Tanaka, "Photometric Redshift with Bayesian Priors on Physical Properties of Galaxies," *ArXiv e-prints*, Jan. 2015.

[89] H. Esmaeilzadeh, A. Sampson, L. Ceze, and D. Burger, "Neural acceleration for general-purpose approximate programs," *Commun. ACM*, vol. 58, no. 1, pp. 105–115, Dec. 2014. [Online]. Available: http://doi.acm.org/10.1145/2589750

[90] R. Nair, "Big data needs approximate computing: Technical perspective," *Commun. ACM*, vol. 58, no. 1, pp. 104–104, Dec. 2014. [Online]. Available: http://doi.acm.org/10.1145/2688072

[91] M. A. Aragon-Calvo, R. van de Weygaert, B. J. T. Jones, and B. Mobasher, "Sub-Megaparsec Individual Photometric Redshift Estimation from Cosmic Web Constraints," *ArXiv e-prints*, Dec. 2014.

[92] B. S. Shen, H. J. Mo, S. D. M. White, M. R. Blanton, G. Kauffmann, W. Voges, J. Brinkmann, and I. Csabai, "Erratum: The size distribution of galaxies in the Sloan Digital Sky Survey," *mnras*, vol. 379, pp. 400–400, Jul. 2007.

[93] F. B. Abdalla, M. Banerji, O. Lahav, and V. Rashkov, "MegaZ-LRG DR6 Catalogue," associated with [84]. Config files for the different photo-z codes available at http://www.ast.cam.ac.uk/ mbanerji/Research/MegaZLRGDR6/Config/. [Online]. Available: http://www.ast.cam.ac.uk/~mbanerji/Research/MegaZLRGDR6/megaz.html

[94] J. Nelson, "Sketching and streaming algorithms for processing massive data," *XRDS*, vol. 19, no. 1, pp. 14–19, Sep. 2012. [Online]. Available: http://doi.acm.org/10.1145/2331042.2331049

[95] C. Sánchez, M. Carrasco Kind, H. Lin, R. Miquel, F. B. Abdalla, A. Amara, M. Banerji, C. Bonnett, R. Brunner, D. Capozzi, A. Carnero, F. J. Castander, L. A. N. da Costa, C. Cunha, A. Fausti, D. Gerdes, N. Greisel, J. Gschwend, W. Hartley, S. Jouvel, O. Lahav, M. Lima, M. A. G. Maia, P. Martí, R. L. C. Ogando, F. Ostrovski, P. Pellegrini, M. M. Rau, I. Sadeh, S. Seitz, I. Sevilla-Noarbe, A. Sypniewski, J. de Vicente, T. Abbot, S. S. Allam, D. Atlee, G. Bernstein, J. P. Bernstein, E. Buckley-Geer, D. Burke, M. J. Childress, T. Davis, D. L. DePoy, A. Dey, S. Desai, H. T. Diehl, P. Doel, J. Estrada, A. Evrard, E. Fernández, D. Finley, B. Flaugher, J. Frieman, E. Gaztanaga, K. Glazebrook, K. Honscheid, A. Kim, K. Kuehn, N. Kuropatkin, C. Lidman, M. Makler, J. L. Marshall, R. C. Nichol, A. Roodman, E. Sánchez, B. X. Santiago, M. Sako, R. Scalzo, R. C. Smith, M. E. C. Swanson, G. Tarle, D. Thomas, D. L. Tucker, S. A. Uddin, F. Valdés, A. Walker, F. Yuan, and J. Zuntz, "Photometric redshift analysis in the Dark Energy Survey Science Verification data," *mnras*, vol. 445, pp. 1482–1506, Dec. 2014.

[96] R. d'Abrusco, G. Longo, M. Paolillo, E. de Filippis, M. Brescia, A. Staiano, and R. Tagliaferri, "The use of neural networks to probe the structure of the nearby universe," *ArXiv Astrophysics e-prints*, Jan. 2007.

[97] National Science Foundation (NSF), "Supernova Caught in the Act - Data-enabled science allowed detection of Type Ia supernova hours after explosion," 2011, press Release 11-261. [Online]. Available: http://www.nsf.gov/news/news_summ.jsp?cntn_id=122537

[98] P. E. Nugent, M. Sullivan, S. B. Cenko, and et.al., "Supernova SN 2011fe from an exploding carbonoxygen white dwarf star," *Nature*, vol. 480, pp. 344–347, Dec. 2011. [Online]. Available: http://www.nature.com/nature/journal/v480/n7377/abs/nature10644.html#supplementary-information

[99] R. S. de Souza, E. Cameron, M. Killedar, J. Hilbe, R. Vilalta, U. Maio, V. Biffi, B. Ciardi, and J. D. Riggs, "The Overlooked Potential of Generalized Linear Models in Astronomy - I: Binomial Regression and Numerical Simulations," *ArXiv e-prints*, Sep. 2014.

[100] S. Cavuoti, M. Brescia, V. De Stefano, and G. Longo, "Photometric redshift estimation based on data mining with PhotoRApToR," *Experimental Astronomy*, Feb. 2015.

[101] M. Obrić, Ž. Ivezić, P. N. Best, R. H. Lupton, C. Tremonti, J. Brinchmann, M. A. Agüeros, G. R. Knapp, J. E. Gunn, C. M. Rockosi, D. Schlegel, D. Finkbeiner,

M. Gaćeša, V. Smolčić, S. F. Anderson, W. Voges, M. Jurić, R. J. Siverd, W. Steinhardt, A. S. Jagoda, M. R. Blanton, and D. P. Schneider, "Panchromatic properties of 99000 galaxies detected by SDSS, and (some by) ROSAT, GALEX, 2MASS, IRAS, GB6, FIRST, NVSS and WENSS surveys," *mnras*, vol. 370, pp. 1677–1698, Aug. 2006.

[102] R. Fassbender, H. Böhringer, A. Nastasi, R. Šuhada, M. Mühlegger, A. de Hoon, J. Kohnert, G. Lamer, J. J. Mohr, D. Pierini, G. W. Pratt, H. Quintana, P. Rosati, J. S. Santos, and A. D. Schwope, "The x-ray luminous galaxy cluster population at $0.9 < z \leq 1.6$ as revealed by the XMM-Newton Distant Cluster Project," *New Journal of Physics*, vol. 13, no. 12, p. 125014, Dec. 2011.

[103] P. Padovani, M. G. Allen, P. Rosati, and N. A. Walton, "Discovery of optically faint obscured quasars with Virtual Observatory tools," *aap*, vol. 424, pp. 545–559, Sep. 2004.

[104] C. Schommer, "An unified definition of data mining," *CoRR*, vol. abs/0809.2696, 2008.

[105] "StatCodes online statistical software for astronomy and related physical sciences," Since 2005. [Online]. Available: http://astrostatistics.psu.edu/statcodes/

[106] Z. Ivezic, A. J. Connolly, J. T. VanderPlas, and A. Gray, *Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data*. Princeton Series in Modern Observational Astronomy, 2014.

[107] D. Pyle, *Data Preparation for Data Mining*. The Morgan Kaufmann Series in Data Management Systems, 1999.

[108] R. Kessler, A. Conley, S. Jha, and S. Kuhlmann, "Supernova Photometric Classification Challenge," *ArXiv e-prints*, Jan. 2010.

[109] R. Kessler, B. Bassett, P. Belov, V. Bhatnagar, H. Campbell, A. Conley, J. A. Frieman, A. Glazov, S. González-Gaitán, R. Hlozek, S. Jha, S. Kuhlmann, M. Kunz, H. Lampeitl, A. Mahabal, J. Newling, R. C. Nichol, D. Parkinson, N. S. Philip, D. Poznanski, J. W. Richards, S. A. Rodney, M. Sako, D. P. Schneider, M. Smith, M. Stritzinger, and M. Varughese, "Results from the Supernova Photometric Classification Challenge," *pasp*, vol. 122, pp. 1415–1431, Dec. 2010.

[110] E. Feigelson and J. Babu, *Modern Statistical Methods for Astronomy With R Applications*. Cambridge University Press, 2012.

[111] K. W. Willett, C. J. Lintott, S. P. Bamford, K. L. Masters, B. D. Simmons, K. R. V. Casteels, E. M. Edmondson, L. F. Fortson, S. Kaviraj, W. C. Keel, T. Melvin, R. C. Nichol, M. J. Raddick, K. Schawinski, R. J. Simpson, R. A. Skibba, A. M. Smith, and D. Thomas, "Galaxy Zoo 2: detailed morphological classifications for 304 122 galaxies from the Sloan Digital Sky Survey," *mnras*, vol. 435, pp. 2835–2860, Nov. 2013.

[112] SDSS Sky Server, "The Petrosian magnitude," 2014. [Online]. Available: http://skyserver.sdss.org/dr7/en/help/docs/algorithm.asp?key=mag_petro

[113] M. Richmond, "Describing the radial profile of light in a galaxy," 2014. [Online]. Available: http://spiff.rit.edu/classes/phys443/lectures/gal_1/petro/petro.html

[114] M. A. Strauss, D. H. Weinberg, R. H. Lupton, V. K. Narayanan, J. Annis, M. Bernardi, M. Blanton, S. Burles, A. J. Connolly, J. Dalcanton, M. Doi, D. Eisenstein, J. A. Frieman, M. Fukugita, J. E. Gunn, Ž. Ivezić, S. Kent, R. S. J. Kim, G. R. Knapp, R. G. Kron, J. A. Munn, H. J. Newberg, R. C. Nichol, S. Okamura, T. R. Quinn, M. W. Richmond, D. J. Schlegel, K. Shimasaku, M. SubbaRao, A. S. Szalay, D. Vanden Berk, M. S. Vogeley, B. Yanny, N. Yasuda, D. G. York, and I. Zehavi, "Spectroscopic Target Selection in the Sloan Digital Sky Survey: The Main Galaxy Sample," *apjl*, vol. 124, pp. 1810–1824, Sep. 2002.

[115] I. Strateva, Ž. Ivezić, G. R. Knapp, V. K. Narayanan, M. A. Strauss, J. E. Gunn, R. H. Lupton, D. Schlegel, N. A. Bahcall, J. Brinkmann, R. J. Brunner, T. Budavári, I. Csabai, F. J. Castander, M. Doi, M. Fukugita, Z. Győry, M. Hamabe, G. Hennessy, T. Ichikawa, P. Z. Kunszt, D. Q. Lamb, T. A. McKay, S. Okamura, J. Racusin, M. Sekiguchi, D. P. Schneider, K. Shimasaku, and D. York, "Color Separation of Galaxy Types in the Sloan Digital Sky Survey Imaging Data," *aj*, vol. 122, pp. 1861–1874, Oct. 2001.

[116] K. Shimasaku, M. Fukugita, M. Doi, M. Hamabe, T. Ichikawa, S. Okamura, M. Sekiguchi, N. Yasuda, J. Brinkmann, I. Csabai, S.-I. Ichikawa, Z. Ivezić, P. Z. Kunszt, D. P. Schneider, G. P. Szokoly, M. Watanabe, and D. G. York, "Statistical Properties of Bright Galaxies in the Sloan Digital Sky Survey Photometric System," *aj*, vol. 122, pp. 1238–1250, Sep. 2001.

[117] Sloan Digital Sky Survery III, "SDSS-III: Four Surveys Executed Simultaneously," 2014. [Online]. Available: https://www.sdss3.org/surveys/

[118] D. J. Eisenstein, D. H. Weinberg, E. Agol, H. Aihara, C. A. Prieto, S. F. Anderson, J. A. Arns, ric Aubourg, S. Bailey, E. Balbinot, R. Barkhouser, T. C. Beers, A. A. Berlind, S. J. Bickerton, D. Bizyaev, M. R. Blanton, J. J. Bochanski, A. S. Bolton, C. T. Bosman, J. Bovy, W. N. Brandt, B. Breslauer, H. J. Brewington, J. Brinkmann, P. J. Brown, J. R. Brownstein, D. Burger, N. G. Busca, H. Campbell, P. A. Cargile, W. C. Carithers, J. K. Carlberg, M. A. Carr, L. Chang, Y. Chen, C. Chiappini, J. Comparat, N. Connolly, M. Cortes, R. A. C. Croft, K. Cunha, L. N. da Costa, J. R. A. Davenport, K. Dawson, N. D. Lee, G. F. P. de Mello, F. de Simoni, J. Dean, S. Dhital, A. Ealet, G. L. Ebelke, E. M. Edmondson, J. M. Eiting, S. Escoffier, M. Esposito, M. L. Evans, X. Fan, B. F. Castell, L. D. Ferreira, G. Fitzgerald, S. W. Fleming, A. Font-Ribera, E. B. Ford, P. M. Frinchaboy, A. E. G. Prez, B. S. Gaudi, J. Ge, L. Ghezzi, B. A. Gillespie, G. Gilmore, L. Girardi, J. R. Gott, A. Gould, E. K. Grebel, J. E. Gunn, J.-C. Hamilton, P. Harding, D. W. Harris, S. L. Hawley, F. R. Hearty, J. F. Hennawi, J. I. G. Hernndez, S. Ho, D. W. Hogg, J. A. Holtzman, K. Honscheid, N. Inada, I. I. Ivans, L. Jiang, P. Jiang, J. A. Johnson, C. Jordan, W. P. Jordan, G. Kauffmann, E. Kazin, D. Kirkby, M. A. Klaene, G. R. Knapp, J.-P. Kneib, C. S. Kochanek, L. Koesterke, J. A. Kollmeier, R. G. Kron, H. Lampeitl, D. Lang, J. E. Lawler, J.-M. L. Goff, B. L. Lee, Y. S. Lee, J. M. Leisenring, Y.-T. Lin, J. Liu, D. C. Long, C. P. Loomis, S. Lucatello,

B. Lundgren, R. H. Lupton, B. Ma, Z. Ma, N. MacDonald, C. Mack, S. Mahadevan, M. A. G. Maia, S. R. Majewski, M. Makler, E. Malanushenko, V. Malanushenko, R. Mandelbaum, C. Maraston, D. Margala, P. Maseman, K. L. Masters, C. K. McBride, P. McDonald, I. D. McGreer, R. G. McMahon, O. M. Requejo, B. Mnard, J. Miralda-Escud, H. L. Morrison, F. Mullally, D. Muna, H. Murayama, A. D. Myers, T. Naugle, A. F. Neto, D. C. Nguyen, R. C. Nichol, D. L. Nidever, R. W. OConnell, R. L. C. Ogando, M. D. Olmstead, D. J. Oravetz, N. Padmanabhan, M. Paegert, N. Palanque-Delabrouille, K. Pan, P. Pandey, J. K. Parejko, I. Pris, P. Pellegrini, J. Pepper, W. J. Percival, P. Petitjean, R. Pfaffenberger, J. Pforr, S. Phleps, C. Pichon, M. M. Pieri, F. Prada, A. M. Price-Whelan, M. J. Raddick, B. H. F. Ramos, I. N. Reid, C. Reyle, J. Rich, G. T. Richards, G. H. Rieke, M. J. Rieke, H.-W. Rix, A. C. Robin, H. J. Rocha-Pinto, C. M. Rockosi, N. A. Roe, E. Rollinde, A. J. Ross, N. P. Ross, B. Rossetto, A. G. Snchez, B. Santiago, C. Sayres, R. Schiavon, D. J. Schlegel, K. J. Schlesinger, S. J. Schmidt, D. P. Schneider, K. Sellgren, A. Shelden, E. Sheldon, M. Shetrone, Y. Shu, J. D. Silverman, J. Simmerer, A. E. Simmons, T. Sivarani, M. F. Skrutskie, A. Slosar, S. Smee, V. V. Smith, S. A. Snedden, K. G. Stassun, O. Steele, M. Steinmetz, M. H. Stockett, T. Stollberg, M. A. Strauss, A. S. Szalay, M. Tanaka, A. R. Thakar, D. Thomas, J. L. Tinker, B. M. Tofflemire, R. Tojeiro, C. A. Tremonti, M. V. Magaa, L. Verde, N. P. Vogt, D. A. Wake, X. Wan, J. Wang, B. A. Weaver, M. White, S. D. M. White, J. C. Wilson, J. P. Wisniewski, W. M. Wood-Vasey, B. Yanny, N. Yasuda, C. Yche, D. G. York, E. Young, G. Zasowski, I. Zehavi, and B. Zhao, "Sdss-iii: Massive spectroscopic surveys of the distant universe, the milky way, and extra-solar planetary systems," *The Astronomical Journal*, vol. 142, no. 3, p. 72, 2011. [Online]. Available: http://stacks.iop.org/1538-3881/142/i=3/a=72

[119] S. S. Server, "Measures Of Flux And Magnitude," 2013. [Online]. Available: https://www.sdss3.org/dr10/algorithms/magnitudes.php

[120] SDSS, "Sloan Digital Sky Survey III SkyServer DR10 Schema Browser," 2014. [Online]. Available: http://cas.sdss.org/dr10/en/help/browser/browser.aspx

[121] Sloan Digital Sky Survery III, "Photometric redshifts algorithms," 2010, [Online; Copyright2010-2013 SDSS III; DR12 URL: http://www.sdss.org/dr12/algorithms/photo-z/]. [Online]. Available: https://www.sdss3.org/dr10/algorithms/photo-z.php

[122] Sloan Digital Sky Survery III, "Photometric redshift estimates for galaxies sample sql queries," 2014, [Online; Copyright2010-2013 SDSS III; DR12 URL: http://skyserver.sdss.org/dr12/en/help/docs/realquery.aspx#photoz]. [Online]. Available: http://skyserver.sdss.org/dr10/en/help/docs/realquery.aspx#photoz

[123] A. Kleiner, A. Talwalkar, P. Sarkar, and M. I. Jordan, "A Scalable Bootstrap for Massive Data," *ArXiv e-prints*, Dec. 2011.

[124] P. J. Bickel, F. Gotze, and W. R. van Zwet, "Resampling Fewer Than n Observations : Gains, Losses, and Remedies for Losses," *Statistica Sinica*, vol. 7, pp. 1–31, 1997. [Online]. Available: http://www3.stat.sinica.edu.tw/statistica/oldpdf/A7n11.pdf

[125] S. M. L. S. Nagarajan, Radhakrishnan, *Bayesian Networks in R with Applications in Systems Biology*. Burlington: Springer New York, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B978012219141150019X

[126] D. Barber, *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012. [Online]. Available: http://web4.cs.ucl.ac.uk/staff/D.Barber/textbook/131214.pdf

[127] F. Leisch, "Sweave: Dynamic generation of statistical reports using literate data analysis," in *Compstat 2002 — Proceedings in Computational Statistics*, W. Härdle and B. Rönz, Eds. Physica Verlag, Heidelberg, 2002, pp. 575–580, iSBN 3-7908-1517-9. [Online]. Available: http://www.stat.uni-muenchen.de/~leisch/Sweave

[128] F. Leisch, "Sweave, part I: Mixing R and LaTeX," *R News*, vol. 2, no. 3, pp. 28–31, December 2002. [Online]. Available: http://CRAN.R-project.org/doc/Rnews/

[129] F. Leisch, "Sweave and beyond: Computations on text documents," in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria*, K. Hornik, F. Leisch, and A. Zeileis, Eds., 2003, ISSN 1609-395X. [Online]. Available: http://www.R-project.org/conferences/DSC-2003/Proceedings/

[130] F. Leisch, "Sweave, part II: Package vignettes," *R News*, vol. 3, no. 2, pp. 21–24, October 2003. [Online]. Available: http://CRAN.R-project.org/doc/Rnews/

[131] E. Klarreich, "In search of bayesian inference," *Commun. ACM*, vol. 58, no. 1, pp. 21–24, Dec. 2014. [Online]. Available: http://doi.acm.org/10.1145/2686734

[132] J. Pearl, "Ucla cognitive systems laboratory," 2015. [Online]. Available: http://bayes.cs.ucla.edu/csl_papers.html

[133] J. Pearl, *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. Morgan Kaufmann, 1988. [Online]. Available: https://www.elsevier.com/books/probabilistic-reasoning-in-intelligent-systems/pearl/978-1-55860-479-7

[134] J. Pearl, "Fusion, propagation, and structuring in belief networks," *Artificial Intelligence*, vol. 29, no. 3, pp. 241 – 288, 1986. [Online]. Available: http://www.sciencedirect.com/science/article/pii/000437028690072X

[135] J. Pearl, "Belief networks revisited," *Artificial Intelligence*, vol. 59, no. 12, pp. 49 – 56, 1993. [Online]. Available: http://www.sciencedirect.com/science/article/pii/000437029390169C

[136] D. Heckerman, "A tutorial on learning with bayesian networks," Learning in Graphical Models, Tech. Rep., 1996. [Online]. Available: http://research.microsoft.com/en-us/um/people/heckerman/tutorial.pdf

[137] I. Ben-Gal, *Bayesian Networks*. John Wiley & Sons, Ltd, 2008. [Online]. Available: http://dx.doi.org/10.1002/9780470061572.eqr089

[138] J. Pearl, *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000. [Online]. Available: http://bayes.cs.ucla.edu/BOOK-99/book-toc.html

[139] S. Conrady and L. Jouffe, "Introduction to bayesian networks & bayesialab," *Bayesia White Paper*, 2013. [Online]. Available: http://www.bayesia.com/en/products/bayesialab/resources/tutorials/IntroBBN.php

[140] S. Conrady and L. Jouffe, "Knowledge discovery in the stock market - supervised and unsupervised learning with bayesialab," *Bayesia White Paper*, 2011. [Online]. Available: http://www.bayesia.com/en/products/bayesialab/resources/tutorials/StockMarket.php

[141] K. Murphy, "A brief introduction to graphical models and bayesian networks," 1998, (Earlier version appears at Murphy K (2001) The Bayes Net Toolbox for Matlab, Computing Science and Statistics, 33, 2001). [Online]. Available: https://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html

[142] E. Charniak, "Bayesian networks without tears: Making bayesian networks more accessible to the probabilistically unsophisticated," *AI Mag.*, vol. 12, no. 4, pp. 50–63, Nov. 1991. [Online]. Available: http://dl.acm.org/citation.cfm?id=122623.122716

[143] E. D. Feigelson and G. J. Babu, *Modern Statistical Methods for Astronomy : With R Applications*. Cambridge University Press, 2012. [Online]. Available: http://site.ebrary.com/lib/georgemason/detail.action?docID=10583257

[144] A. Moore, "Statistical data mining tutorials," 2008, tutorial Topics: Bayesian Networks, Learning Bayesian Networks, Inference in Bayesian Networks, Gaussian Bayes Classifiers, Cross-Validation. [Online]. Available: http://www.autonlab.org/tutorials/

[145] R. E. Neapolitan, *Learning Bayesian Networks*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2003. [Online]. Available: http://www.cs.technion.ac.il/~dang/books/Learning%20Bayesian%20Networks%28Neapolitan,%20Richard%29.pdf

[146] Y. Tang, Y. Wang, K. Cooper, and L. Li, "Towards big data bayesian network learning - an ensemble learning based approach," in *Big Data (BigData Congress), 2014 IEEE International Congress on*, June 2014, pp. 355–357.

[147] C. Bielza and P. Larrañaga, "Discrete bayesian network classifiers: A survey," *ACM Computing Survey*, vol. 47, no. 1, pp. 5:1–5:43, Jul. 2014. [Online]. Available: http://doi.acm.org/10.1145/2576868

[148] F. Flam, "The odds, continually updated," September 2014, [Online; posted 29 September, 2014]. [Online]. Available: http://www.nytimes.com/2014/09/30/science/the-odds-continually-updated.html

[149] P. Szwed and J. R. van Dorp, "A bayesian model for rare event risk assessment using expert judgment about paired scenario comparisons," in *ASEM National Conference Proceedings*, 2002, pp. 444–453. [Online]. Available: http://www.seas.gwu.edu/~dorpjr/Publications/ConferenceProceedings/ASEM2002Szwed%20-%20R1.pdf

[150] P. Szwed, J. R. van Dorp, J. Merrick, T. Mazzuchi, and A. Singh, "A bayesian paired comparison approach for relative accident probability assessment with covariate

information," *European Journal of Operational Research*, vol. 169, no. 1, pp. 157 – 177, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0377221704003911

[151] J. Merrick, V. Dinesh, A. Singh, J. van Dorp, and T. Mazzuchi, "Propagation of uncertainty in a simulation-based maritime risk assessment model utilizing bayesian simulation techniques," in *Simulation Conference, 2003. Proceedings of the 2003 Winter*, vol. 1, Dec 2003, pp. 449–455 Vol.1.

[152] J. van Dorp, T. Mazzuchi, and J. Garciduenas, "A comparison of accelerated life testing designs within a single bayesian inferential framework," in *Reliability and Maintainability Symposium, 2006. RAMS '06. Annual*, Jan 2006, pp. 208–214.

[153] F. Sambo, F. Ferrazzi, and R. Bellazzi, "12 - probabilistic modelling with bayesian networks," in *Modelling Methodology for Physiology and Medicine (Second Edition)*, second edition ed., E. C. Cobelli, Ed. Oxford: Elsevier, 2014, pp. 257 – 280. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780124115576000124

[154] A. Gagin and I. Levin, "A Bayesian approach to removal of incoherent scattering from neutron total-scattering data," *Journal of Applied Crystallography*, vol. 47, no. 6, pp. 2060–2068, Dec 2014. [Online]. Available: http://dx.doi.org/10.1107/S1600576714023796

[155] C. Su, A. Andrew, M. Karagas, and M. Borsuk, "Using bayesian networks to discover relations between genes, environment, and disease," *BioData Mining*, vol. 6, no. 1, 2013. [Online]. Available: http://dx.doi.org/10.1186/1756-0381-6-6

[156] V. ZHARKOVA, S. IPSON, A. BENKHALIL, and S. ZHARKOV, "Feature recognition in solar images," *Artificial Intelligence Review*, vol. 23, no. 3, pp. 209–266, 2005. [Online]. Available: http://dx.doi.org/10.1007/s10462-004-4104-4

[157] R. Yacef, M. Benghanem, and A. Mellit, "Prediction of daily global solar irradiation data using bayesian neural network: A comparative study," *Renewable Energy*, vol. 48, no. 0, pp. 146 – 154, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0960148112002777

[158] C. Voyant, C. Darras, M. Muselli, C. Paoli, M.-L. Nivet, and P. Poggi, "Bayesian rules and stochastic models for high accuracy prediction of solar radiation," *Applied Energy*, vol. 114, no. 0, pp. 218 – 226, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0306261913007988

[159] D. Yu, X. Huang, H. Wang, Y. Cui, Q. Hu, and R. Zhou, "Short-term solar flare level prediction using a bayesian network approach," *apjl*, vol. 710, no. 1, p. 869, 2010. [Online]. Available: http://stacks.iop.org/0004-637X/710/i=1/a=869

[160] T. Hauser, A. Keats, and L. Tarasov, "Artificial neural network assisted bayesian calibration of climate models," *Climate Dynamics*, vol. 39, no. 1-2, pp. 137–154, 07 2012, date revised - 2013-05-01; Last updated - 2013-09-26; SubjectsTermNotLitGenreText - Climate models; Neural networks; Acoustic waves; General circulation models; Humidity; Climatology; Statistical forecasting; Noise

pollution; Future climates; Monte Carlo simulation; Artificial intelligence; Sulfur dioxide; Climate; Temperature; Noise levels; Seasonal variations. [Online]. Available: http://search.proquest.com/docview/1222991229?accountid=14541

[161] M. Catenacci and C. Giupponi, "Potentials and limits of bayesian networks to deal with uncertainty in the assessment of climate change adaptation policies," *Fondazione Eni Enrico Mattei (FEEM)¿Sustainable Development Papers*, 01 2010.

[162] F. N. A., S. S. Qian, H. W. Paerl, K. H. Reckhow, and E. A. Albright, "A study of anthropogenic and climatic disturbance of the new river estuary using a bayesian belief network," *Marine Pollution Bulletin*, vol. 83, no. 1, pp. 107 – 115, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0025326X14002148

[163] A. STASSOPOULOU, M. PETROU, and J. KITTLER, "Application of a bayesian network in a gis based decision making system," *International Journal of Geographical Information Science*, vol. 12, no. 1, pp. 23–46, 1998. [Online]. Available: http://dx.doi.org/10.1080/136588198241996

[164] T. Black and W. Thompson, "Bayesian data analysis," *Computing in Science Engineering*, vol. 3, no. 4, pp. 86–91, Jul 2001.

[165] J. Cheng and R. Greiner, "Comparing bayesian network classifiers," in *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, ser. UAI'99. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999, pp. 101–108. [Online]. Available: http://dl.acm.org/citation.cfm?id=2073796.2073808

[166] C. Huang and A. Darwiche, "Inference in belief networks: A procedural guide," *International Journal of Approximate Reasoning*, vol. 15, no. 3, pp. 225 – 263, 1996. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0888613X96000692

[167] D. Poznanski, D. Maoz, and A. Gal-Yam, "Bayesian single-epoch photometric classification of supernovae," *The Astronomical Journal*, vol. 134, no. 3, p. 1285, 2007. [Online]. Available: http://stacks.iop.org/1538-3881/134/i=3/a=1285

[168] E. Cameron and A. N. Pettitt, "Approximate bayesian computation for astronomical model analysis: a case study in galaxy demographics and morphological transformation at high redshift," *Monthly Notices of the Royal Astronomical Society*, vol. 425, no. 1, pp. 44–65, 2012. [Online]. Available: http://mnras.oxfordjournals.org/content/425/1/44.abstract

[169] G. A. Blanc, L. Kewley, F. P. A. Vogt, and M. A. Dopita, "IZI: Inferring the Gas Phase Metallicity (Z) and Ionization Parameter (q) of Ionized Nebulae Using Bayesian Statistics," *apjl*, vol. 798, p. 99, Jan. 2015.

[170] D. M. Chickering, "Learning Bayesian Networks is NP-Complete," pp. 121–130, 1996. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.21.1322&rep=rep1&type=pdf

[171] D. Maxwell Chickering, C. Meek, and D. Heckerman, "Large-Sample Learning of Bayesian Networks is NP-Hard," *Journal of Machine Learning Research*, vol. 5, pp. 1287–1330, 2004. [Online]. Available: http://jmlr.csail.mit.edu/papers/volume5/chickering04a/chickering04a.pdf

[172] C. Meek, "Finding a Path is Harder than Finding a Tree," *Journal Of Artificial Intelligence Research*, vol. 15, pp. 383–389, Jun. 2001.

[173] I. Tsamardinos, L. E. Brown, and C. F. Aliferis, "The max-min hill-climbing bayesian network structure learning algorithm," *Mach. Learn.*, vol. 65, no. 1, pp. 31–78, Oct. 2006. [Online]. Available: http://dx.doi.org/10.1007/s10994-006-6889-7

[174] G. F. Cooper, "The computational complexity of probabilistic inference using bayesian belief networks," *Artificial Intelligence*, vol. 42, no. 23, pp. 393 – 405, 1990. [Online]. Available: http://www.sciencedirect.com/science/article/pii/000437029090060D

[175] P. Dagum and M. Luby, "Approximating probabilistic inference in bayesian belief networks is np-hard," *Artificial Intelligence*, vol. 60, no. 1, pp. 141 – 153, 1993. [Online]. Available: http://www.sciencedirect.com/science/article/pii/000437029390036B

[176] S. Aji and R. McEliece, "The generalized distributive law," *Information Theory, IEEE Transactions on*, vol. 46, no. 2, pp. 325–343, Mar 2000.

[177] F. Kschischang, B. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *Information Theory, IEEE Transactions on*, vol. 47, no. 2, pp. 498–519, Feb 2001.

[178] M. A. Peot and R. D. Shachter, "Fusion and propagation with multiple observations in belief networks," *Artificial Intelligence*, vol. 48, no. 3, pp. 299 – 318, 1991. [Online]. Available: http://www.sciencedirect.com/science/article/pii/000437029190030N

[179] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul, "An introduction to variational methods for graphical models," *Machine Learning*, vol. 37, no. 2, pp. 183–233, 1999. [Online]. Available: http://dx.doi.org/10.1023/A%3A1007665907178

[180] M. Scutari, "Learning bayesian networks with the bnlearn r package," *Journal of Statistical Software*, vol. 35, no. 3, pp. 1–22, 7 2010. [Online]. Available: http://www.jstatsoft.org/v35/i03

[181] E. de Jonge and M. van der Loo, *An introduction to data cleaning with R.*

[182] M. Scutari and J.-B. Denis, *Bayesian Networks: With Examples in R.*

[183] D. Margaritis, "Learning bayesian network model structure from data," CMU, Tech. Rep., 2003. [Online]. Available: https://www.cs.cmu.edu/~dmarg/Papers/PhD-Thesis-Margaritis.pdf

[184] I. Tsamardinos, C. Aliferis, A. Statnikov, and E. Statnikov, "Algorithms for large scale markov blanket discovery," in *In The 16th International FLAIRS Conference, St.* AAAI Press, 2003, pp. 376–380.

[185] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, pp. 241–259, 1992.

[186] A. Tscher, M. Jahrer, and R. M. Bell, "The bigchaos solution to the netflix grand prize," 2009.

[187] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001. [Online]. Available: http://dx.doi.org/10.1023/A:1010933404324

[188] B. Quost, M.-H. Masson, and T. Denux, "Classifier fusion in the dempstershafer framework using optimized t-norm based combination rules," *International Journal of Approximate Reasoning*, vol. 52, no. 3, pp. 353 – 374, 2011, dependence Issues in Knowledge-Based Systems. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0888613X10001568

[189] D. Ruta and B. Gabrys, "New measure of classifier dependency in multiple classifier systems," in *Proc. of the 3rd International Workshop on Multiple Classifier Systems, number 2364 in Lecture Notes in Computer Science*. Springer Verlag, 2002, pp. 127–136.

[190] UBC Department of Computer Science's Distinguished Lecture Series, May 30, 2013, "Geoff Hinton - Recent Developments in Deep Learning," 2013. [Online]. Available: https://www.youtube.com/watch?v=vShMxxqtDDs

[191] A. Scoică, "Profile geoffrey hinton: Unlocking the language of the brain," *XRDS*, vol. 21, no. 1, pp. 60–61, Oct. 2014. [Online]. Available: http://doi.acm.org/10.1145/2667635

[192] Deep Learning, "Deep Learning," 2015. [Online]. Available: http://deeplearning.net

[193] Yann LeCun Talk @ John Hopkins University, Center for Language and Speed Processing, Nov 18, 2014, "The Unreasonable Effectiveness of Deep Learning," 2014. [Online]. Available: https://www.youtube.com/watch?v=sc-KbuZqGkI

[194] Joe Sill, Ensemble team, Netflix Challenge 2006-2009, "Advances in Ensemble Learning from the Netflix Prize Competition - San Francisco Bay Area Professional Chapter of the ACM Talk," 2010. [Online]. Available: https://www.youtube.com/watch?v=coeak1YsaYc

[195] D. Yu, G. Hinton, N. Morgan, J.-T. Chien, and S. Sagayama, "Introduction to the special section on deep learning for speech and language processing," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 4–6, Jan 2012.

[196] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, Nov 2012.

[197] H. Larochelle, "YouTube Lectures 1-9 On Neural Networks : Deep Learning," 2013. [Online]. Available: https://www.youtube.com/watch?v=vXMpKYRhpmI

[198] R. Salakhutdinov and G. Hinton, "An efficient learning procedure for deep boltzmann machines," *Neural Computation*, vol. 24, no. 8, pp. 1967–2006, Aug 2012.

[199] R. Livni, S. Shalev-Shwartz, and O. Shamir, "An Algorithm for Training Polynomial Networks," *ArXiv e-prints*, Apr. 2013.

[200] J. M. Lotz, H. C. Ferguson, L. Armus, L. F. Barrientos, J. G. Bartlett, M. Blanton, K. D. Borne, C. R. Bridge, M. Dickinson, H. Francke, G. Galaz, E. Gawiser, K. Gilmore, R. H. Lupton, J. A. Newman, N. D. Padilla, B. E. Robertson, R. Roskar, A. Stanford, and R. H. Wechsler, "Galaxy Evolution with LSST," 2010. [Online]. Available: http://www.lsst.org/files/docs/aas/2010/215-RC-963-AAS_Lotz.pdf

[201] J. M. Lotz, P. Jonsson, T. J. Cox, and J. R. Primack, "Galaxy merger morphologies and time-scales from simulations of equal-mass gas-rich disc mergers," *Monthly Notices of the Royal Astronomical Society*, vol. 391, no. 3, pp. 1137–1162, 2008. [Online]. Available: http://mnras.oxfordjournals.org/content/391/3/1137.abstract

[202] K. Kovač, S. J. Lilly, O. Cucciati, C. Porciani, A. Iovino, G. Zamorani, P. Oesch, M. Bolzonella, C. Knobel, A. Finoguenov, Y. Peng, C. M. Carollo, L. Pozzetti, K. Caputi, J. D. Silverman, L. A. M. Tasca, M. Scodeggio, D. Vergani, N. Z. Scoville, P. Capak, T. Contini, J.-P. Kneib, O. Le Fèvre, V. Mainieri, A. Renzini, S. Bardelli, A. Bongiorno, G. Coppa, S. de la Torre, L. de Ravel, P. Franzetti, B. Garilli, L. Guzzo, P. Kampczyk, F. Lamareille, J.-F. Le Borgne, V. Le Brun, C. Maier, M. Mignoli, R. Pello, E. Perez Montero, E. Ricciardelli, M. Tanaka, L. Tresse, E. Zucca, U. Abbas, D. Bottini, A. Cappi, P. Cassata, A. Cimatti, M. Fumana, A. M. Koekemoer, D. Maccagni, C. Marinoni, H. J. McCracken, P. Memeo, B. Meneux, and R. Scaramella, "The Density Field of the 10k zCOSMOS Galaxies," *apjl*, vol. 708, pp. 505–533, Jan. 2010.

[203] Y. Wang, "Model-independent distance measurements from gamma-ray bursts and constraints on dark energy," *Physical Review D*, vol. 78, no. 12, p. 123532, Dec. 2008.

[204] Y. Wang, "A Model-independent Photometric Redshift Estimator for Type Ia Supernovae," *apjl*, vol. 654, pp. L123–L125, Jan. 2007.

[205] B. Ménard, R. Scranton, S. Schmidt, C. Morrison, D. Jeong, T. Budavari, and M. Rahman, "Clustering-based redshift estimation: method and application to data," *ArXiv e-prints*, Mar. 2013.

[206] C. K. Davis, "Beyond data and analysis," *Commun. ACM*, vol. 57, no. 6, pp. 39–41, Jun. 2014. [Online]. Available: http://doi.acm.org/10.1145/2602326

[207] J. L. King and P. F. Uhlir, "Soft infrastructure challenges to scientific knowledge discovery," *Commun. ACM*, vol. 57, no. 9, pp. 35–37, Sep. 2014. [Online]. Available: http://doi.acm.org/10.1145/2644279

[208] C. Staff, "Visualizations make big data meaningful," *Commun. ACM*, vol. 57, no. 6, pp. 19–21, Jun. 2014. [Online]. Available: http://doi.acm.org/10.1145/2601074

[209] Tung, Wei-Fung and Yuan, Soe-Tsyr, "Intelligent Service Machine," *Commun. ACM*, vol. 53, no. 8, pp. 129–134, Aug. 2010. [Online]. Available: http://doi.acm.org/10.1145/1787234.1787268

[210] U. B. Simons Institute for the Theory of Computing, "Theoretical foundations of big data analysis," 2013. [Online]. Available: http://simons.berkeley.edu/programs/bigdata2013

[211] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, Jan. 2008. [Online]. Available: http://doi.acm.org/10.1145/1327452.1327492

[212] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, ser. DMKD '03. New York, NY, USA: ACM, 2003, pp. 2–11. [Online]. Available: http://doi.acm.org/10.1145/882082.882086

[213] J. Lin, E. Keogh, L. Wei, and S. Lonardi, "Experiencing sax: A novel symbolic representation of time series," *Data Min. Knowl. Discov.*, vol. 15, no. 2, pp. 107–144, Oct. 2007. [Online]. Available: http://dx.doi.org/10.1007/s10618-007-0064-z

[214] J. Ouellette, "How Quantum Computers and Machine Learning Will Revolutionize Big Data," 2013. [Online]. Available: http://www.wired.com/2013/10/computers-big-data/all/

[215] S. Lloyd, "Quantum Machine Learning - Google Tech Talks," 2014. [Online]. Available: https://www.youtube.com/watch?v=wkBPp9UovVU

[216] TheAstrophysicsSpectator, "The astrophysics spectator - the structure of our universe," 2009, cosmology History. [Online]. Available: http://www.astrophysicsspectator.com/topics/overview/

[217] NASA/WMAP Science Team, "NASA Wilkinson Microwave Anisotropy Probe (WMAP) Homepage," 2013. [Online]. Available: http://map.gsfc.nasa.gov/

[218] S. Dodelson, Ed., *Modern Cosmology*. Burlington: Academic Press, 2003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B978012219141150019X

[219] NASA/WMAP Science Team, "NASA Wilkinson Microwave Anisotropy Probe (WMAP) Nine Year Microwave Sky Image," 2014. [Online]. Available: http://map.gsfc.nasa.gov/media/121238/index.html

[220] NASA/WMAP Science Team, "NASA Wilkinson Microwave Anisotropy Probe (WMAP) Timeline of the Universe Image," 2012. [Online]. Available: http://map.gsfc.nasa.gov/media/060915/index.html

[221] D. Eisenstein, "Dark energy and cosmic sound," *New Astronomy Reviews*, vol. 49, no. 79, pp. 360 – 365, 2005, wide-Field Imaging from Space Conference on Wide-Field Imaging from Space. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1387647305000850

[222] G. J. D. and J. Wojtusiak, "A natural induction approach to traffic prediction for autonomous agent-based vehicle route planning," George Mason University, Fairfax, VA, Tech. Rep. MLI 08-1, Feb 2008.

Biography

Pragyansmita Nayak graduated from BJB College, Bhubaneswar, India in 1995 (Science stream with minor in Electronics). She received her Bachelor of Engineering (Honors) degree in Computer Science from BITS Pilani, India in 1999. She was employed as a Software Engineer in Wipro Global R&D, Bangalore, India for one year. She worked for three years as an Research Analyst at the ERNET PoP, IIT Madras, India while pursuing her Master of Science (Research) degree in Computer Science. She has been working at CGI Federal Inc. from 2004 till date as a Senior Consultant. She works with the Momentum Financial® Application as part of the Product Development group.