



January 2018

Automatic Approach To Morphological Classification Of Galaxies With Analysis Of Galaxy Populations In Clusters

Madina Renatovna Sultanova

Follow this and additional works at: <https://commons.und.edu/theses>

Recommended Citation

Sultanova, Madina Renatovna, "Automatic Approach To Morphological Classification Of Galaxies With Analysis Of Galaxy Populations In Clusters" (2018). *Theses and Dissertations*. 2358.
<https://commons.und.edu/theses/2358>

This Dissertation is brought to you for free and open access by the Theses, Dissertations, and Senior Projects at UND Scholarly Commons. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of UND Scholarly Commons. For more information, please contact zeinebyousif@library.und.edu.

AUTOMATIC APPROACH TO MORPHOLOGICAL CLASSIFICATION OF
GALAXIES WITH ANALYSIS OF GALAXY POPULATIONS IN CLUSTERS

by

Madina Renatovna Sultanova

Bachelor of Science, Saint Cloud State University, 2013

A Dissertation

Submitted to the Graduate Faculty

of the

University of North Dakota

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Grand Forks, North Dakota

August

2018

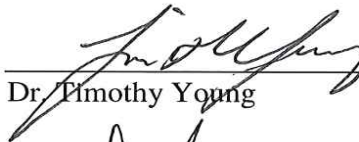
This dissertation, submitted by Madina Renatovna Sultanova in partial fulfillment of the requirements for the Degree of Doctor of Philosophy from the University of North Dakota, has been read by the Faculty Advisory Committee under whom the work has been done and is hereby approved.



Dr. Wayne Barkhouse



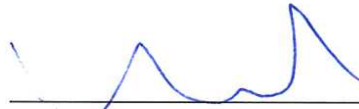
Dr. Kanishka Marasinghe



Dr. Timothy Young

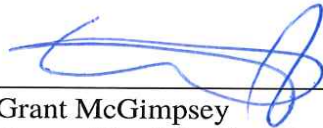


Dr. Yen Lee Loh



Dr. Ronald Marsh

This dissertation is being submitted by the appointed advisory committee as having met all of the requirements of the School of Graduate Studies at the University of North Dakota and is hereby approved.



Grant McGimpsey
Dean of the School of Graduate Studies

July 11, 2018

Date

PERMISSION

Title Automatic Approach to Morphological Classification of Galaxies With
 Analysis of Galaxy Populations in Clusters

Department Physics and Astrophysics

Degree Doctor of Philosophy

In presenting this dissertation in partial fulfillment of the requirements for a graduate degree from the University of North Dakota, I agree that the library of this University shall make it freely available for inspection. I further agree that permission for extensive copying for scholarly purposes may be granted by the professor who supervised my dissertation work or, in their absence, by the Chairperson of the department or the dean of the School of Graduate Studies. It is understood that any copying or publication or other use of this dissertation or part thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of North Dakota in any scholarly use which may be made of any material in my dissertation.

Madina Renatovna Sultanova
June 28, 2018

TABLE OF CONTENTS

| | |
|------------------------------------------------------------|------|
| LIST OF FIGURES | vii |
| LIST OF TABLES | xi |
| ACKNOWLEDGEMENTS | xii |
| DEDICATION | xiii |
| ABSTRACT | xiv |
| CHAPTER | |
| I. INTRODUCTION | 1 |
| 1.1 The Discovery of Galaxies | 2 |
| 1.2 Classification Systems | 9 |
| 1.2.1 The Hubble Tuning Fork | 10 |
| 1.2.2 de Vaucouleurs Revised System | 17 |
| 1.2.3 Morgan's Galaxy Classification System | 19 |
| 1.3 Formation and Evolution of Galaxies | 21 |
| 1.3.1 The Big Bang Theory and the Early Universe | 24 |
| 1.3.2 Relating to Cosmology | 31 |
| 1.3.3 Evolution Mechanisms and Models | 32 |
| II. MORPHOLOGY SOFTWARE | 37 |
| 2.1 Reasons for Automation | 37 |
| 2.2 Images and Charge Coupled Devices | 40 |
| 2.3 Description of the Software | 47 |
| 2.4 Parameters | 51 |

| | | |
|-------|------------------------------------------------------|-----|
| 2.4.1 | Central Concentration | 52 |
| 2.4.2 | Asymmetry | 55 |
| 2.4.3 | Gini Coefficient | 57 |
| 2.4.4 | Theil Index | 61 |
| 2.4.5 | M20 | 63 |
| 2.4.6 | B/D and B/T Ratios | 65 |
| III. | TRAINING DATA SETS | 69 |
| 3.1 | SDSS and Galaxy Zoo: Early Tests | 69 |
| 3.2 | EFIGI Catalog | 71 |
| 3.2.1 | Description | 71 |
| 3.2.2 | Analysis | 73 |
| IV. | HIGH-REDSHIFT DATA | 97 |
| 4.1 | High-redshift CFHT Clusters | 97 |
| 4.1.1 | Analysis | 99 |
| 4.1.2 | Nucleated vs. Non-nucleated Dwarf Galaxies | 112 |
| V. | LOW-REDSHIFT DATA | 119 |
| 5.1 | Low-Redshift Abell Clusters | 119 |
| 5.1.1 | Analysis | 122 |
| 5.2 | WINGS | 126 |
| 5.2.1 | Analysis | 128 |
| VI. | PRINCIPAL COMPONENT ANALYSIS | 134 |
| 6.1 | Introduction | 134 |
| 6.1.1 | Theory | 135 |
| 6.1.2 | When to use PCA | 140 |
| 6.1.3 | PCA using R for the EFIGI | 141 |
| 6.1.4 | How Many Principal Components to Keep? | 145 |

| | | |
|-----------------------------|----------------------------------------------------|-----|
| 6.2 | PCA for CFHT Data Set | 151 |
| 6.3 | PCA for KPNO Data Set | 153 |
| 6.4 | PCA for WINGS Data Set | 156 |
| 6.5 | Minitab: Principle Component Analysis | 159 |
| VII. DISCUSSION | | 161 |
| 7.1 | Success of Morphology Software | 161 |
| 7.2 | Exploring Galaxy Formation and Evolution | 162 |
| 7.3 | Exploring Dwarf Galaxies | 168 |
| VIII. CONCLUSIONS | | 170 |
| 8.1 | Future Work | 172 |
| 8.2 | Final Thoughts | 174 |
| APPENDICES | | 176 |
| BIBLIOGRAPHY | | 180 |

LIST OF FIGURES

| Figure | | Page |
|--------|---------------------------------------------------------------------------------|------|
| 1 | William Herschel’s diagram of the Milky Way | 5 |
| 2 | E and Irr galaxies from <i>The Realm of the Nebula</i> | 11 |
| 3 | Spiral galaxies from <i>The Realm of the Nebula</i> | 12 |
| 4 | The 1936 version of the Hubble tuning-fork diagram. | 15 |
| 5 | The 3D representation of de Vaucouleurs scheme. | 18 |
| 6 | The cosmic timeline of the expanding Universe. | 26 |
| 7 | CMB as imaged by the COBE, WMAP, and Planck satellites. | 29 |
| 8 | Numerical simulation of merging spiral galaxies. | 33 |
| 9 | Galaxy formation and evolution models. | 35 |
| 10 | Sketch and HST images of M51 and NGC 5195 galaxies. | 41 |
| 11 | An analogy diagram for CCDs. | 44 |
| 12 | Flowchart of the morphological software. | 48 |
| 13 | Plot of $\text{Log}(A)$ vs. $\text{Log}(C)$ for Hubble Deep Field data. | 57 |
| 14 | Graphical representation of the Lorenz Curve. | 58 |
| 15 | The Gini vs. Theil plot of EFIGI data | 62 |
| 16 | Examples of various SDSS images used to test and train software. | 70 |
| 17 | Examples of various EFIGI images used to test and train software. | 72 |
| 18 | Histogram of EFIGI Hubble Types as a function of C and A | 75 |
| 19 | Histogram of EFIGI Hubble Types as a function of B/T and M20. | 76 |
| 20 | Histogram of EFIGI Hubble Types as a function of Gini and Theil. | 77 |

| | | |
|----|------------------------------------------------------------------------------------------|-----|
| 21 | Relations between the five nonparametric values (EFIGI data) | 78 |
| 22 | A versus Gini plot of EFIGI data. | 79 |
| 23 | A versus Gini plot of EFIGI data with classification regions. | 80 |
| 24 | A versus Theil plot of EFIGI data with classification regions. | 82 |
| 25 | C versus B/T ratio of 984 galaxies from the SDSS red sequence. | 83 |
| 26 | C versus B/T ratio for 815 galaxies from the EFIGI data. | 83 |
| 27 | C versus Gini plot of EFIGI data. | 86 |
| 28 | Gini versus C plot of SDSS EDR data using the g and i bands. | 87 |
| 29 | C versus Theil plot of EFIGI data. | 89 |
| 30 | $\text{Log}(A)$ versus $\text{Log}(C)$ plot of EFIGI data. | 90 |
| 31 | C versus A plot of EFIGI data. | 91 |
| 32 | EFIGI on Gini vs. Theil plane before and after filtering. | 93 |
| 33 | The C versus Gini \rightarrow Gini versus Theil planes on EFIGI. | 95 |
| 34 | CFHT galaxy postage stamps. | 98 |
| 35 | Relations between five parameters in 421 galaxies from CFHT. | 100 |
| 36 | Histogram of Hubble Types as a function of C and A (CFHT data). | 102 |
| 37 | Histogram of CFHT Hubble Types as a function of Gini and Theil. | 103 |
| 38 | A vs. G and A vs. Theil plots of CFHT galaxies. | 104 |
| 39 | C versus A plot of 421 CFHT galaxies. | 105 |
| 40 | C versus Gini plot of 421 CFHT galaxies. | 106 |
| 41 | C versus Theil plot of 421 CFHT galaxies. | 107 |
| 42 | The A versus Gini \rightarrow A versus Theil planes for 421 CFHT galaxies. | 108 |
| 43 | Histogram of dwarf CFHT galaxies as a function of C | 115 |
| 44 | Ratio of high- C vs. low- C dwarf galaxies as a function of (r/r_{200}) | 117 |
| 45 | KPNO galaxy postage stamps. | 120 |
| 46 | Relations between five parameters for galaxies from KPNO. | 121 |

| | | |
|----|-----------------------------------------------------------------------------------|-----|
| 47 | Histogram of Hubble Types as a function of C and A (KPNO data). | 122 |
| 48 | Histogram of KPNO Hubble Types as a function of Gini and Theil. | 123 |
| 49 | A vs. Gini and C vs. A plot of KPNO galaxies. | 124 |
| 50 | The A versus Gini \rightarrow A versus Theil planes for 259 KPNO galaxies. | 125 |
| 51 | WINGS galaxy postage stamps. | 127 |
| 52 | Relations between five parameters for galaxies from WINGS | 129 |
| 53 | Histogram of Hubble Types as a function of C and A (WINGS data). | 130 |
| 54 | Histogram of WINGS Hubble Types as a function of Gini and Theil. | 131 |
| 55 | A vs. Gini and C vs. A plot of WINGS galaxies. | 132 |
| 56 | The A versus Gini \rightarrow A versus Theil planes for 600 WINGS galaxies. | 133 |
| 57 | Graphical representation of PCA. | 140 |
| 58 | Producing the principle components. | 144 |
| 59 | Scree plot of principal components of the EFIGI data. | 146 |
| 60 | Pareto chart of principal components of the EFIGI data. | 148 |
| 61 | Biplot diagram of the first two principal components in EFIGI. | 149 |
| 62 | Scree and pareto charts for the CFHT data. | 152 |
| 63 | Biplot diagram of the first two principal components in CFHT. | 153 |
| 64 | Scree and pareto charts for the KPNO data. | 154 |
| 65 | Biplot plot of the first two principal components in KPNO. | 155 |
| 66 | Scree and pareto charts for the WINGS data. | 157 |
| 67 | Biplot diagram of the first two principal components in WINGS. | 158 |
| 68 | PCA analysis for CFHT, KPNO, and WINGS data sets | 160 |
| 69 | Galaxy type vs. clustercentric radius and density (Dressler 1980). | 163 |
| 70 | Galaxy type vs. clustercentric radius (Witmore & Gilmore 1993). | 164 |
| 71 | The morphology-radius relation from Goto <i>et al.</i> 2003. | 165 |
| 72 | CFHT morphology as a function of (r/r_{200}) | 166 |

73 Dwarf galaxy morphology-density relation (Ferguson *et al.* 1990). . . 168

LIST OF TABLES

| Table | | Page |
|-------|---------------------------------------------------------------------------|------|
| 1 | Herschel’s “nebula” classification system | 10 |
| 2 | The de Vaucouleurs galaxy classification | 18 |
| 3 | The revised Yerkes classification system | 20 |
| 4 | Information of various test SDSS images. | 70 |
| 5 | The EFIGI Hubble Sequence (EHS). | 73 |
| 6 | 15 CFHT clusters from Rude (2015; <i>et al.</i> 2018 in preparation). . . | 99 |
| 7 | Morphological classification and GALFIT analysis of CFHT clusters. | 111 |

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to my advisor, Dr. Wayne Barkhouse, for his continuous guidance and advice during the course of my doctoral studies. I also want to thank my committee members: Dr. Kanishka Marasinghe, Dr. Timothy Young, Dr. Yen Lee Loh, and Dr. Ronald Marsh for their interest in my research and for taking time to serve on my dissertation committee.

My sincere thanks also goes to the University of North Dakota Department of Physics and Astrophysics and the North Dakota NASA Established Program to Stimulate Competitive Research (EPSCoR) for providing financial support for my research.

To my family: Renat, Lucy, and Liya,
for their encouragement and support.

ABSTRACT

The classification of galaxies based on their morphology (*i.e.* structural properties) is a field in astrophysics that aims to understand galaxy formation and evolution based on their physical differences. Whether structural differences are due to internal factors or a result of local environment, the dominate mechanism that determines galaxy type needs to be robustly quantified in order to have a thorough grasp of the origin of the different types of galaxies (e.g., elliptical, S0, spiral, and irregular). The main subject of this thesis is to explore the use of computers to automatically analyze and classify large numbers of galaxies based on their morphology, and to analyze subsamples of galaxies selected by type to understand galaxy formation and evolution in various environments. I have developed computer software to classify galaxies by measuring specific parameters extracted from digital images. In particular, I have constructed computer algorithms to calculate five classification parameters for a list of galaxies in a single FITS image. This research has important implications for increasing our knowledge of galaxy formation and evolution in dense systems. A diverse range of data sets is studied, primarily focusing on: Rude (2015), Barkhouse *et al.* (2007), WINGS (Fasano *et al.* 2006), and Baillard *et al.* (2011). The data sets include galaxies from a wide range of redshifts, from $0.03 \leq z \leq 0.20$. The different span of redshift allows for comparison of distant clusters with those nearby in order to look for evolutionary changes in the galaxy cluster population.

CHAPTER I

INTRODUCTION

Galaxies are large systems spanning thousands of light years and consisting of millions of stars gravitationally bound together with gas, dust, and dark matter. Galaxies can range from dwarfs of approximately tens of millions of stars to giants with trillions of stars. They can stretch from thousands of parsecs to hundreds of thousands of parsecs in size. Besides their different sizes, galaxies also vary in shape. Classification of galaxies based on their shapes, *i.e.* morphological properties, is a field in astronomy that focuses on organizing galaxies based on their physical differences. The study of galaxy morphology aims to understand how each classification type may (or may not) be related to another. Through the study of galaxies' structure, we can infer the physical processes that are responsible for galaxy formation and evolution, as well as answer questions such as what causes these morphological differences — whether the changes are caused by internal factors of the galaxies themselves or by their surrounding environment (or a combination thereof) — all of which are important questions about galactic research in the field of astrophysics.

Galaxies are the building-blocks of the large-scale structure of the Universe, therefore, a better understanding of the properties of galaxies and their evolution can improve our knowledge of the Universe as a whole. From observations, it is evident that distant galaxies (*i.e.* galaxies at high redshift) have different structural compositions than ones at lower redshift, therefore, studying distant galaxies allows us to look at the Universe at an earlier time and thus help to uncover clues of galaxy formation

and evolution. The study of galaxy morphology has been gaining attention since the 1920's. In this chapter, we discuss the history of galactic studies, as well as the current visual classification systems and current theories of galaxy formation and evolution.

1.1 The Discovery of Galaxies

The development of the notion that our Solar System — and later, the Milky Way — is not the complete Universe but a system within a larger structure was gradual. Before any attention could be directed towards other galaxies, the Milky Way was the main focus of astronomical studies. This bright band of light scattered across the night sky has fascinated people across the world for thousands of years. The name, “Milky Way,” derives from the Greek “galaxias k'uklos” (*galaxias kyklos*, literally meaning “milky circle” in English), where the word “g'ala” (*gála*) is “milk”. The Romans translated the Greek name to “via lactea” — for which the literal translation is “the road of milk” in English. Thus, the Milky Way is the English translation of the Latin “via lactea.”

Democritus (460 — 370 BC) was one of the first philosophers recorded to propose that this band must be made up of stars. His description of the Milky Way can be found in Plutarch's *Moralia*, a collection of essays and speeches published in 100 AD. In *Moralia*, the Milky Way is described as “a cloudy circle, which continually appears in the air, and by reason of the whiteness of its colors is called the galaxy, or the milky way” (Plutarch 1878). It further states that Democritus proposed the idea that the Milky Way “[...] is the splendor which ariseth from the coalition of many small bodies, which, being firmly united amongst themselves, do mutually enlighten one another” (Plutarch 1878).

However, it wasn't until the seventeenth century that Galileo Galilei — an Italian

astronomer, physicist, philosopher, and mathematician — provided proof of this fact. Galileo Galilei was one of the first to develop a telescope for astronomical use. Though it is not known who can solely be credited for the creation of the telescope, it is evident that Galileo’s refinement of this device and its use significantly advanced astronomic knowledge (Timmons 2012). In his book, *The Starry Message*, published in 1610, Galileo wrote:

“I have observed the nature and the material of the Milky Way. With the aid of the telescope this has been scrutinized so directly and with such ocular certainty that all the disputes which have vexed philosophers through so many ages have been resolved, and we are at last freed from wordy debates about it. The galaxy is, in fact, nothing but a congeries of innumerable stars grouped together in clusters. Upon whatever part of it the telescope is directed, a vast crowd of stars is immediately presented to view. Many of them are rather large and quite bright, while the number of smaller ones is quite beyond calculation” (Galileo 1610).

Besides stars, other objects have been observed in the night sky. Unlike stars — which appear as bright, compact dots of light — these objects appear as small, faint, elliptical patches and were thus referred to as “nebula” stars, from the Latin word “nebula”, meaning “fog”. Galileo also writes:

“But it is not only in the Milky Way that whitish clouds are seen; several patches of similar aspect shine with faint light here and there throughout the aether, and if the telescope is turned upon any of these it confronts us with a tight mass of stars. And what is even more remarkable, the stars which have been called ‘nebulous’ by every astronomer up to this time turn out to be a group of very small stars arranged in a wonderful

manner” (Galileo 1610).

He believed (though perhaps not strongly) that these nebulosities could be resolved into systems of stars. However, not all agreed. Since not all nebula patches in the sky could be resolved into groups of stars, some regarded nebulosities as stars that became blurred due to an optical effect of the telescope. Some nebulosities showed no signs of star clusters at all (Whitney 1971).

By mid-1700’s, some astronomers, such as Abbe Nicolas Louis de La Caille, proposed that there existed various classes of nebula, and therefore, some nebulosities were “nothing but [...] vaguely terminated whitish space, more or less luminous and frequently of very irregular form,” while others comprised of stars and were “only nebula in appearance and to the unaided eye, but which one sees at the telescope as a cluster of distinct stars, quite close together” (Whitney 1971). But others, such as Thomas Wright, introduced their own theories of the Universe based on Galileo’s initial description of the nebulosities. Wright proposed that the Universe was a thin, spherical shell of stars distributed in a way that “fill[s] up the whole medium with a kind of regular irregularity of objects” (Berendzen *et al.* 1976), and the center of this shell was the location of a supernatural spirit. He supposed that the Milky Way is a flat, rotating structure of stars. In his book, *An Original Theory, Or New Hypothesis of the Universe*, he also speculated that there were other structures like the Milky Way, meaning, it is just one among many (Whitney 1971).

Later, the 18th-century philosopher Immanuel Kant expanded Wright’s idea that the Milky Way was a flat, rotating, collection of stars to suggest that it is a flat rotating *disk* of stars. In order to explain the nebula stars, Kant proposed that there may be other universes outside of our own. In *General History of Nature and Theory of the Heavens*, he writes that he regards the nebulosities as “being not such enormous

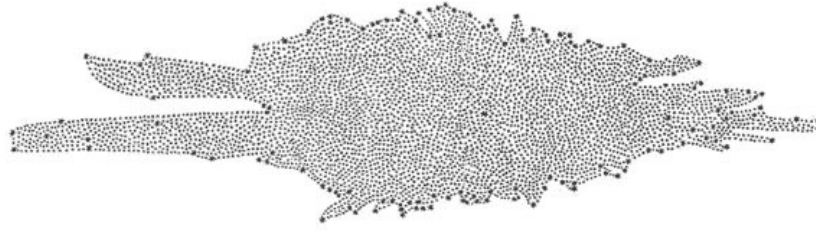


Figure 1: William Herschel’s diagram of the Milky Way published in *Philosophical Transactions of the Royal Society* in 1785. The Sun is located near the center of the system

single stars but systems of many stars” (Whitney 1971). His theory to describe these star systems, one of which is our own Milky Way, is frequently referred to as the “island universes” theory.

With further advancement of telescopes, astronomers were able to take a closer look at the nebula. William Herschel (1737 — 1822) was one of the leaders in the study of stellar systems and telescope development at the time. He devoted most of his life to the study of the “structure of the heavens” (de Vaucouleurs 1957). One of the telescopes he constructed was the “Great Forty-Foot” telescope — a 47-inch diameter primary mirror and 40-foot focal length reflecting telescope. With the assistance of his sister, Caroline, he observed and recorded over two thousand nebula stars (Whitney 1971). One of his other goals was to outline the structure of the Milky Way. He set an assumption that all stars had the same brightness, and therefore, stars that appeared faint had to be distant and bright stars had to be nearby. Then by counting the number of stars (which he called “star gauges”) in 683 regions of the sky, Herschel came up with a diagram representation of the Milky Way as seen in Figure 1 (Whitney 1971; Carroll 2006).

In order for the Milky Way to have the structure he recorded from his observations,

Herschel speculated that the Sun had to be located near the center of the Milky Way. Using the Great Forty-Foot, Herschel was able to resolve stars in various nebulosities. He also found a number of faint nebula stars in darker regions of the sky away from the Milky Way. He concluded that the nebulosities that could not be resolved must be distant. In 1785, he wrote:

“As we are used to call the appearance of the heavens, where it is surrounded with a bright zone, the Milky Way, it may not amiss to point out some other very remarkable nebula which cannot well be less, but are probably much larger than our own systems; and, being also extended, the inhabitants of the planets that attend the stars which compose them must likewise perceive the same phenomena. For which reason they may also be called milky ways by way of distinction” (Berendzen *et al.* 1976).

Herschel’s observations appeared to confirm Kant’s/Wright’s theory — that these objects were outside of the Milky Way system (Whitney 1971; de Vaucouleurs 1957).

Later, in 1845, William Parsons used a 72-inch diameter mirror and 52-foot focal length reflecting telescope to observe the nebula systems. This telescope — also called the “Leviathan of Parsonstown” — was an advancement from William Herschel’s Great Forty-Foot. Parsons discovered that a number of these objects possessed spiral structure and was able to resolve individual stars in some. The nature of these spirals was questioned — some believed them to be solar systems such as our own. Therefore, until about the early 20th century, there was still a debate about whether the so-called nebula stars were objects located within the Milky Way; and perhaps the Milky Way included all stellar objects (*i.e.* was the entirety of the Universe), or if the nebulae were separate island universes of their own (Berendzen *et al.* 1976; Hetherington 1993). At the time, there was still a widespread disagreement within

the astronomical community about the size and structure of the Milky Way, mainly because the question of how to accurately determine distances to stars and nebulae did not have a clear answer.

Discussion over the “scale of the universe” occurred during the National Academy of Sciences annual meeting at the Smithsonian Institution in Washington, D.C. in April of 1920, between two teams of astronomers led by Harlow Shapley and Heber Curtis. At the meeting, Shapley presented his argument that the Milky Way was tremendously large (having a diameter of about 100 kpc), therefore the nebulae had to be located within the Milky Way. His argument also proposed that our Solar System was located towards the edge of our galaxy/Universe. Meanwhile, Curtis claimed that the nebulae had to lay outside of the Milky Way, which he estimated had a diameter of 10 kpc, and the Sun was located at the center of our galaxy. This discussion is now referred to as the Shapley-Curtis debate or the “Great Debate”, though in 1969 Shapley writes: “I don’t think the word ‘debate’ was used at the time. Actually it was a sort of symposium, a paper by Curtis and a paper by me, and a rebuttal apiece” (Whitney 1971). This “symposium” presented two opposing theories about the Universe at the time. Though the issue of the real scale of the Universe was not solved right at that moment, the Great Debate remains as a record of the modern process of scientific thinking (Shu 1982).

The resolution to the Great Debate did not start until Edwin Hubble provided evidence from his observations that the nebula stars were in fact not part of the Milky Way galaxy but separate galaxies of their own (Hubble 1926). Using the 100-inch Hooker Telescope, Hubble used Cepheid variable stars to determine the distance to object Messier 31 (M31), otherwise known as the Andromeda galaxy.

Cepheid stars are variable stars (stars that change brightness over time) whose pulsation periods are proportional to their luminosities. The correlation between pul-

sation period and luminosity of these variable stars was made by Henrietta Leavitt, who observed that the longer the pulsation period, the greater the average luminosity of the star. From the period and the light curve (*i.e.* a graph of apparent magnitude versus time) of a cepheid star, its absolute magnitude and average apparent magnitude can be found, after which, the distance to that star can be calculated from the distance-modulus formula: $m - M = 5 * \text{Log}(d/10)$, where m , M , and d are the average apparent magnitude, absolute magnitude, and distance (in parsecs), respectively.

By resolving and identifying cepheid variable stars in M31, Hubble was able to calculate the distance to the galaxy as approximately 275 kpc (Hubble 1929a). Hubble's measurement placed M31 well outside of Shapley's estimate and thus offered strong support for Curtis' argument. Today, it is estimated that M31 is about 778 kpc away from the Milky Way.

The properties of nebulae (now called galaxies) have been studied extensively since. As in most fields of science, one of the first steps to understand the physics of a phenomena is to classify it in some order (Sandage 1961; Buta 2011). As astronomer Allan Sandage (1926 — 2010) writes:

“The master problem in cosmology is to understand the distribution and motions of galaxies as they relate to the origin and evolution of the universe. Two distinct approaches are possible and necessary. First, the stellar content of galaxies must be described, classified, and studied. The classification should relate class properties of the objects by finding a continuous sequence of forms. This is possible if the galaxies have really evolved and if both the old and the new forms exist at the present time. The problem is analogous to proving biological evolution by reading

the fossil record and classifying the bones in a continuous sequence. The second approach is a study of the way galaxies, as systems, define the large-scale distribution and motion of matter in the universe” (Sandage 1961).

After describing the characteristics and structures of a certain group of objects from observations, one can attempt to explain the physics of how they work. A meaningful classification system is one that is based on continuously varying parameters (*e.g.* luminosity, temperature, etc.) that can be related to theories that explain the physics of the phenomena. Various parameters continue to be explored and tested today. A classification system helps ease the identification of individual objects wherever they may be found, and paint a clearer picture of how different groups of such objects are interrelated. It is through classification that astronomers can build a better understanding of formation and evolution of galaxies.

1.2 Classification Systems

In the early 20th century, Edwin Hubble wrote about the knowledge of galaxies at the time: “Extremely little is known of the nature of nebula, and no significant classification has yet been suggested, not even a precise definition has been formulated” (Hubble 1920). At the time, astronomy was limited to the optical waveband, therefore galaxies were only studied visually or by their spectra. Visual classification of galaxies’ morphology consists of observing the galaxies using their optical band images (which were sensitive to the blue region of the light spectrum) and categorizing them based on the observer’s judgement of their appearance. During the late 18th century, Herschel classified the then-called nebulae in terms of their brightness and size using capital and lower-case letters as given in Table 1 (Berendzen *et al.* 1976).

Herschel’s son, John Herschel, later expanded this system.

Table 1: Herschel’s “nebula” classification system

| | |
|-----------|-----------------|
| B. Bright | v. very |
| F. Faint | c. considerable |
| L. Large | p. pretty |
| S. Small | e. extremely |

However, the first classification system that gained acceptance worldwide was published by Hubble (1926). Dubbed as “the Hubble tuning fork” or the “Hubble sequence”, this classification system arranges galaxies into a few broad categories: ellipticals, spirals, and irregulars. The Hubble system has its disadvantages and numerous modifications to the original Hubble scheme have been proposed (*e.g.* Morgan 1958, 1959; de Vaucouleurs 1959; van den Bergh 1960a,b,c). However, it currently remains the most popular classification method. It is important to note that at this time no classification method has been found to be ideal.

1.2.1 The Hubble Tuning Fork

The Hubble classification scheme visually arranges galaxies into three bins: ellipticals, spirals, and irregulars. The observer visually judges which classification bin a galaxy belongs to based on various parameters such as the size of the galaxy, the comparison of the concentration of light in the bulge of the galaxy to its disk, and the prominence of spiral arms. The main classification bins are further broken down into subcategories. Hubble’s initial classification categories are described as follows:

- Ellipticals (E): The main parameter used to judge the E galaxies is ϵ , which is ten times the ellipticity. It is defined as $\epsilon = 10(a - b)/a$, where a and b are the major and minor axes on the image of the galaxy, respectively. Ellipticity subcategories range from zero to seven, where E0 galaxies appear round and

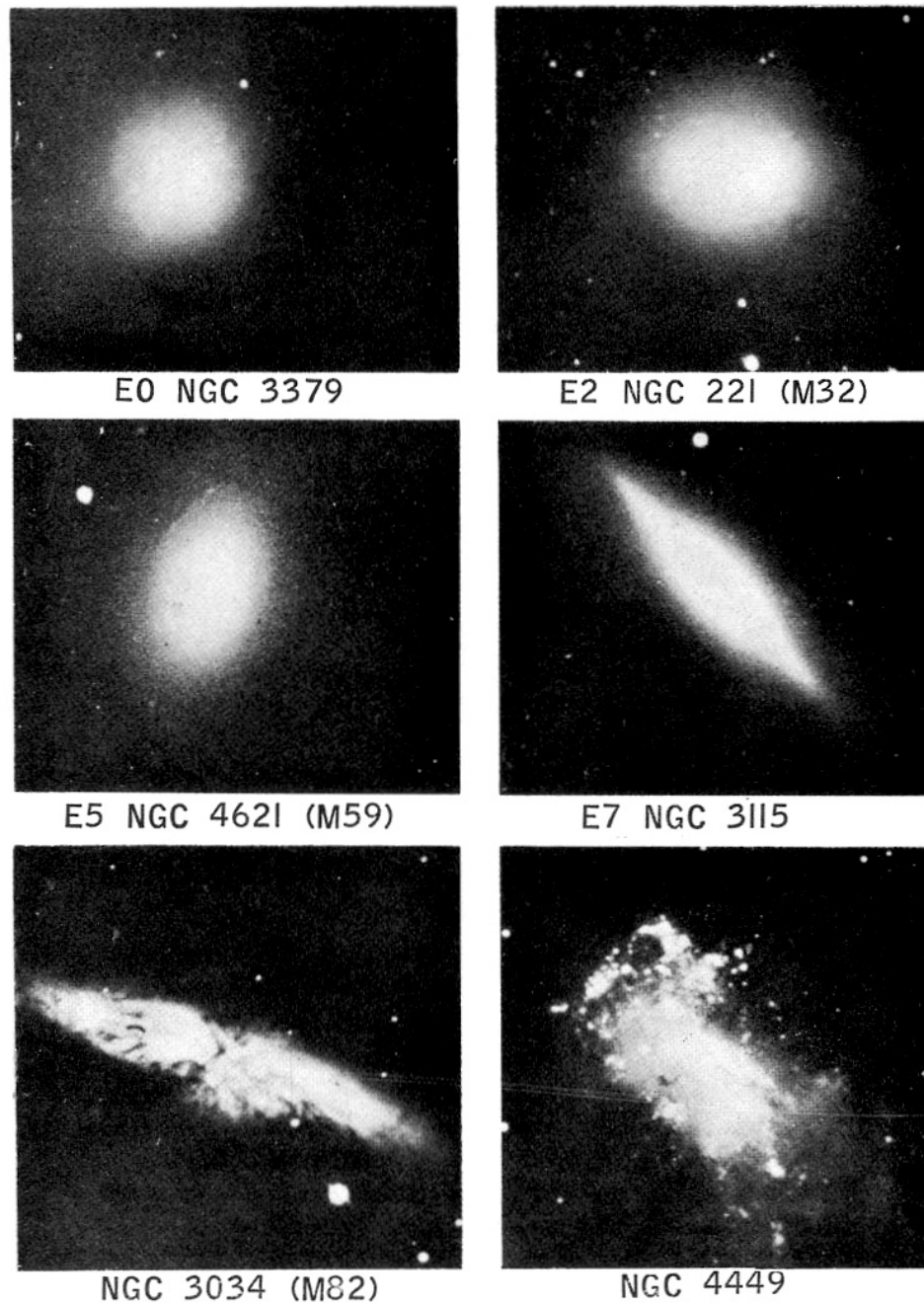


Figure 2: Elliptical and Irregular galaxies published in *The Realm of the Nebula* in 1936.

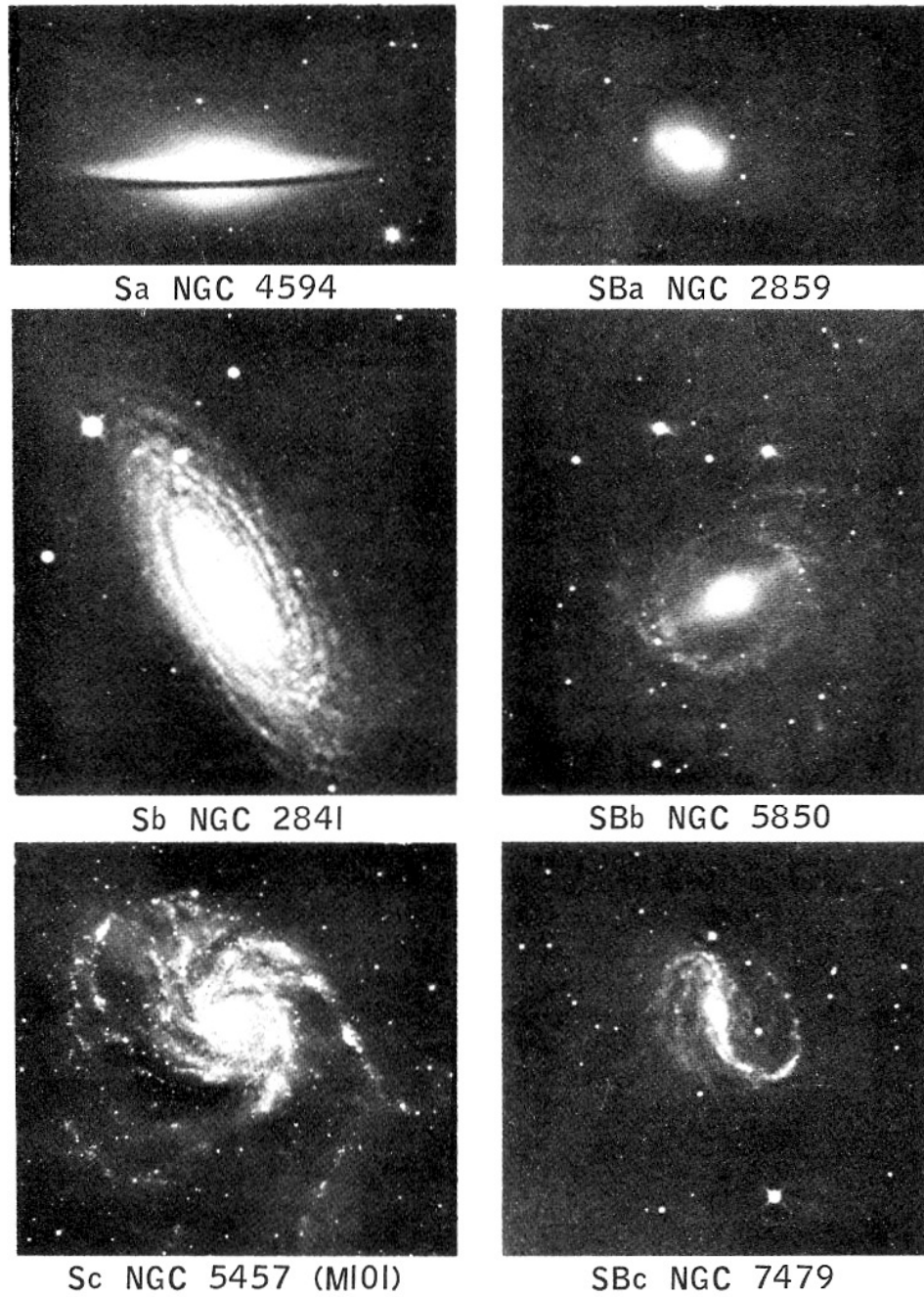


Figure 3: Normal and barred spiral galaxies published in Hubble's *The Realm of the Nebula*.

spheroidal in shape while E7 galaxies look more like flattened ellipsoids with an ellipticity of 0.7. Hubble found that no E galaxies have ellipticity greater than 0.7. For E galaxies, light distribution varies smoothly from the highly concentrated center to the dimmer outer edges. E galaxies do not possess disks or spiral arms, unlike the spirals.

- Spirals (S): This category has two forms – ordinary spirals (S) and barred spirals (SB). S galaxies are classified by the presence of disk-like, spiral pattern usually around a bright center (also referred to as the nucleus). The size of the nucleus, the tightness of the wounding of the arms around the nucleus, and the density of the spiral arms make up the criteria used to classify these galaxies. Depending on the inclination of an S galaxy, the ellipticity of its disk may be measured. Normal spirals appear on the top arm of the Hubble tuning fork diagram in Figure 4, where the subclasses for normal spirals are listed as: $Sa - Sb - Sc$. Type Sa have prominent central nuclei and tightly-wound spiral arms, whereas type Sc have an insignificant central nuclei and loosely-wound spiral arms. Type Sb range somewhere in-between. Spiral galaxies that have a bar-like structure present in their center are categorized beneath the normal spirals on the tuning fork, and their subclasses are similar to the normal spirals: $SBa - SBb - SBc$. As with the normal spirals, type SBa spirals also have more prominent nuclei and tightly wound arms, whereas SBc have smaller nuclei and loosely wound, patchy arms.
- Irregulars (Irr): This category is separate from the tuning fork. The Ir galaxies generally exhibit a patchy structure, show no evidence of rotational symmetry, and no obvious spiral arms.

In 1936, Hubble modified his original classification system to add lenticular (S0) and

barred lenticular (SB0) categories to account for the details that his initial classification bins lacked (Hubble 1936). The lenticular class was introduced as a transition type from ellipticals to the spiral (S and SB) classes. Referring to positioning the lenticular class in the junction between the E and S galaxies, Hubble writes:

“The junction may be represented by the more or less hypothetical class S0 — a very important stage in all theories of nebular evolution. Observations suggest a smooth transition between E7 and SBa, but indicate a discontinuity between E7 and Sa in the sense that Sa spirals are always found with arms fully developed [...] At the present, the suggestion of cataclysmic action at this critical point in the evolutionary development of nebula is rather pronounced” (Hubble 1936).

Hubble’s speculation about an evolutionary connection between class-types will be discussed later in this chapter. Thus, the final Hubble sequence consisted of four classification bins: elliptical (E), lenticular (S0), spiral (S or SB), and irregular (Irr). Examples of these galaxies can be seen in Figure 2 and Figure 3. In Figure 2, the four E galaxies range from spherical E0 to flattened E7. Galaxies that are flatter than E7 have a disk and are considered spiral or lenticular. NGC 3034 and NGC 4449 are irregular galaxies. In the case of the spirals shown in Figure 3, the galaxies in the left column are the normal spiral galaxies and the three in the right column are barred spirals. These images were taken with either the Hooker telescope (100-inch reflector) or the 60-inch reflector at the Mount Wilson Observatory of the Carnegie Institution of Washington.

Later, Shapley & Paraskevopoulos (1940) introduced the Sd group into the Hubble scheme, creating the following classification sequence for spirals: $Sa - Sb - Sc - Sd - Irr$, and similarly for the barred spirals. Holmberg (1958) further subdivided the

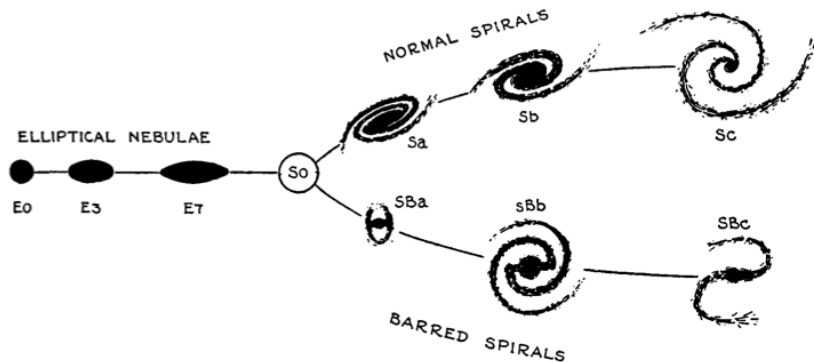


Figure 4: The 1936 version of the Hubble tuning-fork diagram, revised to include lenticular (S0) galaxies in order to transition from ellipticals to spirals.

classification bins into: $Sa - Sb^- - Sb^+ - Sc^- - Sc^+$. Using this classification scheme, Holmberg was able to show that classification types correlated with the mean colors of galaxies.

The color index is defined as $C = m_{pg} - m_{pv}$, where m_{pg} and m_{pv} are the photographic (\sim blue) and photovisual (yellow-green) magnitudes of an object, respectively. Holmberg (1958) showed that the mean intrinsic colors of galaxies change gradually from red to blue as one progressed down the following classification sequence: $E - Sa - Sb^- - Sb^+ - Sc^- - Sc^+ - Irr$. This was the first physical significance of a classification scheme.

Color is the difference between the magnitude of a celestial object in one filter versus another filter. It is an important physical property of celestial objects because it relates to the color of light that is emitted by the celestial object. A galaxy is made up of different colors, depending on the types of stars found in its stellar population. Stars that are cooler appear red in color while hotter stars are generally blue, therefore the light from spiral and irregular type galaxies is dominated by younger, hotter stars, and are overall blue in color. In contrast, elliptical and S0 galaxies lack young, massive stars, and thus are redder in color.

Classification of galaxies by color is a useful method because it organizes large number of galaxies together by type rather than individually. This allows for inspection and comparison of properties of similar galaxies to each other or to other types, and builds an understanding of their development. Since the mean color of a galaxy also reflects its general stellar population, this provides information for theories about the formation and evolution of the galaxy. However, classifying galaxies by color is not always accurate in the sense that there are cases where some S galaxies occupy the cluster red-sequence (*i.e.* are similar in color to the red E/S0 galaxies). The reddening of the color of some S galaxies may be due to a larger than average dust content, or that these red spirals have had their gas removed, thus quenching star formation.

Various criticism of the Hubble scheme has emerged throughout the years. Reynolds (1927) referred to Hubble’s classification system as being “too simple” and suggested a more detailed classification for the spiral galaxies. In return, Hubble argued that, because there is a great range of structural details to be dealt with when it comes to classification of galaxies, “a first general classification should be as simple as possible” (Hubble 1927). He also stated that the nuances in the features of certain galaxies are small compared to their prominent similarities with the broad classification bins he specified. Hubble admits that “some interesting details of structure” are ignored in his scheme, but nonetheless, he states that his “simple homogeneous system” offers a clear and definite system of classifying a large number of galaxies (Hubble 1927).

Nonetheless, the Hubble scheme is not a complete model for describing all galaxies observed in the Universe. It is effective in only classifying galaxies in the field and in nearby small clusters but unsuccessful in describing galaxies at moderate or high redshift, *i.e.* $z \geq 0.1$, or faint and small galaxies. Since its classification bins are broad, the Hubble scheme is unable to resolve galaxies in dense galaxy clusters, where the

majority of the galaxies appear to fall into either the elliptical or lenticular classes.

1.2.2 de Vaucouleurs Revised System

One modification to the Hubble classification system was proposed by de Vaucouleur (1959) who introduced finer divisions to Hubble’s classification bins, particularly to Hubble’s broad spiral galaxy class. The system keeps most of the same classification classes but aside from adding more detail, it can be viewed three-dimensionally. The system’s main axis is as follows: $E0 - S0 - Sa - Sb - Sc - Sd - Sm - Im - Irr$, where the index m refers to a galaxies’ resemblance to the Magellanic clouds. In Figure 5, the “A” at the top of the main axis indicates an absence of a bar, while the “B” below indicates a bar in the nuclues. The scheme is set up to differentiate between galaxies with no bars, labeled as “SA”; an intermediate class of galaxies with mixed characteristics, the weakly barred “SBA”; and galaxies with a presence of a prominent bar, labeled as “SB”. The r and s located on the sides of the main axis stand for ring and spiral, respectively. A galaxy may be categorized as having a mixture of a ring and spiral structure by the symbol rs , also. In Figure 5, the bottom-left diagram is a cross-section of the main diagram at the top, while the diagram on the bottom-right is a cross-sectional example of the various stages a galaxy can be classified into.

Besides it’s thorough approach to classifying spiral galaxies, the sequence also has finer subcategories such as $E, E^+, S0^-, S0^0, S0^+, etc.$, where the minus superscript refers to galaxies with smooth appearance (these galaxies can be also referred to as “late”) and the plus superscript refers to galaxies with a patchy appearance (*i.e.* “early”). The lower case $a, b, c,$ and d associated with the spiral galaxies stand for “early”, “intermediate”, “late” and “very late”. In this way, the de Vaucouleurs system offers many classification bins for spiral galaxies.

Later, de Vaucouleurs introduced a new numerical parameter called the T -type,

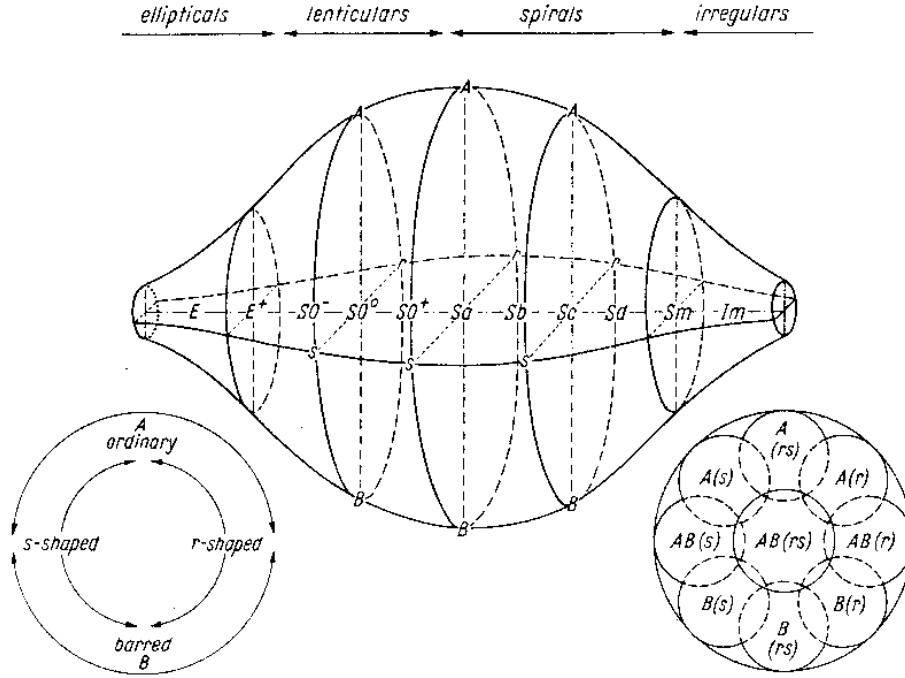


Figure 5: The three-dimensional representation of de Vaucouleurs' revision to the Hubble tuning fork diagram (de Vaucouleurs 1959).

which is correlated to the main axis of his three-dimensional classification system (see Table 2). The definitions for the T values follow directly from de Vaucouleurs main axis definitions stated above (de Vaucouleurs *et al.* 1991). The T parameter ranges from $T = -6$, compact ellipticals (cE), to $T = 10$, magellanic irregulars (Im).

Table 2: The de Vaucouleurs galaxy classification with corresponding de Vaucouleurs T type (de Vaucouleurs 1994).

| | | | | | | | | | |
|----------------------------|-----------|------------|-----------------------|------------------------|------------------------|------------------------|-------------|-----------|------------|
| de Vaucouleurs | <i>cE</i> | <i>E</i> | <i>E</i> ⁺ | <i>S0</i> ⁻ | <i>S0</i> ⁰ | <i>S0</i> ⁺ | <i>S0/a</i> | <i>Sa</i> | <i>Sab</i> |
| T Type | -6 | -5 | -4 | -3 | -2 | -1 | 0 | 1 | 2 |
| de Vaucouleurs | <i>Sb</i> | <i>Sbc</i> | <i>Sc</i> | <i>Scd</i> | <i>Sd</i> | <i>Sdm</i> | <i>Sm</i> | <i>Im</i> | |
| T Type | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |

The physical significance of the de Vaucouleurs scheme can be viewed similarly to the Hubble system. Just as with the Hubble classification, the main axis of de Vaucouleurs three-dimensional scheme correlates with the mean color of galaxy type,

which can be related to the temperature and age of the stellar population of the galaxy type (van den Bergh 1998). The “A” and “B” forms are found to not significantly differ in color, which implies that both forms contain stars of similar ages. It has also been found that the r and s forms occur in similar frequencies among the early-type spirals, while among the late-type spirals the s form occur much more frequently than the r (van den Bergh 1998). The de Vaucouleurs system is also not one without problems. Along the $Sc - Sd - Sm$ sequence, galaxies are found to become fainter and bluer at the same time, which shows that the de Vaucouleurs system does not separate luminosity and color effects for these galaxies.

1.2.3 Morgan’s Galaxy Classification System

Morgan’s Classification System (also known as the Yerkes System) classifies galaxies based on the concentration of light in their centers and their spectral features. It was produced from the author’s earlier work of classifying galaxies from their spectra (Morgan *et al.* 1957).

Through visual inspection of galaxies, classification based on this system can be done by relying on a sequence of fundamental parameters: $a - af - f - fg - g - gk - k$, where “a” represents the group of galaxies with little or no central concentration of light and “k” represents the group of galaxies with high concentration of light in the center. The groups in between “a” and “k” are referred to as intermediate categories. In this manner, relation between central concentration and galaxy morphology can be observed from the Morgan classification system. Galaxies with high central concentration of light tend to have older stellar populations than galaxies with low central concentration of light. These high central concentration galaxies would be referred to as class E or S0 on the Hubble scale. They would be located on the “k” side of the Yerkes classification scale. Unlike the Yerkes classification system, the Hubble

Table 3: The revised “form families” of the Yerkes classification system (Morgan 1958, 1975)

| Form Family | Description |
|-------------|-------------------------------------------------------------------------------------------------|
| S | Spirals |
| B | Barred Spirals |
| E | Ellipticals |
| I | Irregular systems |
| R | Systems without clearly marked spiral or elliptical structure but that show rotational symmetry |
| N | Systems with small, bright nuclei, and faint background |
| N- | Less pronounced N galaxies, <i>i.e.</i> weak nuclei |
| N+ | Very pronounced N galaxies, <i>i.e.</i> bright nuclei |
| Q | Unresolved objects with starlike appearance and large redshifts |
| C | Small galaxies with large surface-brightness |
| D | Galaxies with an elliptical-like nucleus and extensive envelope |
| cD | Supergiant D galaxies |
| db | Dumbbell-shaped galaxies with two distinct nuclei |

classification scheme does not adequately describe galaxies in rich clusters, where galaxies are mostly E’s and S0’s. The Yerkes classification system offers more options to classify galaxies in such environments (van den Bergh 1998).

Secondary parameters are also introduced in the Yerkes system and are referred to as “form families”. The system also includes a purely geometrical parameter to describe the approximate degree of tilt of each system — referred to as an “inclination class” and is analogous to Hubble’s use of ellipticity. A number between one through seven is assigned to a galaxy, in which a circular face-on galaxy would be noted by a “1” and a highly elliptical edge-on galaxy by a “7”. The system was revised in 1975 and the form families are stated in Table 3.

It has been found that measurements of galaxies’ central concentration is possible using digital images (Abraham *et al.* 1994). For an automatic computer classification, central concentration for each galaxy can be expressed as a parameter usually labeled by the letter C . This parameter is found by measuring the ratio of flux at two different

radii. Different definitions of C have been used in the literature and this, along with other parameters, will be discussed in Chapter II.

1.3 Formation and Evolution of Galaxies

Today, the process of galaxy formation and galaxy evolution is a subject with more questions than answers. In the early 1900's, Hubble built his classification scheme from "simple inspection of photographic images" (Hubble 1922). However, before attention could be turned to the study of galaxy formation and evolution, astronomers studied the development of stars and planets. Mathematical physicist and astronomer Pierre-Simon Laplace (1796 — 1827) suggested that the Sun formed from a large, rotating cloud that gradually shrank down to its present size. He writes in *The System of the World*:

"Whatever the sun's nature, it must have encompassed all of the planets; and considering the enormous distances separating these bodies, it must have been a fluid of an immense extent. In order to have given the planets almost circular motions in the same direction, this fluid must have surrounded the sun like an atmosphere. The consideration of the planetary motions thus leads us to think that [...] the solar atmosphere originally extended beyond the orbits of all the planets and that it progressively shrank to its present limits" (Whitney 1971).

William Herschel was one of the first to propose theories of evolution of astronomical objects, which Laplace had also analyzed in later editions of *The System of the World*. Herschel believed that stars formed from the peculiar looking "planetary" nebula — objects that appeared to be gaseous but with a single star at the center (Whitney 1971) — though most astronomers today would disagree with this concept.

Other theories, such as the mathematical Jeans-Jeffreys tidal hypothesis (Jeans 1917; Jeffreys 1918), suggested by James Jeans and Harold Jeffreys, initially proposed to explain the formation of the solar system and planets. The Jeans-Jeffrey theory described the formation of the planets in the solar system as a result of a tidal interaction between the Sun and a nearby star rather than through rotational kinematics. As was mentioned previously in this chapter, at this time in history, the size of the Universe was still a subject of debate and many believed that the nebulae were objects within our Milky Way. Jeans argued that rotation played a role in forming the different shapes of spiral galaxies but not in forming solar systems (Milne 2013). He believed that the spiral galaxies originally formed from a large, rotating nebulous cloud that gradually condensed down to form stars (Hetherington 1993) and that it was the rotation of these systems that gave rise to the different spiral-shapes. Objections to his tidal theory came later and even the author himself wrote: “This vague sketch of the tidal theory will, it is hoped, to be read as an indication of the possibilities open to the tidal theory [...] The theory is beset with difficulties, and in some respects appears to be definitely unsatisfactory.”

In the 1900’s, Hubble also explored the possibility of an evolutionary progression between the different shapes of galaxies. He states in *American Section Report*: “We seem to be succeeding with the evolutionary sequence classification of the stars, and we may look forward with some hope to a time when something of the sort can be attempted with the nebula” (Berendzen *et al.* 1976).

Hubble referred to galaxies on the left of the tuning fork in Figure 4 (*i.e.* E galaxies) as “early” type and ones on the right (*i.e.* S galaxies) as “late” (Carroll 2006; Hubble 1936). In *The Realm of the Nebula*, he describes the Hubble tuning fork:

“The progression throughout the complete sequence thus runs from the

most compact of the elliptical nebula to the most open of the spiral — a progression in dispersion or expansion. The terms ‘early’ and ‘late’ are used to denote relative position in the empirical sequence without regard to their temporal implications. These explanations emphasize the purely empirical nature of the sequence of classification. The consideration is important because the sequence closely resembles the line of development indicated by the current theory of nebular evolution as developed by Sir James Jeans” (Hubble 1936).

Due to lack of definite proof, Hubble was hesitant to push the evolutionary element of his theory, nonetheless he also wrote: “There is [...] some grounds for using the terms early type and late type spirals and considering the elliptical nebula and spirals as a single evolutionary sequence” (Hart 1971).

Today, the nomenclature of “early” and “late” type galaxies still remains, as well as the questions, “How do galaxies form?” and “Do they evolve? If so, how?” Though it is not universally accepted by everyone in the astrophysical community as of yet, Allan Sandage suggested in 1961 that galaxies evolve the opposite way along the tuning fork (from “late” to “early”), citing the fact that young stars are observed in late-type galaxies. He writes: “There is an almost one-to-one correspondence between the presence of dust and the presence of bright, blue O and B stars. Such stars are known to be very young because their nuclear energy sources can last for only a few million years. Since they are visible today, they must have been created within the last several million years. It is invariably the Irr, Sc, and SBc galaxies that contain these young stars,” meanwhile early-type spirals and E galaxies “show little or no resolution into bright stars or HII regions. Star formation has apparently stopped completely, because all the necessary dust has been used up. These galaxies contain

stars that are very old [...]” (Sandage 1961).

1.3.1 The Big Bang Theory and the Early Universe

To get to the roots of galaxy formation and evolution it is important to note the current knowledge about the early stages of the Universe. In the 1910’s, before it became clear that spiral nebulae were objects outside of the Milky Way, Vesto Slipher was the first astronomer to measure the radial velocities for a number of galaxies using the 24-inch telescope at Lowell Observatory in Arizona, USA. He found that most of the spectra showed redshifted spectral lines, meaning that the galaxies were receding from Earth.

Einstein published his theory of General Relativity in 1916, however, at this time, astronomers believed that the Universe is neither expanding nor contracting, meaning, the Universe is static. However, in 1917, after discovering that his equations suggest a dynamic Universe in which galaxies would gravitationally influence one another, Einstein introduced a constant to balance the attractive force of gravity and match the popular belief of a static Universe. In his 1917 paper, “Cosmological Considerations in the General Theory of Relativity” he states:

“Thus the theoretical view of the actual universe, if it is in correspondence with our reasoning, is the following. The curvature of space is variable in time and place, according to the distribution of matter, but we may roughly approximate to it by means of a spherical space. At any rate, this view is logically consistent, and from the standpoint of the general theory of relativity lies nearest at hand; whether, from the standpoint of present astronomical knowledge, it is tenable, will not here be discussed. In order to arrive at this consistent view, we admittedly had to introduce

an extension of the field equations of gravitation which is not justified by our actual knowledge of gravitation. [...] That term is necessary only for the purpose of making possible a quasi-static distribution of matter, as required by the fact of the small velocities of the stars” (Engel 1997).

This constant — also known as the “cosmological constant” and usually characterized as Λ — represents a repelling force that balances the gravitational attraction between galaxies in such a way that the Universe remains static.

Other’s have also speculated that the Universe might be expanding (or contracting) and early evidence of it can be found from Alexandre Friedmann’s and Georges Lemaître’s solutions of Einstein’s field equations published independently in their 1922 and 1927 papers, respectively. In his book “The World as Space and Time” published in 1923, Friedmann writes:

“The non-stationary type of Universe presents a great variety of cases: for this type there may exist cases when the radius of the curvature of the world, starting from some magnitude, constantly increases with time; there may further exist cases when the radius of curvature changes periodically: the Universe contracts into a point (into nothingness), then again, increases its radius from a point to a given magnitude, further again reduces the radius of its curvature, turns into a point and so on” (Evans 2015).

However, the equation that demonstrated that the Universe is expanding is credited to Edwin Hubble, who supplied observational evidence for this law, now called the “Hubble Redshift-Distance” relation or the “Hubble’s law.” As was mentioned previously in this chapter, Hubble measured the distance to galaxy M31 in 1929 and continued to determine distances to other galaxies. Later, he and his assistant, Milton

Humason, found a linear correlation between distances and radial velocities (Hubble 1929b, 1931): $v = H_0 d$.

The variables in Hubble’s law are as follows: v is the radial velocity of galaxies observed and d is the proper distance to the galaxies. H_0 is the present-day constant of proportionality known as the Hubble constant. The value of the Hubble constant remains the subject of study today. After Edwin Hubble presented his observations of an expanding Universe, Einstein referred to the cosmological constant he introduced before as his “biggest blunder”. But with the development of our knowledge about dark matter and dark energy, we now see it may not have been so.

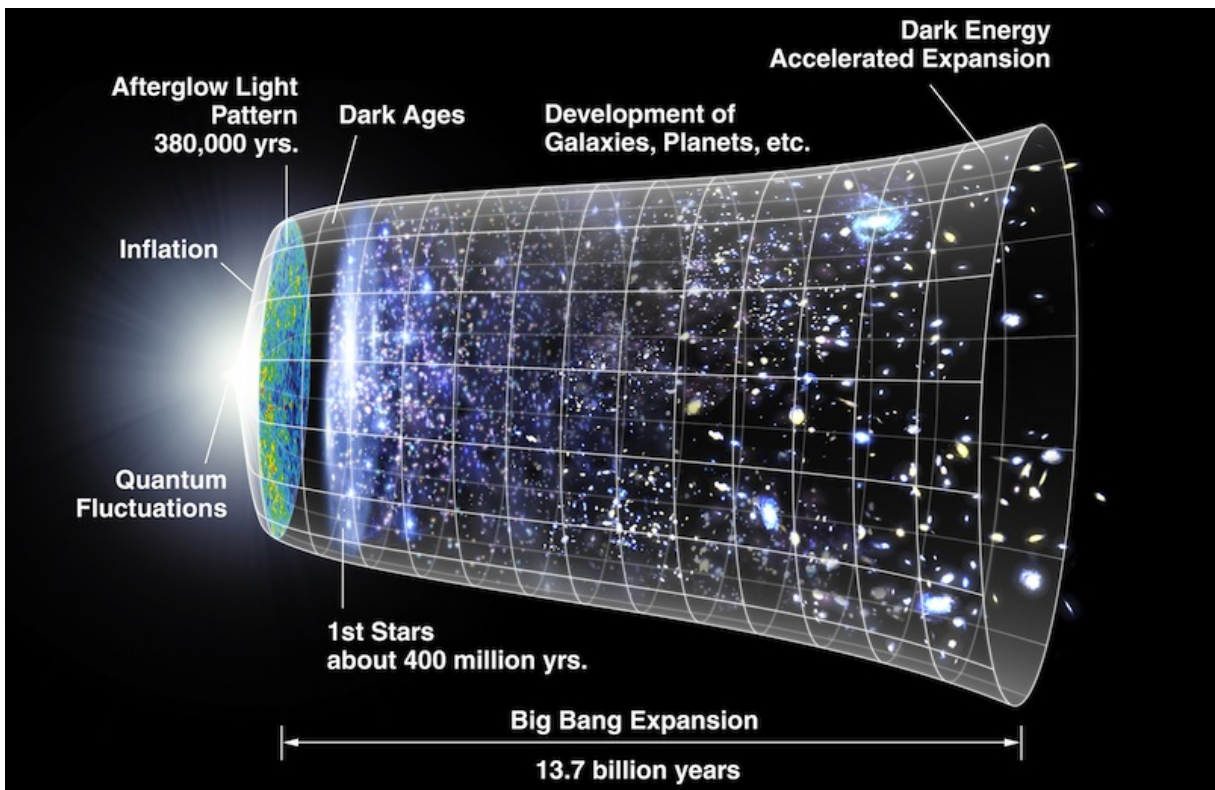


Figure 6: The cosmic timeline of the expanding Universe according to the Big Bang model.

One logical implication of an expanding Universe is that it must have been smaller in the past. Lemaître was one of the first to suggest that the Universe started as a

“primeval atom”, stating in 1931: “[...] at the origin, all the mass of the universe would exist in the form of a unique atom; the radius of the universe, although not strictly zero, being relatively small. The whole universe would be produced by the disintegration of this primeval atom. It can be shown that the radius of space must increase” (Luminet 2011).

Another implication of an expanding Universe is that if it was smaller in the past, then it must have been hotter also (of order 10^9 — 10^{10} K). This idea came to be known as the “Big Bang” theory. The cosmic timeline of the Big Bang theory can be visualized in Figure 6¹. In 1896, Henri Becquerel accidentally discovered radioactivity from his experiments with phosphorescent material and thus began the intensive study of this subject. His discovery later led Ernest Rutherford to develop the concept of radioactive half-life in which an element can change into another through an emission of an alpha or beta particle. Expanding upon these ideas, George Gamow, while studying the abundances and formation of elements in 1948, predicted that since the early, dense Universe must have been hot, the gases of that time should have emitted strong blackbody radiation.

Gamow, along with Ralph Alpher and Robert Herman, were some of the first astrophysicists to theoretically predict the existence of this blackbody radiation, now called the cosmic microwave background (CMB) radiation, *i.e.* the earliest radiation in the Universe. Alpher and Herman estimated the present temperature of the CMB to be ~ 5 K (Alpher & Herman 1948; Evans 2015). However, though it wasn’t understood at the time, CMB was noticed as early as 1941 by radio astronomers (*e.g.* Adams 1941; McKellar 1941) who observed that cyanogen molecules (CN) found in different parts of space all have faint absorption lines at the first excited energy state. In theory, empty

¹Image Source: NASA/WMAP Science Team - Original version: NASA; modified by Ryan Kaldari.

space should have been absolutely cold, but McKellar writes that the interstellar space has a higher temperature: “[...] several sharp lines of interstellar origin in the spectra of distant stars are due to transitions from the lowest energy states of the diatomic molecules CH and CN. Thus not only has it been shown that diatomic molecules exist in interstellar space but also the presence there of the hitherto undetected elements hydrogen, carbon, and nitrogen has been demonstrated. [...] Also from Adams’ results on the interstellar CN lines, it can be calculated that the ‘Rotational’ temperature of interstellar space is about 2°K” (McKellar 1941).

When a molecule absorbs a photon, it moves from a ground state to an excited energy state. Molecules located in the space between Earth and distant stars produce absorption lines in the spectra of these stars. Most of these absorption lines are found to be from the ground state of the molecules, but this was not the case for CN. Still, it wasn’t until 1965 that the CMB became known for what it is today. Princeton University’s astronomer Robert Dicke and his team were in the process of building a telescope that could detect the hypothetical “cosmic background radiation” from the early Universe in the early 1960’s. However, Arno Penzias and Robert Wilson, from Bell Laboratories, unknowingly were quicker. Penzias and Wilson used highly sensitive equipment — a 20-foot horn-antenna in Holmdel, NJ, USA, thus dubbed the “Holmdel horn” antenna — while studying microwave signals from the Milky Way and noticed they had unaccountable noise. When the equipment was aimed at the zenith, Penzias and Wilson found that the antenna picked up a microwave signal that remained even after natural microwave radiation from the Earth’s atmosphere was subtracted out. As Penzias and Wilson report: “Measurements of the effective zenith noise temperature of the 20-foot horn-reflector antenna [...] at 4080 Mc/s have yielded a value about 3.5°K higher than expected” (Penzias & Wilson 1965).

Dicke and his team at Princeton University published their paper that same year,

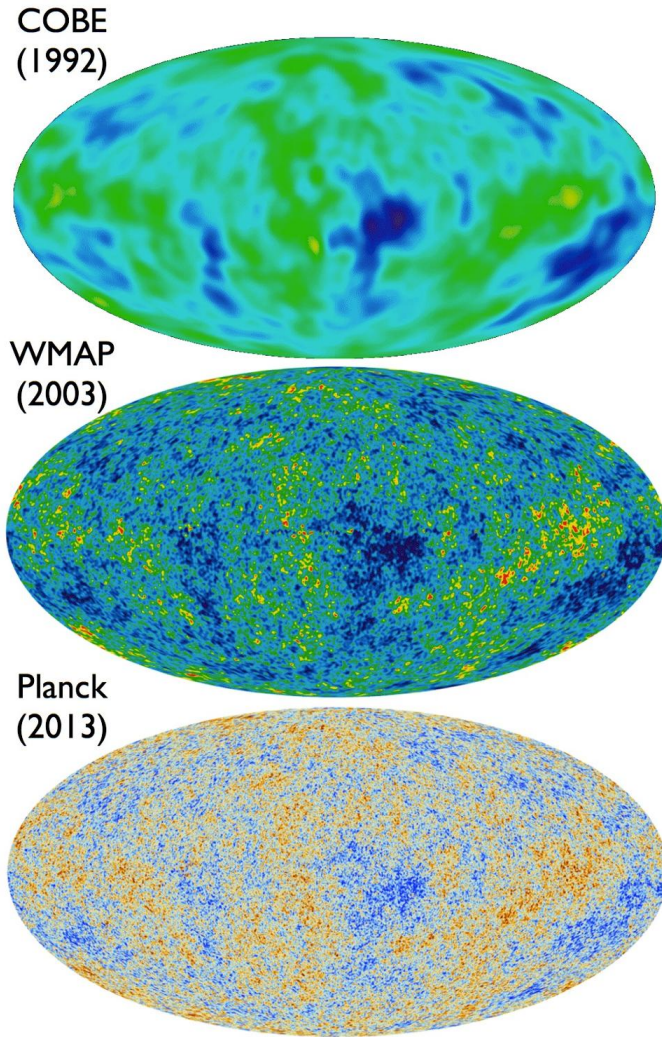


Figure 7: Comparison of improvement in resolution of the CMB imaged by the COBE, WMAP, and Planck satellites.

titled “Cosmic Black-Body Radiation,” stating that though they have yet to obtain their own results from their instruments, “Penzias and Wilson (1965) of the Bell Telephone Laboratories have observed background radiation at 7.3-cm wavelength. In attempting to eliminate every contribution to the noise seen at the output of their receiver, they ended with a residual of $3.5^\circ \pm 1^\circ\text{K}$. Apparently this could only be due to radiation of unknown origin entering the antenna. [...] A temperature in excess of 10^{10} °K during the highly contracted phase of the universe is strongly implied by a

present temperature of 3.5°K for black-body radiation” (Dicke *et al.* 1965). Penzias and Wilson received the Nobel Prize in Physics in 1978 for the discovery of the CMB.

However, there was another hypothesis for the origin of the Universe: the Steady State theory. Unlike the Big Bang theory, which states that the Universe is expanding and was much denser and hotter in the past, the Steady State theory — proposed by Fred Hoyle, Hermann Bondi, and Thomas Gold at Cambridge University in mid-1900’s — states that the Universe is always expanding, maintains a constant overall density, and remains unchanging with time. The Universe, according to this theory, has no beginning or end, but as it expands, matter is created in order to maintain the same average density. However, the Steady State theory had no plausible explanation for the CMB, and thus, as there was strong evidence of the existence of this radiation in the Universe, this theory has fallen out of favor.

In November 1989, NASA launched the Cosmic Background Explorer (COBE) satellite to study the CMB. It was the first satellite to take precise measurement of this radiation. The COBE satellite was operational from 1989 till 1993. It’s successors were the Wilkinson Microwave Anisotropy Probe (WMAP), which operated from 2001-2010, and the Planck spacecraft (2009-2013). The improvements of the resolution of these instruments can be seen in Figure 7². The angular resolution of COBE³ was 7° , meaning that only features larger than this were detected. In 2001, WMAP⁴ had a resolution of 0.23° and improved measurement accuracy of temperature variations. The WMAP objective was also to measure the temperature differences in the CMB radiation. The European Planck satellite⁵ had a resolution of $5' = 0.0833^\circ$. It had also mapped the anisotropies of the CMB at microwave and

²Image compiled by Ethan Siegel at Forbes Online.

³<https://science.nasa.gov/missions/cobe>

⁴<https://map.gsfc.nasa.gov/>

⁵<http://planck.caltech.edu> & <http://www.esa.int/planck>

infrared frequencies.

1.3.2 Relating to Cosmology

Physical cosmology is the study of the origin and evolution of the Universe as a whole. Much literature (*e.g.* Chiosi *et al.* 2014; Houjun *et al.* 2010; Seeds *et al.* 2001) is being written about this subject as it is currently an important topic of study in astronomy. Modern physical cosmology is based upon the cosmological principle and Albert Einstein’s theory of General Relativity. The cosmological principle states that on large-scales, the Universe is both isotropic and homogeneous, *i.e.* spatially the Universe is uniform, while the theory of General Relativity describes how massive objects distort space-time.

The current belief in cosmology is that the following exist in the Universe: baryonic matter, dark matter, and dark energy. Baryonic matter in astronomy refers to ordinary matter from which everything we see around us, including objects such as stars and galaxies, is made. It is also known as “visible” matter. In the Standard Model of particle physics, baryonic matter refers to matter composed of three quarks. Particles such as protons and neutrons are considered baryonic, whereas electrons and neutrinos are not. However, in astronomy, electrons are included in the term (while neutrinos are usually not). At this time, dark matter and dark energy are not well understood but there is a number of observational evidence to suggest their existence. Among the evidence to support the concept that the Universe contains another sort of matter, the so-called “dark” matter, are the flat rotation curves of various galaxies, including the Milky Way. Galaxy rotation curves (*i.e.* plot of orbital speed versus distance from the center of a galaxy) have been recorded from analyzing the motions of stars and speeds of clouds of hydrogen gas in galaxies. These curves have been found to flatten out with distance. Since most of the mass in a galaxy is

concentrated at the center, it was expected that the velocity of stars should decrease with the square root of the radius (this is also called the “Keplerian” rotation curve). However, very few galaxies seem to follow this trend. Most galaxies have rotation curves where velocity remains more-or-less constant with distance, which implies that mass continues to increase linearly with radius. One explanation for this has been dark matter — matter that seems invisible to us but one that exerts a gravitational force. Gravitational lensing has been found to further support the existence of dark matter.

Finally, there is also observational proof of “dark energy” to suggest its existence. The Universe has been found to have accelerated expansion, and since some sort of unknown energy must be causing this, it is now called the “dark” energy. The standard model for our Universe, which is based on the Big Bang theory, is the Λ CDM model, *i.e.* the Lambda Cold Dark Matter model, which states that our Universe appears to be spatially flat and that only $\sim 5\%$ of the Universe is baryonic matter while the rest is dark matter ($\sim 25\%$) and dark energy ($\sim 70\%$).

1.3.3 Evolution Mechanisms and Models

Due to improvements in technology, our understanding of galaxy formation and evolution is advancing and changing rapidly. There are a number of mechanisms that have been used to explain the different shapes of galaxies that we see in our Universe.

Firstly, the environment can play an important role in forming galaxies. From observations, we know that most galaxies are found in groups. Isolated galaxies are rarely found in space. Rich clusters of galaxies contain thousands of galaxies (mostly E type), therefore these environments are very dense. Most of the galaxies in rich clusters also tend to be concentrated in the center of the cluster. Poor clusters, on the other hand, contain fewer galaxies and these galaxies are widely spread throughout

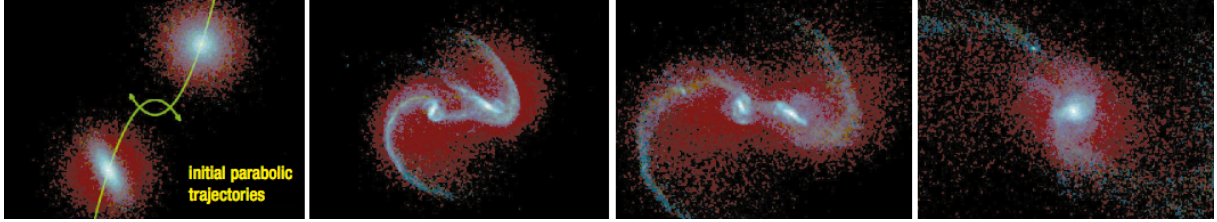


Figure 8: The left-to-right progression of these diagrams show a numerical simulation by Joshua Barnes from the University of Hawaii published in Ellis *et al.* 2000.

their systems. These are not highly dense environments (Seeds *et al.* 2001; Pasachoff *et al.* 2007).

In high density environments, galaxies may collide and/or merge more frequently than in low density environments. Interactions can distort the galaxies' shapes or possibly form a different class of galaxy. Since galaxies are large systems of stars, when these systems collide they essentially will pass through one another. Unlike clouds of gas (which will interact), the distances between stars are much greater than the sizes of the stars, thus the probability of stars colliding with each other is small. However, tidal forces due to the gravitational fields of the systems can distort their shape (causing tidal tails) or even merge the two galaxies together.

Though it remains a subject of debate, there is evidence to believe that an elliptical galaxy can be produced by the collision and merging of two or more disk-type galaxies, for example, as described in Lutz 1991. For the disk-type galaxies, gravitational interactions of some form may also be necessary to produce their particular shapes (Seeds *et al.* 2001). Numerical simulations, such as seen in Figure 8, demonstrates how elliptical galaxies might form when two spirals merge. Simulations indicate that the tails (seen in the diagrams) may be transient structures, and the final outcome of the merging is an elliptical galaxy.

Secondly, since about the 1970s, there have been studies on the possibility of

galactic evolution due to internal processes or instabilities of the galaxy, such as galactic winds, black holes, and dark matter halos, or the movements of spiral arms and bars. This slow and steady evolution, otherwise known as “secular evolution,” may also be responsible for the formation of certain, particularly disk-type, galaxies. But evolution of galaxies may be a combination of secular and environmental results. For example, Kormendy *et al.* (2004) write: “At early times, galactic evolution was dominated by hierarchical clustering and merging, processes that are violent and rapid. In the far future, evolution will mostly be secular: the slow rearrangement of energy and mass that results from interactions involving collective phenomena such as bars, oval disks, spiral structure, and triaxial dark halos.” After gaining angular momentum, spiral galaxies may undergo gradual accretion of material into their systems from their environment.

One of the questions in the study of galaxy evolution is whether galaxies form and develop in isolation *or* whether they form in clusters. But perhaps it is not a question of “or.” It may be that both methods are responsible for forming and developing a galaxy of the same sort. There are currently two proposed models for galaxy formation:

- The Classical Model: also known as the “monolithic” collapse model, seen in Figure 9 on the left hand side. This model proposes that galaxies develop from the rapid collapse of large gas clouds, *i.e.* a top-down approach to formation. There is little impact on galaxies’ development from the surrounding environment. This model can be used to explain the formation of elliptical galaxies and spiral bulges.
- The Hierarchical Model: on the right hand side of Figure 9, proposes that galaxies gradually form and evolve through a series of mergers, *i.e.* a bottom-up

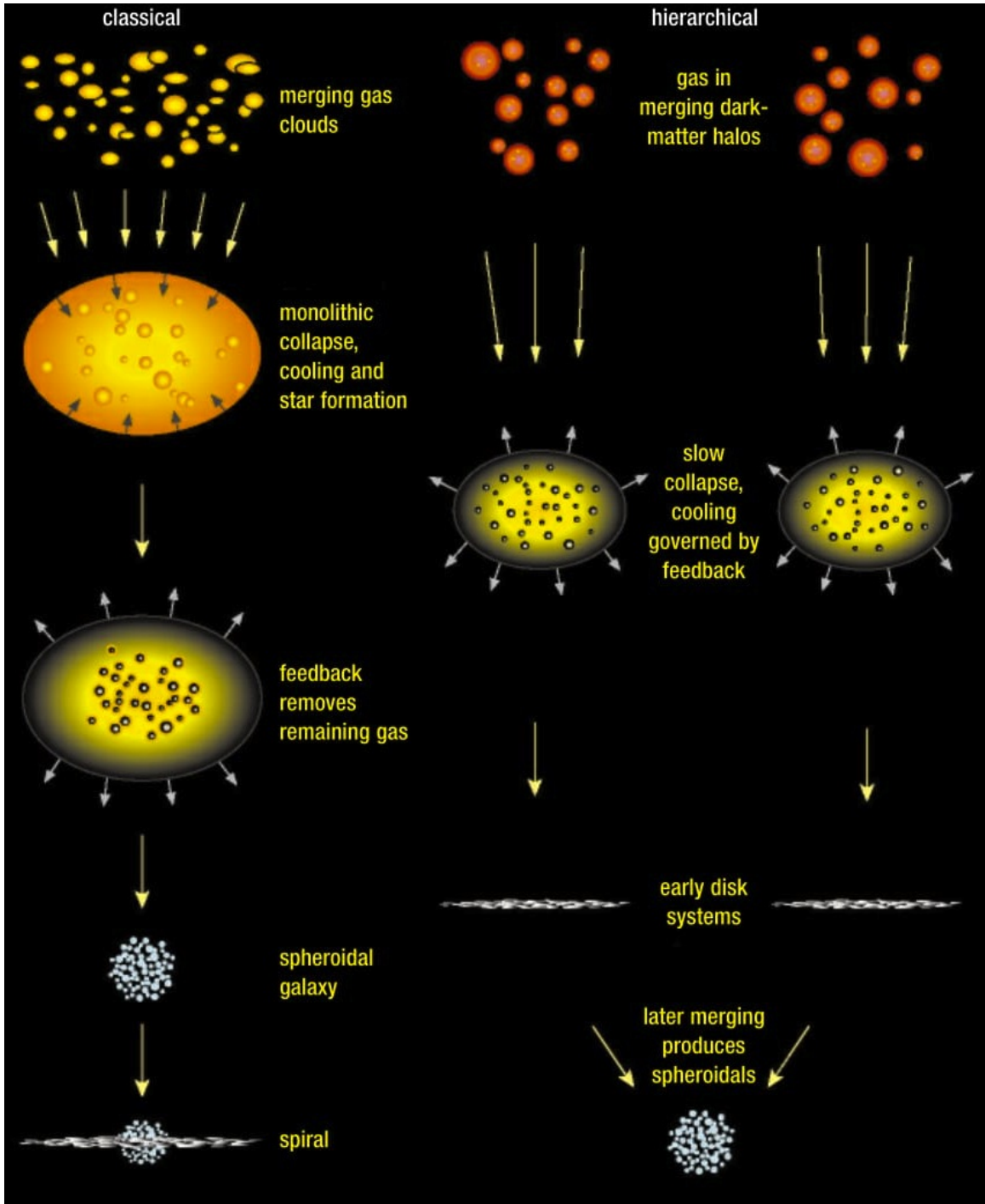


Figure 9: The two well-known models for galaxy formation and evolution. Image reproduced from Ellis *et al.* 2000.

approach. In this theory, galaxies' formation is dependent on the environment surrounding them. This model can explain the existence of galaxies shortly after the Big Bang. This model can also be used to explain the formation of disk galaxies.

The “monolithic” collapse was the popular model for galaxy formation before the existence of dark matter became known. Nowadays, the hierarchical model seems to be the acceptable model for galaxy formation, however, it still has room for improvement.

In the next chapter we introduce the software we have developed to read galaxy images and measure a number of parameters in order to perform galaxy classification and analysis.

CHAPTER II

MORPHOLOGY SOFTWARE

2.1 Reasons for Automation

Most of the existing galaxy classification systems require observers to visually classify individual galaxies. However, this task has a number of limitations. In recent years, one organized venture into classifying a large number of galaxies has been the online citizen project called the Galaxy Zoo¹. This project started in 2007 as a means to aid in the classification of galaxies from the data collected mostly by the Sloan Digital Sky Survey (SDSS)². In the Galaxy Zoo project, the SDSS galaxies are classified by eye by a large number of people throughout the world. For small groups of researchers, it is a tedious and time-consuming task to manually classify thousands of galaxies, but by establishing the Galaxy Zoo, the mission of classifying thousands of galaxies has become manageable. Nevertheless, the data available from ongoing and future surveys — such as the Dark Energy Survey (DES) or the Large Synoptic Survey Telescope (LSST) — will be immense, since these surveys will cover large areas of the sky (for the LSST, approximately 50% of the total sky). It will be difficult for researchers to analyze the structure of galaxies in a timely fashion even with the aid of citizen projects. Therefore, the fact that modern surveys produce large amounts of data in a short amount of time, and with a small number of personnel available to manage it, is one of the limits of the visual classification of galaxies.

¹<https://www.galaxyzoo.org/>

²<http://www.sdss.org/>

Additionally, classifying galaxies by eye can introduce human bias. When observers look at galaxies, they may insert parameters into their classification that have no specific relation to the galaxy itself. For example, it has been noted that galaxies that appear fainter and smaller tend to be classified as early-type by observers (e.g. Lintott 2010; Deng 2013; O’Leary 2013). However, though it is not ideal, the human eye still remains the best device for noticing structural patterns and detecting low surface brightness features in objects.

Another challenge of classifying galaxies today occurs with classifying ones at higher redshifts. Since the current classification systems, such as the Hubble system, rely on specific parameters, e.g. the bulge-to-disk ratio or the tightness of spiral arms, it is often difficult to resolve such structures from the images of these galaxies. Since the classification of galaxies nowadays is done through analyzing digital images in some manner, it is limited by the angular resolution of those images. At higher angular resolution, finer details of the galaxy can be recognized. However, distant galaxies often appear small and faint even on high-resolution images taken by the Hubble Space Telescope (HST). Therefore, visual classification of distant galaxies generally becomes difficult. As Willett *et al.* (2013) write, the fraction of votes from observers who noticed “finer morphological features (such as identification of disk galaxies, spiral structure, or galactic bars) decreases at higher redshift.”

Also, the more distant the galaxies, the less they seem to possess the regular features (such as spiral structure or ellipticity) that many low-redshift galaxies have (Abraham *et al.* 1996a; Driver *et al.* 1998; Cheng 2009), which makes it difficult to classify them into the current Hubble bins. The shapes of galaxies become more chaotic with increasing distance, therefore, the classification categories of most classification systems today, such as the Hubble tuning fork, do not generally apply to high-redshift galaxies.

For many galaxies, their structure also depends on the wavelength at which they are viewed. As was mentioned in the previous chapter, younger, hotter stars emit most of their light as ultraviolet radiation (*i.e.* short wavelengths of light) and thus appear bluer in color, while cooler, older stars appear more red (*i.e.* emit longer wavelengths of light). Therefore, when imaged at shorter wavelength, the number of blue stars dominates over the red stars. Thus, this may give the impression that there are more blue stars than red, and the human mind may create a bias towards this feature during visual classification (Wirth 1984).

Also, imaging a galaxy at shorter wavelengths can give it a “patchy” appearance, since hot stars tend to cluster in groups, rather than be evenly spread out across a galaxy (Buta 2000; Ellis 2000). The presence of dust — which is mainly a collection of fine particles — can also affect the appearance of galaxies. Dust scatters short-wavelengths more effectively than long-wavelengths and thus it disturbs images taken in the blue-wavelength more than in the red. However, in the infrared, dust also produces thermal emission that can conceal the true shape of the galaxy (Buta 2000).

In recent years, methods of automatic classification have been receiving more attention (e.g. Odewahn 1995; Abraham *et al.* 1996a,b; de la Calleja 2004; Shamir 2009; Viral *et al.* 2015). One of the main goals of automatic galaxy classification is to develop “objective algorithms to produce classification parameters directly from digital images of galaxies” (Wirth 1984). Classifying galaxies through the use of computer algorithms can offer an alternative classification system that may be applicable to all types of galaxies in the Universe. Automated classification methods can enable us to assign morphologies to a large number of galaxies, which will be beneficial using large surveys.

The principal of automated classification is to determine parameters that can separate galaxies into various physical categories based on their structure. As Wirth

(1984) writes, “[...] rather than attempt to interpret these derived parameters in terms of present classification systems or our current ideas of galaxy structure,” the main purpose of automatic classification is to develop a new classification system for galaxies.

Unlike people, machines are capable of classifying large amounts of data in short amounts of time, provided they have the processing power to do so. Computer-automated classification methods may be compared across different wavelengths and redshifts as well, meaning that an automatic classification system can be applied to high-redshift galaxies, which, as was mentioned, tend to not fall into any classification bins of most visual classification systems used today (Holwerda *et al.* 2014).

Computer classification is generally free of biases that often occur with human visual classification. But such classification is also limited by the resolution of images it uses. Also, currently computers are not capable of distinguishing subtle features that human classifiers can (Odewahn 1995). Developing software that can notice structural patterns and detect low surface brightness features is one of the challenges today. Nonetheless, terms of efficiency, universality, and impartiality, the use of computer software to classify and analyze the morphology of galaxies will be extremely important in the future.

2.2 Images and Charge Coupled Devices

Retaining visual evidence of galaxies has played an important role in their study. Currently, the method to preserve images of astronomical objects has been through the use of charged coupled devices (CCDs). Prior to CCDs, galaxies were observed by eye through telescopes and sketched out by hand. As was mentioned previously, William Parsons was one of the first to observe spiral structure in nebulae with his

“Leviathan of Parsonstown” telescope. In Figure 10(a.), his sketch of the Whirlpool galaxy (M51) and its companion, dwarf galaxy NGC 5195, can be seen at the top (Image reproduced from Bailey *et al* 2005) and compared to Figure 10(b.) below of the galaxies observed with a CCD camera aboard the HST³ in 2005.

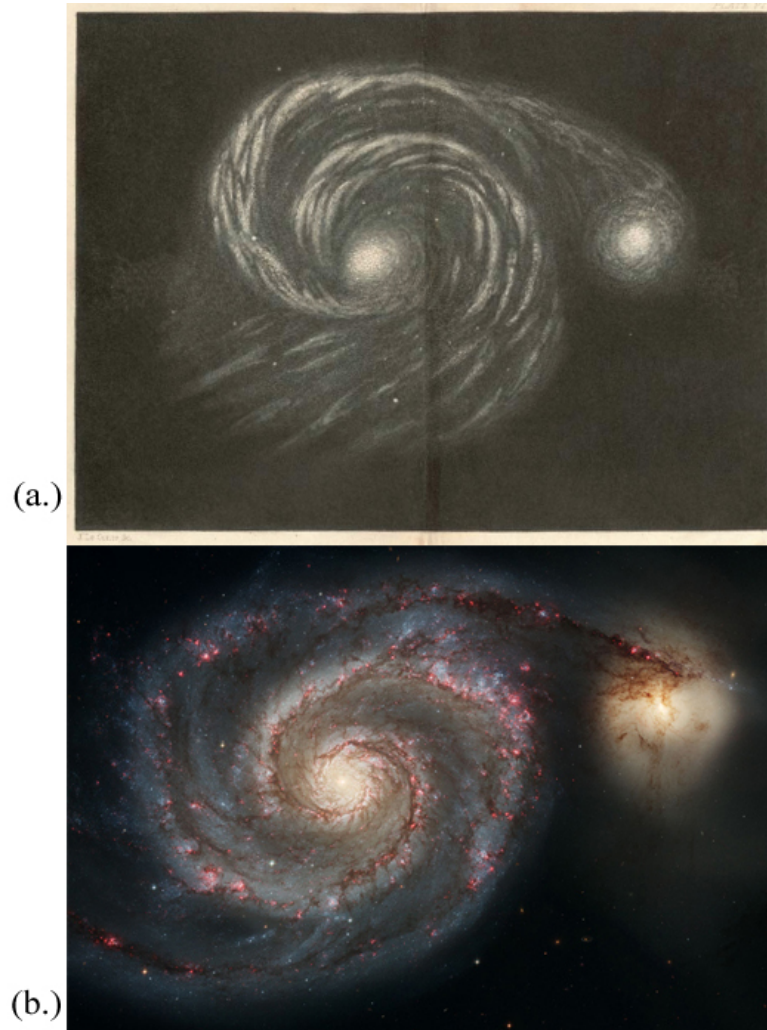


Figure 10: Comparison of William Parsons sketch of M51 and NGC 5195 galaxies (a.) to the image of the same galaxies observed by HST (b.).

Image recording of galaxies advanced to photography in the mid-19th century. Photography was the primary method of recording and analyzing astronomical ob-

³Image Source: NASA/ESA

jects till CCDs were introduced in the late 1970s. Initially, photographic images at the optical wavelengths were taken on silvered copper plates and required long exposure times of about 15 minutes (Belkora 2003). This was the so-called “dageurreotype” process. Later, the “wet collodion process” became more preferred for such photography. This process involves coating glass plates with collodion, a light-sensitive substance that becomes less light-sensitive as it dries. However, since this technique has to be done quickly and thus requires a portable darkroom, it was mostly replaced by gelatin dry plates in the late 1800s. When roll-films were introduced a few years later, glass plates still remained in use in astronomy due to their sturdiness. Plates do not bend or shrink like film and can therefore supply accurate images for measurements or other analysis.

Among the advantages of photographic emulsion on glass plates is their wide field of view, however, they have many disadvantages too. Besides requiring long exposure times, plates are also nonlinear, meaning that the number of developed crystal grains is not proportional to number of photons falling on the plate. Additionally, they have a low quantum efficiency of about 3%, meaning that for every 100 photons that fall on the plate only about three trigger a photochemical reaction. Photographic plates are also inherently more sensitive to blue wavelengths (*i.e.* shorter than 500 nm), but can be made red-sensitive through the use of certain dyes. Their blue-sensitivity may cause objects in images to appear not as they actually are.

Though photographic plates continued to be used by some observatories even till about the 1990’s, it was in the late 1950s when the forerunners to the modern imaging devices were developed. These photomultiplier tubes (PMTs) are generally non-imaging devices, meaning they do not record the spatial position of photons. Through the process of the photoelectric effect, these tubes convert incoming photons, that strike the surface of their photoemissive material, into electrons. In such a way,

measurement of an object's flux, or intensity of its light, can be accomplished. This technique is called "photometry."

When a single photon hits the detector surface, it ejects an electron which then gets accelerated through the tube. The initial electron dislodges additional electrons down the tube, creating a "cascading" effect that produces a strong signal at the final electrode of the tube. Therefore, PMTs can be used to measure even the faint light of astronomical objects. Similar to photographic plates, PMTs also have a higher sensitivity to blue-regions of the light spectrum than the red. The typical quantum efficiencies of PMTs are relatively high ($\sim 30\%$), therefore even today, these tubes are used in high-energy astronomy (study of objects that release high-energy photons). Position sensitive PMTs can also be used in nuclear particle detection physics.

CCDs were first developed in the late 1960's at Bell Laboratories as memory storage devices, but about a decade later began to be used in astronomy as well. CCDs are position-sensitive devices, made up of arrays of small metal oxide semiconductors (MOSs), that can be used to create and store images of objects or transfer electrical charge. Their sensitivity to light varies across a wide range of wavelengths (from X-rays to infrared) and compared to the quantum efficiency of the human eye — which is approximately 1%, meaning that only about one photon out of one hundred is detected — or that of photographic plates, they have a quantum efficiency generally over 80%. This makes them very useful in recording images of faint astronomical objects.

A typical CCD chip is arranged into rows and columns (*i.e.* an array) of small, light-sensitive MOSs called pixels, as seen in Figure 11(b.). Light falling on a pixel gets converted into an electron. When the CCD is not saturated by light (*i.e.* when the finite charge capacity of the individual pixels is not exceeded) it behaves as a perfect linear detector, meaning that the number of electrons collected on a pixel is

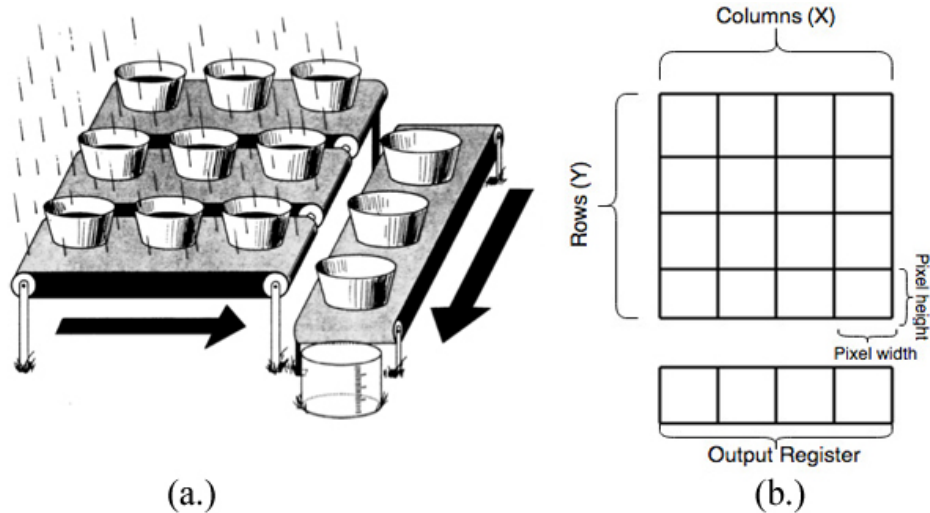


Figure 11: An analogy for CCDs is an array of buckets collecting rain water as seen in (a.) on the left. The CCD chip is an array of pixels (b.).

proportional to the amount of photons that fall on that pixel. A common analogy of the CCD chip is that of an array of buckets collecting rain water, as seen in Figure 11(a.) (Image reproduced from Howell 2006). The buckets are the individual pixels and the rain drops are the incoming photons.

There is currently a large amount of literature devoted to discussing the physics, function, and properties of CCDs (e.g. Martinez 1997; Carroll 2006; Howell 2006; McLean 2008; Clements 2014), but briefly the function of CCD chips can be summarized as follows: when photons from an astronomical object travel through the telescope’s optics and reach the CCD chip, they are absorbed by the (typically) silicon pixels of the chip. Through the process of the photoelectric effect, absorption of photons creates free negatively-charged electrons and leaves behind positively-charged “holes”. Each photon creates such an electron-hole pair. As the CCD remains exposed to light, electrons and holes created by the photons are stored in different areas of the pixel. Each pixel has a voltage applied to it in such a manner as to create a potential well, which will hold the freed electrons until the end of the exposure. After

the exposure ends, the shutter closes and the CCD readout begins. By varying the voltage across each pixel, the electrons are moved, row by row, to the output register. In these output electronics, the charge collected within each pixel passes through an amplifier, which boosts the electron signal and measures it as an output voltage. The output voltage is then converted to a digital number (DN) by the analog-to-digital converter (A/D converter). The DN uses the units of counts, otherwise known as analog-to-digital units (ADUs). The output voltage (meaning, the number of collected electrons) needed to produce 1 ADU is determined by the gain of the CCD. For example, if the gain of a CCD is 10 electrons/ADU, this means that for every 10 electrons collected by a pixel, the output from that pixel will have a DN value of 1 ADU. Finally, the shifting of each CCD row into the output register/output electronics, which convert each pixel's stored charge into a DN value, continues until the whole array of pixels is read out. For large pixel arrays, the readout time may be as long as several minutes.

Besides their high quantum efficiency, generally strong linearity, and sensitivity to a broad wavelength range (which is usually accomplished by introducing impurities into the silicon surface of the chip), CCDs also have very low noise. But one disadvantage of these chips is their usually small size. However, in order to capture a large area of the sky, this issue can be bypassed by placing many CCD chips in a grid-like formation (a mosaic) at the focal plane of a telescope. CCDs are now predominantly used in many fields of astronomy, including imaging, photometry, and spectroscopy.

The size of the CCDs varies. It is defined by the number of rows and columns of its pixel array. The size of pixels can also vary (they typically range from $10\text{-}30\mu\text{m}^2$). After all the pixels on a CCD detector have been read-out into a computer, the information is stored in a file. In astronomy, the standard format to save images from the CCDs is the FITS format (Flexible Image Transportation System). The FITS format

was developed in the 1970s and standardized in Wells *et al.* 1981. It was designed to store, transmit, and analyze n -dimensional data arrays, such as one-dimensional spectra, two-dimensional images, or three-or-more-dimensional data cubes. FITS files all conform to the same file structure — the Header and Data Units (HDUs) structure — meaning, the standard FITS file contains a primary ASCII header usually followed by a primary, uncompressed, data array. The header contains information about the size of the data array, the date it was recorded, the telescope, exposure time, gain of the CCD, and so on. The information is supplied in keywords. However, every FITS file is required to have the following five keywords:

1. SIMPLE: is a logical variable (True (T) or False (F)), which states if the file is a standard FITS file.
2. BITPIX: is an integer value that specifies the number of bits in the data values. The BITPIX value can be either 8, 16, 32, -32, -64.
3. NAXIS: specifies the dimension of the data array. If the FITS file contains just the header and no data that follows, then the NAXIS value is zero. If it contains image data, then NAXIS would be 2 (meaning it is a two-dimensional array). The maximum value of NAXIS is 999.
4. NAXIS n : designates the length of each axis in BITPIX units. For a two-dimensional data array, the number of columns (Y-axis) in the array would go to NAXIS1 and rows (X-axis) to NAXIS2.
5. END: necessary in order to indicate the end of the header, after which the data array begins.

The (x, y) location of the pixel on the CCD detector has a matching (x, y) position in the image data array. The image header starts in the following way: “the first image

pixel value will occur in the first pixel position in [the] first data array record” (Wells *et al.* 1981) and the same idea is applied to the rest of the data array. Among the software that can view FITS images are SAOImage DS9 and the ESA/ESO/NASA FITS Liberator. The software developed in this thesis reads images in the FITS format but does not view them like imaging viewing software.

2.3 Description of the Software

The morphological software developed for this thesis is a combination of routines and modules written in FORTRAN 90. The digital image of an object is input into the program and measurements of five different parameters are output. The program is capable of measuring classification parameters from a single postage-stamp of a galaxy, as well as from an image of a full cluster of galaxies. In order to read data files in FITS format, the program uses the CFITSIO library: <https://heasarc.gsfc.nasa.gov/docs/software/fitsio/fitsio.html>

CFITSIO is a library of C and FORTRAN subroutines for reading and writing FITS data files. By default, CFITSIO library can read FITS files up to 2.1 GB in size. Since a number of our data sets include image files larger than this limit, the library had to be modified. As described in the CFITSIO User’s Reference Guide (2010), additional compiler flags on Linux systems need to be included in the CFITSIO Makefile in order to achieve large file support. The flags are ‘-D_FILE_OFFSET_BITS=64’ and ‘D_LARGEFILE_SOURCE’. After compiling the library with these additional flags, the maximum FITS file size supported by CFITSIO is 6 terabytes (containing 2^{31} FITS blocks, each 2880 bytes in size).

As seen in Figure 12, after the program opens the image in FITS format, the image is read as an array of pixels. A separate DAT file containing descriptive information

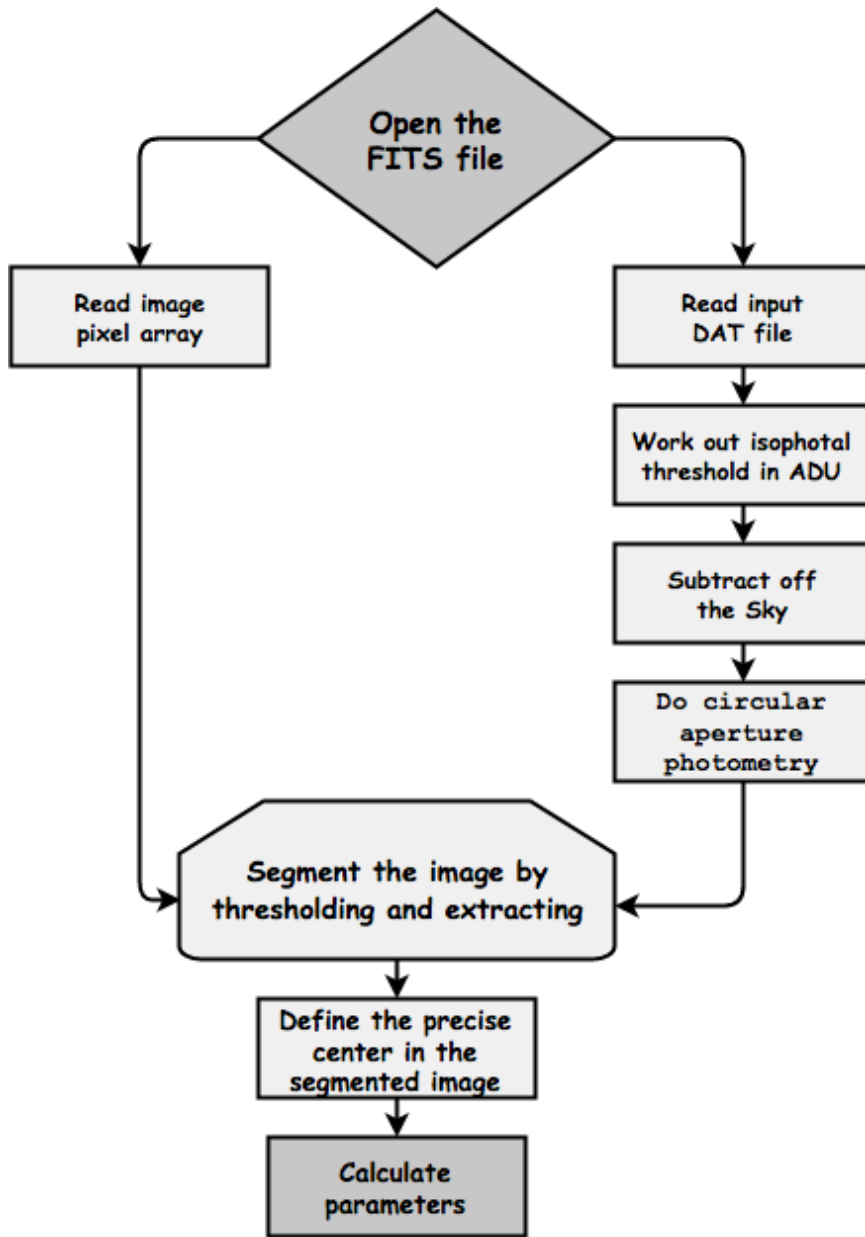


Figure 12: Flowchart of the main procedures of the morphological software developed for this thesis.

about the image is also read. The DAT file must be manually assembled for each galaxy cluster or galaxy studied. The information supplied in the DAT file is as follows:

1. Sky level (SKY): the pixel values (in units of counts) of the background around the object. This value is found for each galaxy cluster or single galaxy FITS image using the “m” feature of the *imexamine* command in the Image Reduction and Analysis Facility (IRAF)⁴ software. The “m” feature measures the statistics of a 5x5 pixel region of the sky in each image. It prints the following values: image section, the number of pixels, the mean counts, the median counts, the standard deviation, the minimum, and the maximum count values. To find the SKY value for a FITS image, we take the average of the mean values of each image section measured.
2. Sky rms (σ): the root-mean-square variation of the sky level (in units of counts). This value is also found using the “m” feature of the *imexamine* command in IRAF. To find σ , we average the standard deviation of each section of the sky measured.
3. Pixel scale: the measurement of the amount of sky covered in one pixel (in units of arcseconds/pixel). This value is found in the header of the image or in the telescope description.
4. Zero Magnitude (ZEROMAG): the counts required to give the object an apparent magnitude of zero. This value is found by using the equation: $m = K - 2.5 * \text{Log}(\text{FLUX})$, where K is the photometric zero point and $\text{FLUX} = (\text{GAIN} * \text{ZEROMAG}) / \text{EXPTIME}$. The photometric zero point of an image is the magnitude that produces one count per second. Additional to GAIN and EXPTIME, *i.e.* the exposure time of the image, photometric zero point can be found in the header of the FITS image. Setting the magnitude, m, to zero, this equation can then be solved for the ZEROMAG (which is in units of counts).

⁴<http://iraf.noao.edu/>

5. Object X-center: the center of the object along the x-axis, in units of pixels.
6. Object Y-center: the center of the object along the y-axis, in units of pixels.
7. Surface Brightness Threshold (μ_T): The surface brightness is the measure of brightness per area on the sky, in units of magnitude/arcseconds². It defines the minimal brightness threshold, *i.e.* isolates pixels that are brighter than this threshold. We use our morphology software to measure the value of 1σ above the sky level. Once the value of 1σ above the sky is known (also called μ), the surface brightness threshold is found by subtracting two from μ , meaning: $\mu_T = \mu - 2$. This definition of μ_T has been chosen empirically.

From the DAT file, the program then uses the defined isophotal threshold around the specified center of the object to isolate the object, meaning, the pixels belonging to the object are separated from the sky and other objects surrounding it. The threshold is found in units of counts (ADUs) above the specified sky level. Furthermore, the SKY values are subtracted from the data, leaving just the pixel values of the observed object. Aperture photometry is then performed to further narrow down the pixels belonging just to the object being measured.

Aperture photometry is the measurement of light that falls inside a particular aperture. In this case, the circular aperture radius is empirically set to 3 arcseconds and then divided by the pixel scale specified in the DAT file to convert the radius to units of pixels. After the circular photometry subroutine is completed, the image of the object is segmented (partitioned into sets of pixels that either belong to the object, the sky, or background objects) by thresholding and extracting. The pixels of the object are separated from the sky and background pixels by using the pre-defined isophotal threshold. A new center of the object is found by precisely locating the pixel with the maximum value, and the classification parameters are then calculated.

Each classification parameter is a subroutine that reads the processed image as an array of pixels.

For large clusters, the x and y positions of the center of each galaxy, along with the rest of the information such as the SKY, σ , etc., can be stated in the DAT file. Once the software measures the five parameters for one of the galaxies in the cluster, it will move on to the next object in the list contained in the DAT file. Finally, the results can be printed to the terminal screen or saved to a file.

2.4 Parameters

The computer software used in this thesis has been developed to classify galaxies by measuring specific parameters extracted from digital images. It uses nonparametric methods to measure galaxy light distributions. Nonparametric methods are used when much is not known about the structure of astronomical objects studied or their processes. In our software, we implement computer algorithms to calculate five physical parameters: the central concentration, asymmetry, the Gini coefficient, the Thiel index, and M20. These parameters are defined in ways that stress major structural features of astronomical objects without making assumptions about their structure, such as the existence of bulges or disks in their systems.

However, we also use a parametric method by finding the light ratio between the bulge and the disk components of galaxies through GALFIT (Peng *et al.* 2002; Peng 2003). The parametric method relies on fitting light profiles to galaxies and their features. In this method, profiles of galaxies are found by first measuring how the average intensity of their light changes as a function of radius and matching these profiles to theoretical models, such as the Sèrsic profile. The bulges of galaxies can be fit by certain functions and the disks by others, and in this manner constraints are placed

on the object’s morphology. In this thesis, we focus mainly on the nonparametric method of measuring the five physical parameters.

2.4.1 Central Concentration

The central concentration is a ratio of light measured between an inner radius and an outer radius of the galaxy, and is usually represented by the symbol C . The lengths of the radii can vary, therefore several definitions for this index exist. The inner radius must be large enough to enclose the pixels in the central part of the object, while the outer radius must not enclose objects outside the system observed. The C parameter is correlated with the bulge-to-disk ratio, which in turn, is correlated with the morphological class of galaxies. Therefore, C is a popular parameter to measure during automatic classification.

One of the definitions of the central concentration is as follows (Bershady *et al.* 2000; Lotz 2004):

$$C = 5 * \text{Log} \left(\frac{r_o}{r_i} \right), \quad (2.1)$$

where r_o and r_i are the outer and inner circular radii, respectively. Usually, the outer radius, r_o , is defined to contain 80% of the total flux, while r_i encloses the inner 20%, or 90% for r_o and 50% for r_i (Ferrari *et al.* 2015). The other common ratio is 70% for r_o and 30% for r_i .

In our Morphological software, we apply the Abraham *et al.* (1994) definition. In this definition, the apertures around the object studied are elliptical, rather than circular. The sky background, which contains the foreground and any objects in the foreground that do not belong to the object being measured, has to first be subtracted

from the image of the object and then the second-order image moments are used to determine an elliptical light distribution. In this way, first the second-order moments of the object in the image are calculated and then an ellipse with the same second-order moments is found. An image moment is defined as a weighted average, *i.e.* a moment, of the pixel intensities in a digital image. The second-order image moments are then defined as follows:

$$M_{xx} = \frac{\sum_i \sum_{j \in A} x^2 I_{ij}}{\sum_i \sum_{j \in A} I_{ij}}, \quad (2.2)$$

$$M_{xy} = \frac{\sum_i \sum_{j \in A} xy I_{ij}}{\sum_i \sum_{j \in A} I_{ij}}, \quad (2.3)$$

$$M_{yy} = \frac{\sum_i \sum_{j \in A} y^2 I_{ij}}{\sum_i \sum_{j \in A} I_{ij}}, \quad (2.4)$$

where x and y are positions of pixels relative to the center of the object in the image, A is the area of the galaxy that is enclosed by an isophote at about 2σ above the sky level, and I_{ij} is the intensity of a pixel in position (i, j) . The second-order image moments are then used to determine an elliptical light distribution that has the following equation:

$$r^2 = M'_{xx}x^2 - 2M'_{xy}xy + M'_{yy}y^2, \quad (2.5)$$

where r is the normalized radius, and the primes on the second-order moments indicate that they have been divided by a normalization constant. The second-order image moments have been normalized in such a way that $r = 1$ when the area en-

closed by the ellipse is equal to A , *i.e.* $E(r = 1) = A$. Then the definition of central concentration becomes:

$$C = \frac{\sum_i \sum_{j \in E(\alpha)} I_{ij}}{\sum_i \sum_{j \in E(1)} I_{ij}}, \quad (2.6)$$

where I_{ij} is the intensity of a pixel in position (i, j) , $E(\alpha)$ is the inner normalized elliptical radius, and $E(r = 1)$ is the outer elliptical radius that has been normalized to 1. The inner radius isolates the flux within the cores of galaxies, which Abraham *et al.* (1994) have found empirically that the “shrinking factor” of $\alpha=0.3$ produces the best results based on simulations. A value of $\alpha=0.3$ signifies that the inner radius contains about 30% of the total object flux. Too small of a value for α would make C sensitive to seeing (size of PSF).

A high C value indicates that the object has a high density of light in its central region. It is found that many early-type galaxies have high values of central concentration, while many late-types have low values. In order to chose the radii for the calculation of this parameter, the center of the galaxy must be specified. Finding the center of an object is one of the challenges of automated classification, because there is not a clear definition of what the center should be. In most cases, the center is defined to be the brightest pixel in the galaxy, however, for irregularly shaped objects, the position of such a pixel is not always the visual center. This definition also creates a problem for the classification of distant galaxies, many of which are irregularly shaped. In order to find C , we define the center to be the geometric center of a galaxy.

From Equation 2.6, values of C range from a minimum of $C=0$ and a maximum of $C=1$. Values of C that are closer to one indicate high central concentration of

light in the galaxy. Early-type galaxies tend to have high C values because the bulk of their light is located in a small radius around their nucleus. These galaxies have bright central regions and extended outer envelopes, unlike many late-type galaxies. Values of C that are closer to zero tend to represent late-type galaxies.

2.4.2 Asymmetry

Asymmetry (A) of galaxies refers to how evenly light is distributed throughout the system. Just as with C , the center of the object is first found, *i.e.* the geometric center of a galaxy. The object is then rotated by a 180° angle about this center. The center of the rotated object is aligned with the center of the non-rotated image of the same object, and the rotated image is subtracted from the non-rotated image. The result of this process is an image of the asymmetric components in the object. These residuals are the pixel-by-pixel differences between the original non-rotated image and the rotated image, *i.e.* $|I(i, j) - I_{180}(i, j)|$ as defined below.

On average, half of the residual values will be positive and half negative, therefore, one mathematical way to define A is through Conselice (1997; *et al.* 2000), where A is the square root of the sum of squares of the residuals divided by twice the sum of the intensity distribution of the non-rotated image.

However, in our software, we follow the Abraham *et al.* (1996b), Brinchmann *et al.* (1998), and Povic *et al.* (2015) definition, where the sum of squares is replaced by the absolute values:

$$A = \frac{\sum_{i,j} |I(i, j) - I_{180}(i, j)|}{2 \sum_{i,j} |I(i, j)|} - \frac{\sum_{i,j} |B(i, j) - B_{180}(i, j)|}{2 \sum_{i,j} |B(i, j)|}, \quad (2.7)$$

where $I(i, j)$ is the intensity distribution of the non-rotated image, $I_{180}(i, j)$ is the

intensity distribution of the image rotated by 180° about its center, $B(i, j)$ is the original background sky within the area of the object being analyzed, and similarly, $B_{180}(i, j)$ is the rotated background sky.

In both methods, A has a maximum value of $A=1$ and a minimum value of $A=0$. Objects with $A = 1$ are completely asymmetric, since it means that $|I(i, j) - I_{180}(i, j)| = I(i, j)$. On the other hand, $A=0$ represents objects that are completely symmetric, since in this case $|I(i, j) - I_{180}(i, j)| = 0$. Most objects have A values that are in-between these two extremes. Galaxies with values closer to $A = 1$ tend to appear more spiral or irregular in shape, while ones with A values closer to zero tend to be elliptical.

Similar to Nair (2009), we find that each galaxy type spans a wide range of A values, therefore asymmetry alone cannot be used to distinguish galaxy types. For example, the wide range of A values for various galaxy types can be seen in the histogram in Figure 18 (b). Therefore, asymmetry must be compared to other parameters, as demonstrated in Figure 13.

Comparison of the A parameter to the C parameter is frequently done in galaxy morphological studies. Figure 13 shows a plot of asymmetry vs. central concentration for a sample of classified galaxies in the Hubble Deep Field (HDF) analyzed by Abraham *et al.* (1996a). The HDF is a 2.6×2.6 arcminute² region of the sky located in the constellation of Ursa Major and observed by the HST. The figure represents data observed through the I-band (F814W) and shows a sample of approximately 300 visually classified HDF galaxies. Most of these galaxies are fairly distant ($z > 0.3$), however, HST offers enough resolution for accurate visual classification. Using diagonal lines, three sectors were defined on the diagram by Abraham *et al.* (1996a) from their visual classification of the galaxies. Visual classifications were performed by eye and are indicated by the plot symbols, where the early-type galaxies (*i.e.* E/S0

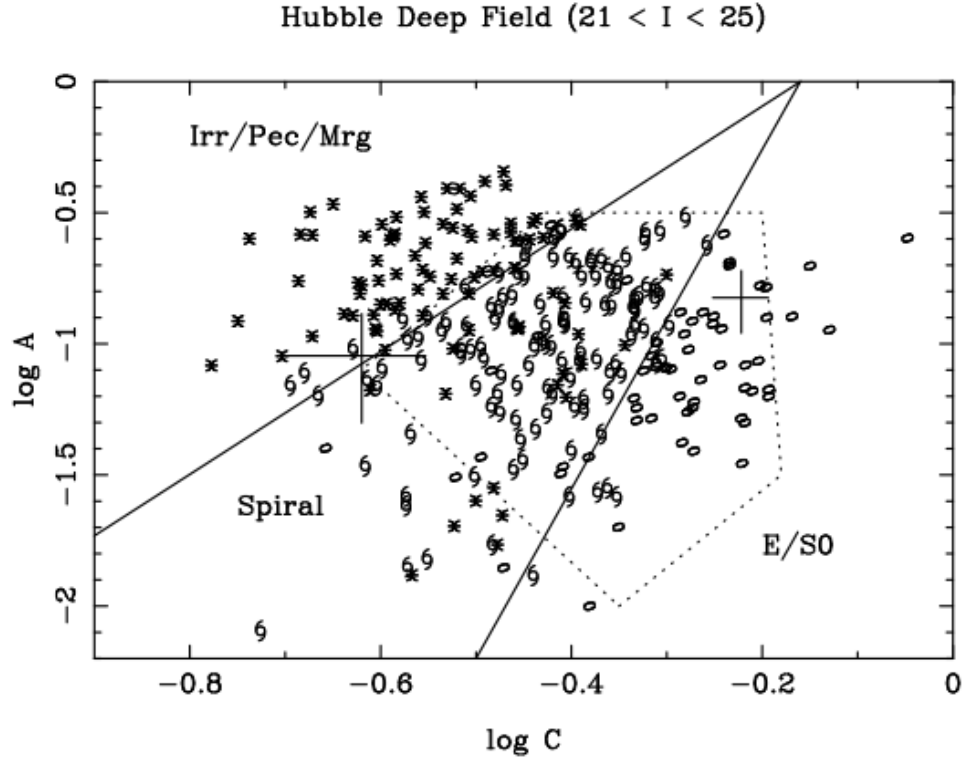


Figure 13: Plot from Abraham *et al.* (1996a) of $\text{Log}(A)$ versus $\text{Log}(C)$ for Hubble Deep Field data.

galaxies) are shown as ellipses, spirals earlier than Sd are shown as spirals, and irregulars/peculiar/mergers are represented by asterisks. The sectors divide the data into three broad categories: irregular/peculiar/mergers (Irr/Pec/Mrg), spirals (S), and ellipticals/lenticulars (E/S0). Using a similar method, we will introduce our own classification regions for data-sets throughout this thesis.

2.4.3 Gini Coefficient

In astrophysics, the Gini coefficient (G) is a statistic that describes how uniformly light is distributed among the pixels of a galaxy, *i.e.* the inequality in the distribution of galaxy pixel values. It is based on the Lorenz curve (Lorenz 1905), which is used to describe the economic inequality in a population's distribution of wealth. The

Lorenz curve shows the proportion of total income earned by a given percentage of the population. Given a sample of n individuals, let X_i be the wealth (income) of the individual i , with $i=1\dots n$. Graphically, the Lorenz curve is then the cumulative proportion of wealth of individuals plotted against the population sorted from lowest to highest income, as shown in Figure 14.

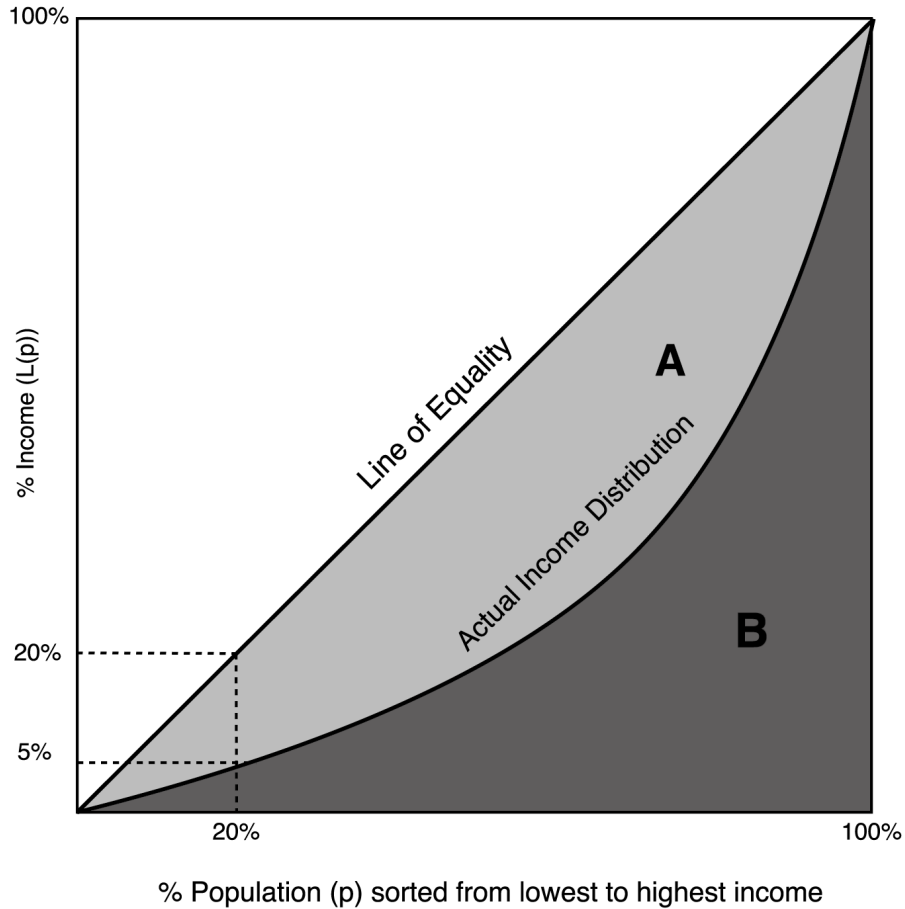


Figure 14: Graphical representation of the Lorenz Curve.

If all individuals in a population have exactly the same wealth (for example, if the poorest 20% of a population have 20% of the total wealth), then the Lorenz curve is a straight diagonal line with a slope of one (also called the line of equality). However, in reality, the actual income distribution is not usually this equal, therefore the Lorenz

curve of the actual income distribution will fall below the line of equality. In this case, it may be that the poorest 20% of a population only earns 5% of the wealth.

The further away the Lorenz curve is plotted from the line of equality, the more unequal the distribution of income among the population. The shaded region A, located between the line of equality and the curve of actual income distribution in Figure 14, represents the inequality gap. Mathematically, the Lorenz curve, $L(p)$, is defined as:

$$L(p) = \frac{1}{\bar{X}} \int_0^p F^{-1}(u) du, \quad (2.8)$$

where \bar{X} is the average of all individual incomes (or galaxy pixel values) X_i , p is the percentage of the population (or pixels), and $F(x)$ is the cumulative distribution function of a positive random variable X .

The Gini coefficient measures the degree of wealth inequality in a population. There are various ways of defining G . From Figure 14, G can be thought of as the ratio between the region A and the total area beneath the line of inequality, (A+B). Thus, the ratio is: $G = A / (A + B)$

The value of G varies from $G=0$ to $G=1$, where $G=0$ represents total equality, meaning every individual in the population has the same income and $G=1$ represents total inequality, meaning all the wealth belongs to a single person. In terms of astronomy, $G=0$ indicates that the galaxy's light is evenly distributed among the galaxy's pixels, while $G=1$ means that all the light in the galaxy is concentrated in a single pixel.

Another way to define Gini is as half of the relative mean of the absolute difference

between all X_i :

$$G = \frac{1}{2\bar{X}n(n-1)} \sum_{i=1}^n \sum_{j=1}^n |X_i - X_j|, \quad (2.9)$$

where \bar{X} is, again, the average of all individual incomes (or galaxy pixel values) X_i , and n is the total number of individuals in a population (total number of galaxy pixels).

However, a more efficient way to calculate G is by first sorting the pixels of the galaxy (X_i) by increasing intensity and then summing:

$$G = \frac{1}{\bar{X}n(n-1)} \sum_{i=1}^n (2i - n - 1) X_i. \quad (2.10)$$

It has been suggested that Gini can be used as an alternative to the central concentration parameter (Abraham *et al.* 2003). This may be very beneficial in the automatic classification of galaxies since — because of the way it is defined — the Gini coefficient is applicable to galaxies of arbitrary shape. Unlike with the central concentration parameter, in order to measure the Gini coefficient the center of the galaxy need not be defined, *i.e.* Gini is independent of the spatial distribution of a galaxy’s light (Abraham *et al.* 2003; Lotz *et al.* 2004).

In our software, we measure the G values for each galaxy in our data-sets by using Eq. 2.9 and Eq. 2.10. As in Abraham *et al.* (2003), these values are measured for galaxy images that are sky-subtracted and have background objects removed. Gini is then measured for the galaxy pixels that are above a constant surface brightness threshold.

2.4.4 Theil Index

Theil index (T_T) is a new statistic that we have adopted for morphological measurement in this thesis. Historically, the Theil index is an alternative to the Gini coefficient. It is a statistic that is commonly used to measure income inequality in a population. However, in the context of this thesis, it is used to measure the inequality in the distribution of light within individual galaxies. The index is based on information theory (Theil 1967; Alison 1978), and is defined as:

$$T_T = \frac{1}{N} \sum_{i=1}^N \left(\frac{x_i}{\bar{x}} \cdot \ln \frac{x_i}{\bar{x}} \right), \quad (2.11)$$

where, in our context, \bar{x} is the average of the flux (light) of the pixels in the galaxy, x_i is the flux of a single pixel, and N is the number of pixels in a galaxy.

In terms of economics, if every individual in a population has the same income, then $T_T=0$ (considered the maximum disorder), while if only one individual has all the income, $T_T=1$ (maximum order). Higher values of T_T indicate more order, while lower values of T_T indicate less order. In other words, it is the measure of “entropic distance the population is away from the “ideal” egalitarian state of everyone having the same income.”⁵

In astronomy, this index can be used to measure lack of balance in the light distribution within galaxies. The inequality in light distribution within galaxies can then be compared to the central light concentration and asymmetry, and contrasted with results from the analysis with the Gini index.

Besides the mathematical differences in their definitions, the Gini coefficient and the Theil index describe slightly different schemes economically. As we have seen in

⁵https://en.wikipedia.org/wiki/Theil_index

EFIGI data: Gini vs Theil

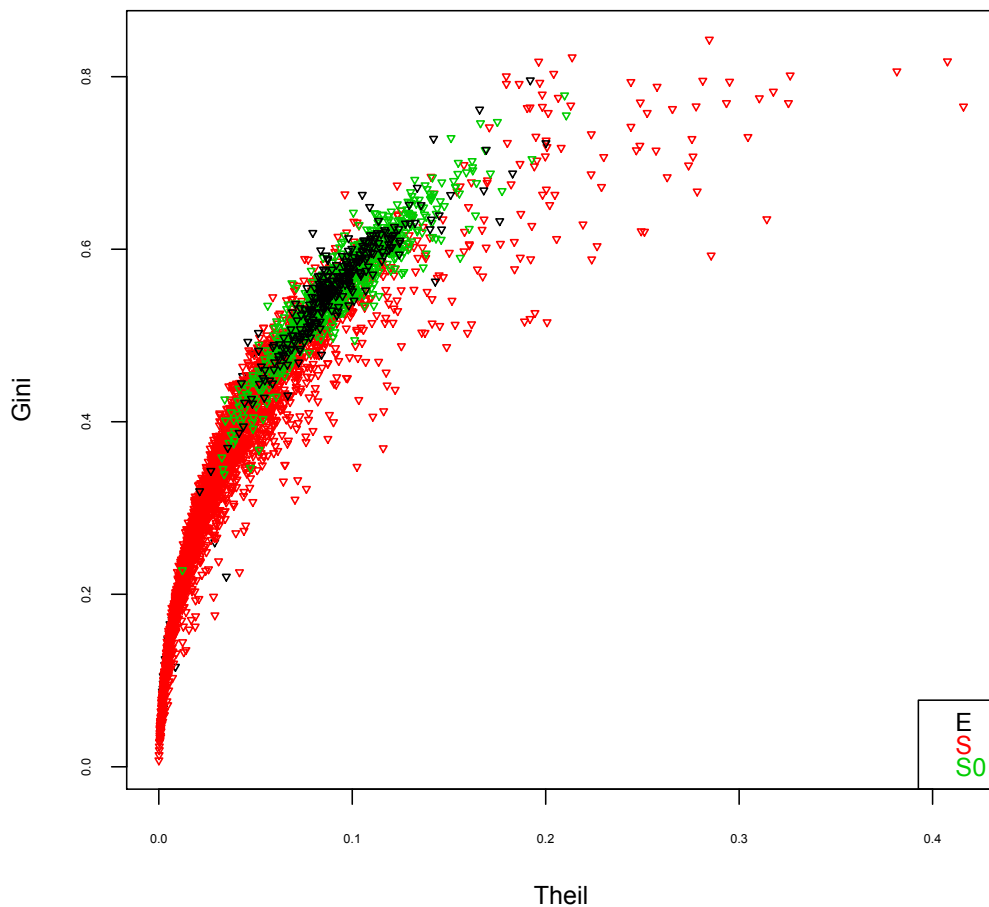


Figure 15: The Gini vs. Theil plot of EFIGI data

the last section, the Gini coefficient measures the distribution of wages/income (or light) over a population (or galaxy). However, in economics, one drawback of the Gini coefficient is that it does not offer any information about how the population influences the distribution of income. The Theil index, on the other hand, takes into account how the distribution of wealth among various groups in the population is affected by the size of the groups. Basically, “the Theil Index captures the same information that the Gini coefficient would if applied to wages, but in addition provides information on

the contribution of each group to the total level of wage inequality in the population” (Cozzens & Bobb 2003). In this thesis, we are testing the usage of the Theil index in galaxy morphological classification. We include a sample of the FORTRAN code used to measure the Theil index in our software in Appendix A.

We compare the Thiel index to the Gini coefficient and other parameters throughout this thesis. In Figure 15, we plot the Gini versus the Theil values of 4,352 galaxies from the Extraction de Formes Idealisées de Galaxies en Imagerie (EFIGI)⁶ project, which we will describe in more detail in a later chapter. We group the eighteen EFIGI morphological types (Baillard *et al.* 2011; de Lapparent *et al.* 2011) into three broad categories: E, S, and S0, where the E class includes the EFIGI Hubble sequence (EHS) types (described in Chapter III) 10 and -6 through -4, the S class includes EHS types 0 through 10, and S0 are the EHS types -3 through -1. From the figure, it can be noted that the Gini coefficient and the Theil index are strongly correlated for $T_T < 0.1$ and $G < 0.3$.

2.4.5 M20

The normalized second-order moment of the brightest 20% region of a galaxy (usually referred to as M20) is the most recent parameter we have incorporated into the software. It is the relative contribution of the brightest 20% of the pixels in the galaxy. The parameter was introduced by Lotz *et al.* (2004). In order to measure M20 for a galaxy, the total second-order moment (M_{total}) — which is the summation of each individual pixel flux (I_i) multiplied by its squared distance to the center of

⁶<https://www.astromatic.net/projects/efigi>

the galaxy — is found first:

$$M_{total} = \sum_i^n M_i = \sum_i^n I_i [(x_i - x_c)^2 + (y_i - y_c)^2], \quad (2.12)$$

where I_i is the intensity of pixel i in the image and (x_c, y_c) are the coordinates of the center pixel of the galaxy. The center can be found by varying the values of (x_c, y_c) such that M_{total} is minimized (e.g. Lotz *et al.* 2004; Bendo *et al.* 2007), however we follow an approach similar to Holwerda *et al.* (2014) where variation in the center are treated as a source of uncertainty. Our program, therefore, uses the geometric center of each galaxy.

Afterwards, M20 can be computed by ranking the galaxy pixels in descending order (*i.e.* I_1 is the brightest pixel, I_2 is the second brightest, etc.) and summing M_i over the brightest pixels until the sum of the brightest pixels equals 20% of the total flux of the galaxy. Finally, in order to have a parameter that is independent of total galaxy flux or size, the sum of M_i is normalized by M_{total} :

$$M_{20} = \sum_i \frac{M_i}{M_{total}}, \text{ while } \sum_i I_i < 0.2I_{total}, \quad (2.13)$$

where I_{total} is the total intensity of the pixels in a galaxy. Note that we follow the Nair (2009) definition of M20 where we do not take the log of the ratio. The value of 20% was chosen empirically. It was found that using brighter flux thresholds, such as 5% of I_{total} , produces second-order moment values that are unreliable at low spatial resolutions (Lotz *et al.* 2004).

Similarly to C , M20 correlates with the light distribution throughout the galaxy.

It shows how light is distributed in features like the center of the galaxy, but unlike C , it also accounts for off-center light sources such as bars, spiral arms, and so on, since it is weighted towards the luminous regions (Povic *et al.* 2015; Lotz *et al.* 2004; Scarlata *et al.* 2007). Unlike C , the M20 parameter is not measured through circular or elliptical apertures. This makes M20 more sensitive to merging features of galaxies compared to C (Lotz *et al.* 2004).

2.4.6 B/D and B/T Ratios

Historically, two important morphological features of galaxies have been the bulge and the disk. Not all galaxies exhibit these features, but the extent to which they possess them can be applied to classification. In the last chapter, we mention that the comparison of the concentration of light in the bulge of the galaxy to its disk is one of the characteristics used to classify galaxies in schemes such as the Hubble classification system. Therefore, another set of parameters that can be found for each galaxy are its bulge-to-disk (B/D) and bulge-to-total (B/T) ratios. We will briefly define galactic bulges and disks in this section.

As we know, galaxies are generally defined as large systems consisting of millions of stars gravitationally bound together with gas, dust, and dark matter. For many galaxies — particularly ones nearby — stars, gas, dust, etc., organize into bulge and/or disk system. A bulge of a galaxy is a round, tightly-packed structure containing primarily old stars, gas, and dust. Visually, it is generally evident that these structures are located at the center of most galaxies and may have formed via monolithic collapse (see Figure 9). Disks, on the other hand, are flat structures that can sometimes surround bulges. They are supported by the rotation of the galaxy and contain much higher levels of gas and dust than bulges. Star formation usually occurs in disks, therefore these structures tend to contain a larger fraction of young stars.

In some cases, there may be structures that appear to be a combination of disk and bulge — the so-called disk-like bulges. Disk-like bulges are flatter than the classical bulge structures and may contain star formation. The study of the properties and development of bulges and disks is still on-going, with much literature currently being published on the subject (e.g. Kormendy & Kennicutt 2004; Athanassoula 2005; Gadotti 2009; Houjun 2010). Disks may have formed as part of the hierarchal collapse scenario (Figure 9).

In this thesis, we use the GALFIT software to measure (B/D) and (B/T) ratios of galaxies in our data-sets. GALFIT⁷ (Peng *et al.* 2002; Peng 2003) is a two-dimensional fitting algorithm that analyzes light from astronomical objects in digital images. It uses functions to fit a galaxy’s light profile. The functions are defined by a set of free fitting parameters that are selected to fit the light distribution of the galaxy in an image. In this study, we use the Sèrsic profile to fit the data, which is a power law of the form:

$$I(r) = I_e \cdot \exp \left[-\kappa \left(\left(\frac{r}{r_e} \right)^{\frac{1}{n}} - 1 \right) \right], \quad (2.14)$$

where I_e is the pixel surface brightness at effective radius (r_e), which contains half the total luminosity, and n is the concentration parameter (also known as the “Sèrsic” index), which controls the degree of curvature of the profile. κ is a positive parameter that is defined in terms of n .

The Sèrsic profile is a general mathematical function that describes how the intensity of a galaxy varies with distance from the center. By varying the n value in the exponent, we can derive different profiles that describe distinct morphological features

⁷<https://users.obs.carnegiescience.edu/peng/work/galfit/galfit.html>

of galaxies. Large values of n produce profiles that fit a more centrally concentrated structure (such as a bulge of a galaxy), while small values of n produce less centrally concentrated profiles. Setting $n=1$ gives a special case of the Sèrsic profile, called the exponential disk profile, that best describes the distribution of light in the disk component of galaxies, while $n=4$ gives the de Vaucouleurs profile that describes the distribution of light in the bulge component of galaxies. The de Vaucouleurs profile is also a special case of the Sèrsic profile (Peng *et al.* 2002; Concejo 2009).

In GALFIT, there are two ways to input initial parameters: manually or by an input file. Since we measure thousands of galaxies, it is practical to use an input file, as well as Perl scripts, to apply GALFIT to every galaxy in our data sets. Using Perl, we define the individual galaxies' FITS image, the SKY values of the images ($\$sky_value$), zero-point magnitude values of the images ($\$mag_ZP$), and pixel scale of the images ($\$pixscale$) in the GALFIT input file, which returns an output data image block for each galaxy ($\$galaxy.out.fits$). An example of the GALFIT input file is given in Appendix B.

The input file in Appendix B contains two sections: image parameters (items labeled A-P, which relate to the FITS image of a galaxy), and the initial object and sky fitting parameters (that contain the x, y position of the galaxy, the object profile (*i.e.* the object type), position angle, and so on). In order to find the (B/D) and (B/T) ratios, we specify two objects for each galaxy — the “expdisk” that checks the extend to which a galaxy possess a disk structure in its system, and the “sersic” object that represents fits to the bulge of a galaxy.

After applying GALFIT to each galaxy image, we compute the contribution of each component (the bulge and disk) to the total flux of the galaxy. From the relation between absolute magnitude (M) and flux (Φ), we can derive the (B/D) and (B/T)

ratios as follows (Concejo 2009; Delgado 2010):

$$\frac{B}{D} = \frac{\Phi_{bulge}}{\Phi_{disk}} = 10^{-0.4(M_{bulge} - M_{disk})}, \quad (2.15)$$

$$\frac{B}{T} = \frac{\Phi_{bulge}}{\Phi_{bulge} + \Phi_{disk}} = \frac{10^{-0.4M_{bulge}}}{10^{-0.4M_{bulge}} + 10^{-0.4M_{disk}}}. \quad (2.16)$$

In general, it has been found that $B/D > 1$ for elliptical galaxies and $B/D < 1$ for spiral and irregular galaxies (Weinzirl 2009; Graham 2001; Trujillo *et al.* 2000). Similarly for the bulge-to-total ratio, objects with higher prominence of bulges (such as elliptical systems) tend to have larger values of B/T , while late-type systems tend to have smaller values of B/T . A pure disc has $B/T = 0$.

CHAPTER III

TRAINING DATA SETS

3.1 SDSS and Galaxy Zoo: Early Tests

The morphological software developed in this thesis has been checked for accuracy by comparing results from the software with classifications for the same galaxies based on published visual classifications. Among the data used to test the program’s ability to classify were galaxies from the Sloan Digital Sky Survey (SDSS), which is a project that has created the most comprehensive three-dimensional map of a section of the Universe to date. The SDSS consists of a multitude of data and digital images, which are also utilized by the Galaxy Zoo Project — the citizen project in which galaxies from the SDSS are classified by eye by a large number of people throughout the world. The classification results from my morphology code were compared to classifications made by Galaxy Zoo. This method has been found to be effective in developing and training the code.

In order to test and train the software, we obtained r -band images from the SDSS Image List Tool¹ and the SDSS Science Archive Server (SAS)². Figure 16 displays a few examples of the multi-band color JPEG thumbnails of several individual galaxies tested. Table 4 provides the coordinate information and the nonparametric values measured for these galaxies by our software. These values were compared to Nair (2009). For their measurement of central concentration (referred to as C_{Nair} in Table

¹<http://cas.sdss.org/dr7/en/tools/chart/list.asp>

²<https://dr9.sdss.org/>



Figure 16: Examples of various SDSS images used to test and train the software. Each thumbnail includes the galaxies’ spID and the official SDSS object designation number.

4), Nair (2009) utilized the Abraham *et al.* (1994) definition of the central concentration, which is the same definition we employ in our software. Our definition for the Gini coefficient (Gini) is also the same as Nair (2009), which we refer to as G_{Nair} in Table 4. From Table 4, it can be seen that there is strong agreement with the values measured by our software and that of Nair (2009).

To further train, test, and develop our code, another set of data studied in this thesis is the EFIGI catalog. In this chapter, we discuss the method and the analysis of the results from the testing and training of the program.

Table 4: Image and galaxy information of several SDSS galaxies used in the testing and training of our morphology software. C , A , Gini, and Theil are the nonparametric quantities measured by our software, while C_{Nair} and G_{Nair} values are from Nair (2009).

| spID | RA | DEC | C | A | Gini | Theil | C_{Nair} | G_{Nair} |
|---------------|---------|--------|-------|-------|-------|----------|-------------------|-------------------|
| 277-51908-2 | 166.289 | -0.796 | 0.377 | 0.067 | 0.522 | 6.95E-02 | 0.396 | 0.554 |
| 280-51612-5 | 171.397 | -0.768 | 0.543 | 0.037 | 0.602 | 0.106 | 0.577 | 0.671 |
| 285-51930-309 | 178.909 | -0.77 | 0.375 | 0.053 | 0.560 | 9.34E-02 | 0.401 | 0.55 |
| 285-51930-103 | 180.367 | -0.718 | 0.389 | 0.072 | 0.339 | 2.58E-02 | 0.27 | 0.418 |

3.2 EFIGI Catalog

3.2.1 Description

In order to train and test our morphological software — as well as serve as an aid in the analysis of our data — we include the Extraction de Formes Idealisées de Galaxies en Imagerie (EFIGI) catalog into our study. The catalog data tables, SDSS images through various filters, PSF images, and color PNG images are publicly available on the project website: <https://www.astromatic.net/projects/efigi>. Further description of the catalog can be found in Baillard *et al.* (2011), de Lapparent *et al.* (2011).

The EFIGI project is a catalog of 4,458 galaxies, each with their own SDSS digital images and detailed visual morphological information. The aim of the catalog is to supply morphological data for the purpose of training supervised learning machines for the development of automatic galaxy classification systems. The catalog contains the following information for each galaxy: the Principal Galaxy catalog (PGC) designation, the EFIGI Hubble type, and 16 attributes that describe the galaxy shape, as well as the attributes' lower and upper bounds of confidence intervals. The EFIGI Hubble type and 16 attributes were estimated visually from the composite “*gri*” color images by 10 astronomers. The attributes describe the different components of a galaxy and are as follows: bulge (B/T ratio), spiral arms (the strength, curvature, and rotation of the arms), bars, rings, perturbation, presence of dust, flocculence, hotspots, inclination, arm rotation, and environment contamination and multiplicity.

The data for the catalog was extracted from the Third Reference catalog of Bright Galaxies (RC3; de Vaucouleurs *et al.* 1991, 1995), the Principal Galaxy catalog (PGC; Paturel *et al.* 1989, 1995), the Sloan Digital Sky Survey (SDSS³), the New York

³<http://www.sdss.org>

University Value-Added Galaxy catalog (NYU-VAGC), HyperLeda, and the NASA Extragalactic Database (NED). Various conditions relating to features such as the surface brightness limit, apparent diameter, recessional velocity, and magnitude of the galaxies were applied to the data in order to narrow down the sample. From the RC3, one of the conditions imposed on the data is that the majority of the galaxies selected for the EFIGI catalog possess an apparent diameter larger than 1 arcmin at the $\mu_B = 25$ magnitude/arcseconds² isophotal level. Due to these methods of sampling, the EFIGI catalog contains an over-population of late spiral and irregular galaxies.

After sampling, the 4,458 galaxies that make up the EFIGI catalog are those that have reliable RC3 morphological types and imaging in all 5 bands (*ugriz*) in the SDSS DR4 photometric survey. The photometric and spectroscopic data is from the SDSS DR5 catalog.

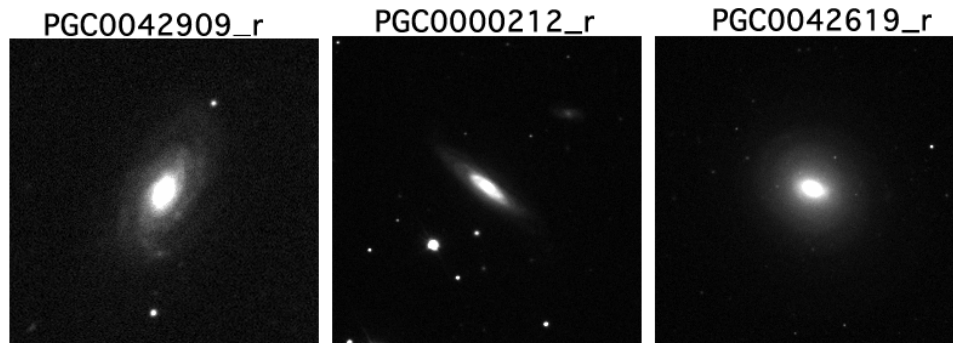


Figure 17: Examples of various EFIGI images used to test and train the software. Each FITS thumbnail is from the *r*-band and includes the galaxies' PGC name.

Nonetheless, the EFIGI catalog contains a large number of diverse morphological types of galaxies over the full 6670 deg² of the SDSS DR4. The FITS images of the galaxies are 255x255 pixels in size and have been visually checked to exclude images affected by artifacts (such as bright stars or satellite trails) or images that

have certain data missing. However, images partially contaminated by artifacts have been kept in order to provide realistic sampling of survey conditions. Most of the galaxies in the catalog (4365 out of 4458) have a redshift $z \leq 0.05$, which is due to selection conditions (Baillard *et al.* 2011). A few examples of the galaxies FITS postage-stamps are included in Figure 17. In this research, we focus on the r -band images.

3.2.2 Analysis

We demonstrate that we are able to emulate the visual classifications of the galaxies in the EFIGI catalog using a set of automated parameters measured for each galaxy by our morphological software and GALFIT. As seen in Table 5, the EFIGI Hubble sequence (EHS) is closely based on RC3. These morphological types were previously given in Table 2. In our study, we categorized the EHS types into broader Hubble classes: -6 through -4 as Ellipticals, -3 through -1 as Lenticulars, 0 through 9 as Spirals, 10 as Irregular type galaxies, and 11 as dwarf galaxies.

Table 5: The EFIGI Hubble Sequence (EHS) as seen in Baillard *et al.* (2011)

| | | | | | | | | | |
|---------------------|-----------|------------|-----------|-----------------------|-----------------------|-----------------------|-------------|-----------|------------|
| Literal Type | <i>cE</i> | <i>E</i> | <i>cD</i> | <i>S0⁻</i> | <i>S0⁰</i> | <i>S0⁺</i> | <i>S0/a</i> | <i>Sa</i> | <i>Sab</i> |
| EHS Type | -6 | -5 | -4 | -3 | -2 | -1 | 0 | 1 | 2 |
| Literal Type | <i>Sb</i> | <i>Sbc</i> | <i>Sc</i> | <i>Scd</i> | <i>Sd</i> | <i>Sdm</i> | <i>Sm</i> | <i>Im</i> | dE |
| EHS Type | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |

We apply our morphological software and GALFIT to the EFIGI catalog FITS images in the SDSS r -band. From the 4458 galaxies in the sample, 106 galaxies did not match the surface brightness threshold⁴ criteria of $\mu_T = 21.8$ magnitude/arcseconds², computed for this data set. Our software was unable to measure the parameters for these objects, thus they are excluded from our analysis. The resulting parameters for

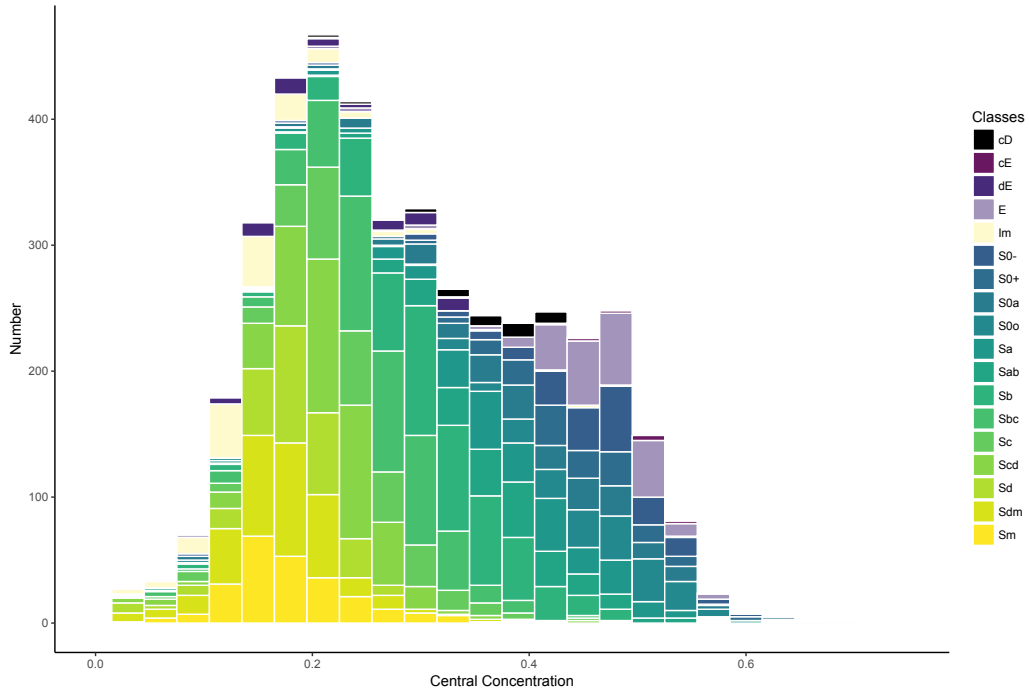
⁴The surface brightness threshold used in our software is defined in Chapter II.

the remaining 4352 galaxies are examined and their relations plotted.

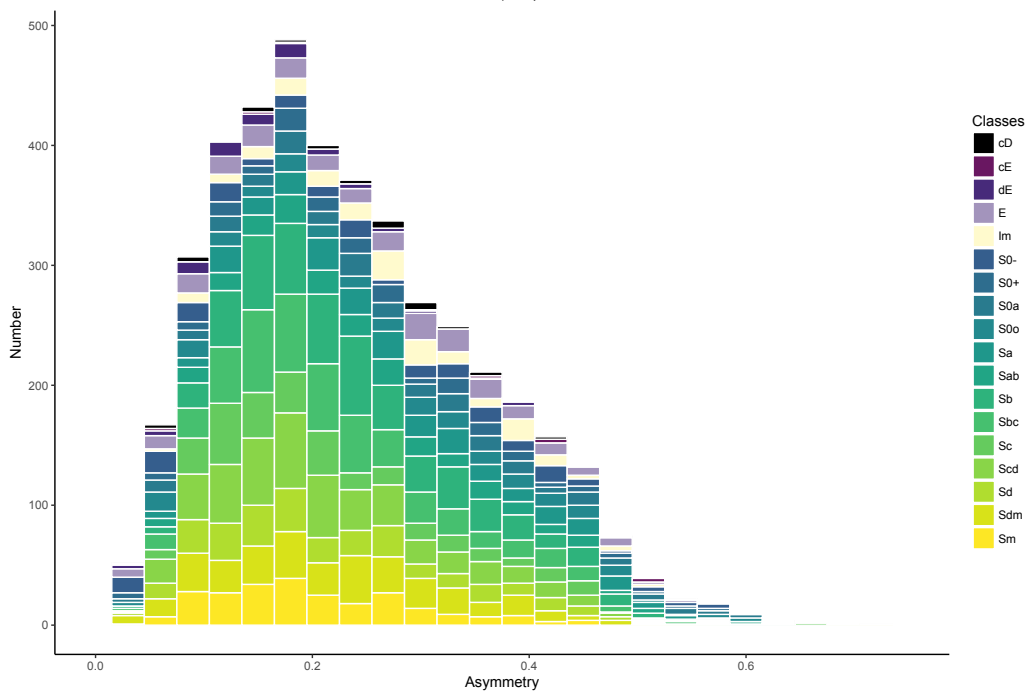
We measured the parameters described in the previous chapter by computing them from each individual EFIGI galaxy FITS images remaining in our study. Figures 18, 19, and 20 display histograms for central concentration, asymmetry, M20, Bulge-to-Total ratios, Gini coefficients, and Theil indexes measured for the EFIGI data. The histogram in Figure 18 (a) displays the detailed Hubble Types defined in the EFIGI catalog as a function of central concentration binned by values of $C = 0.03$. Figure 18 (b) is the histogram of the Hubble Types as a function of asymmetry, also binned by 0.03. Figure 19 (a) is a histogram of Hubble Types as a function of B/T , binned by 0.03 and Figure 19 (b) represents the Hubble Types as function of M20, binned by $M20 = 0.005$. Figure 20 (a) shows the Hubble Types as functions of the Gini coefficient, binned by $Gini = 0.03$. Figure 20 (b) shows the Hubble Types as functions of the Theil index, binned by 0.01.

We compare the relationships of the parameters with one another and defined the regions enclosing the most EFIGI galaxies of a specific class. Figure 21 illustrates the relationship between each quantity as a function of the others. Based on the EHS types, the galaxies depicted in Figure 21 are grouped into five large classes — Dwarfs, Ellipticals, Irregulars, Lenticulars, and Spirals — which were derived from the EHS types. We study the different parameter spaces in order to develop an accurate, automatic system for galaxy classification.

The distribution the 4352 EFIGI galaxies in the A versus Gini plane are shown in Figures 22 and 23. From the figures, it can be seen that early-type galaxies are predominantly located at the right portion of the graph, while the late-type galaxies dominate the left portion. Using the visual classifications of the EFIGI catalog, we divide the plot into two classification regions as shown in Figure 23. We mark the region bounded by $Gini \geq 0.45$ as the region containing mainly the early-type “E/S0”

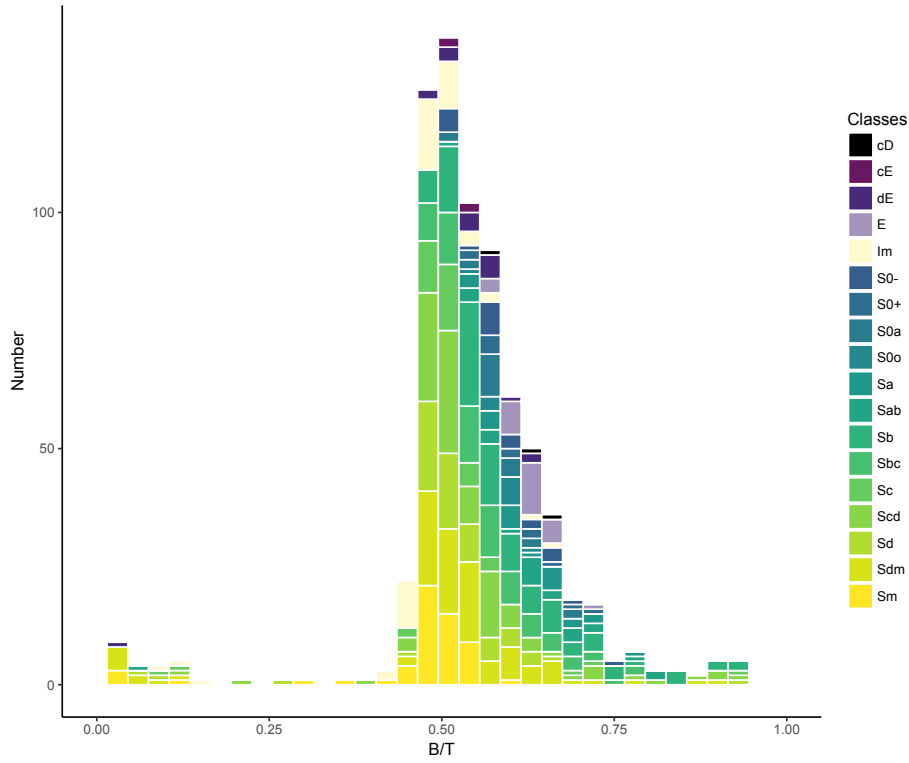


(a.)

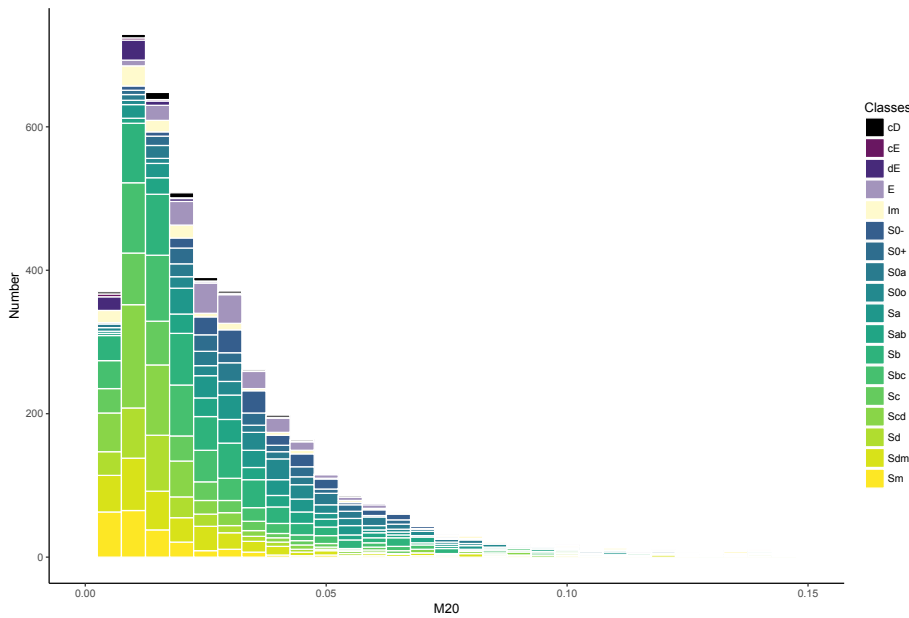


(b.)

Figure 18: Histogram of various Hubble Types, based on the EHS definitions in Table 5, as functions of central concentration (top) and asymmetry (bottom). See text for details.

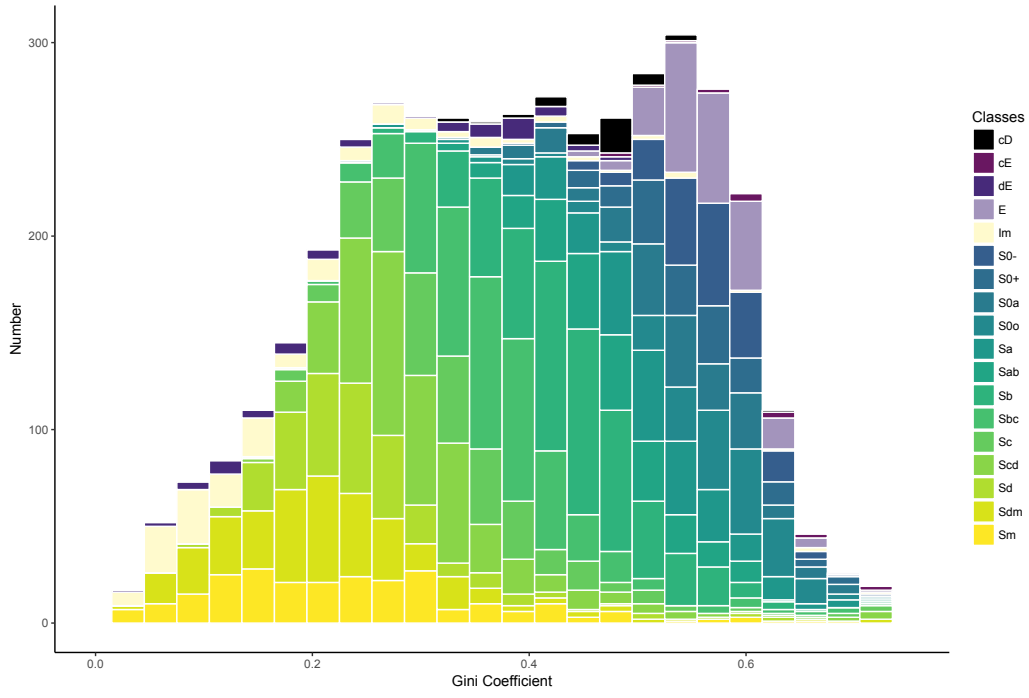


(a.)

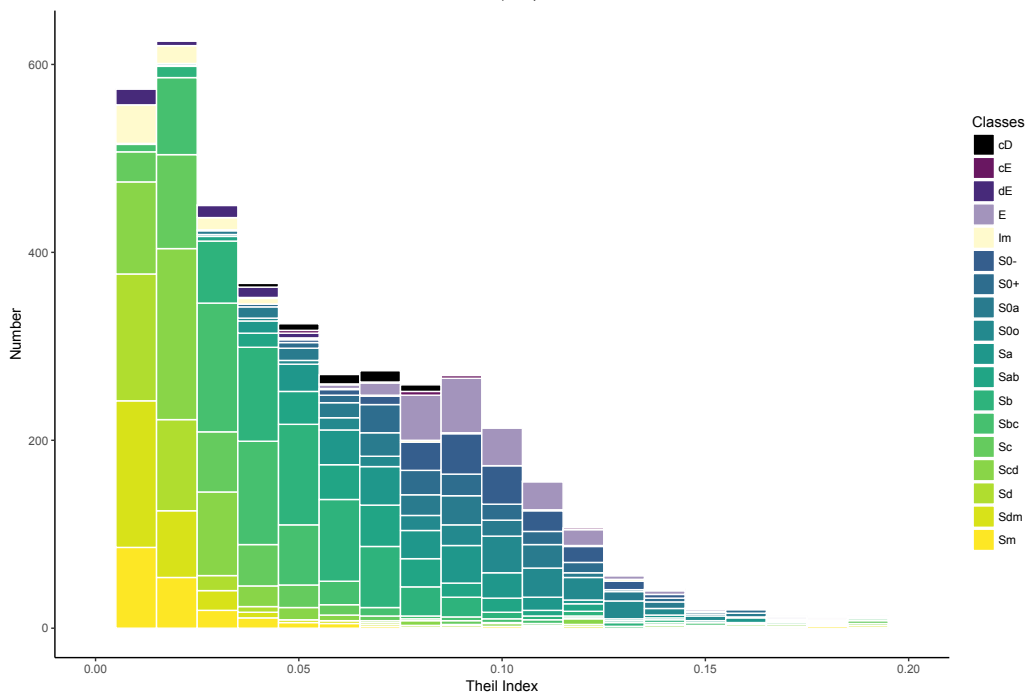


(b.)

Figure 19: Histogram of various Hubble Types, based on the EHS definitions in Table 5, as functions of Bulge-to-Total ratio (top) and M20 (bottom). See text for details.



(a.)



(b.)

Figure 20: Histogram of various Hubble Types, based on the EHS definitions in Table 5, as functions of the Gini coefficient (top) and Theil index (bottom). See text for details.

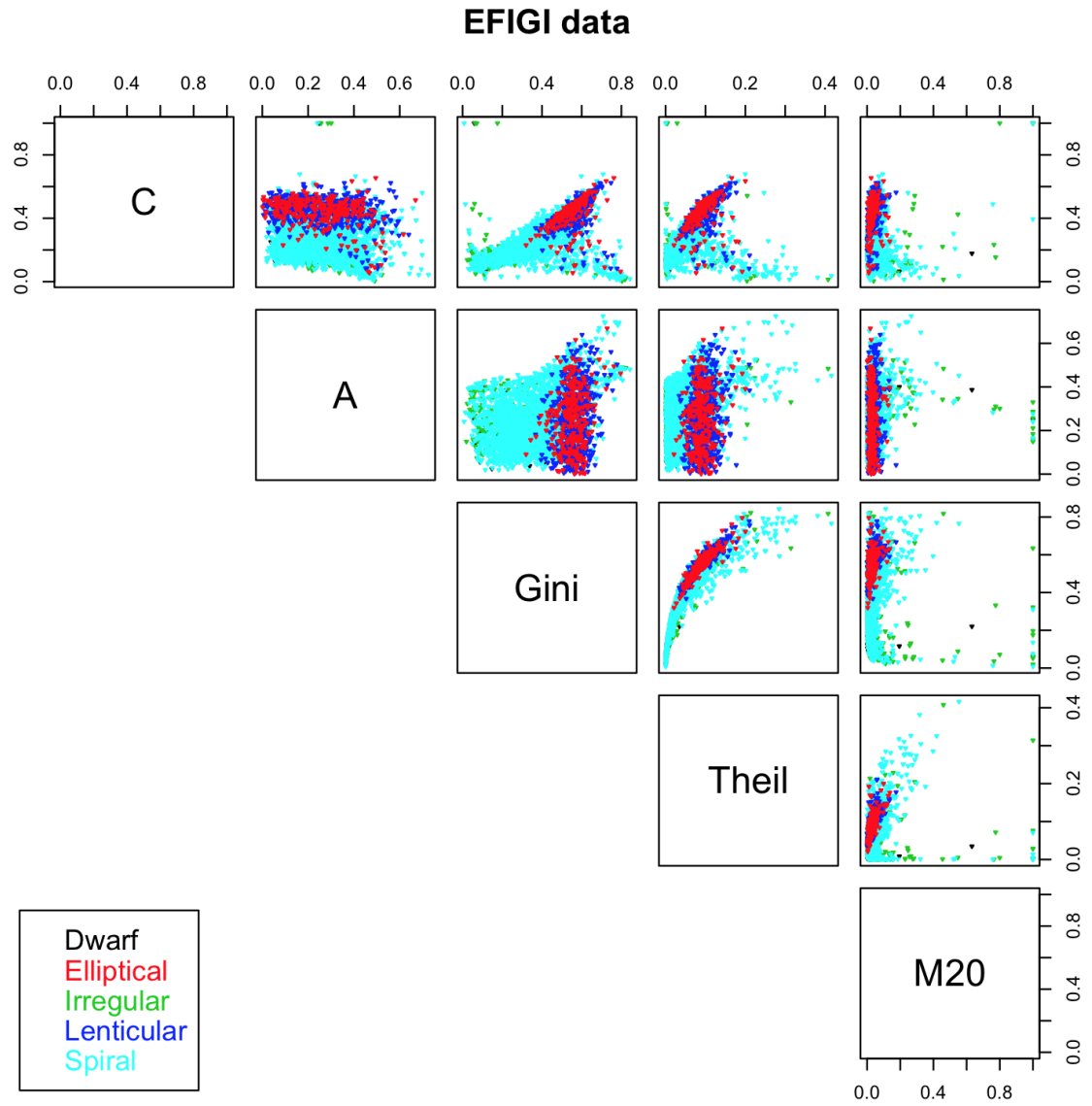


Figure 21: Relations between the five nonparametric values using the 4352 galaxies from the EFIGI data.

EFIGI data: A vs Gini

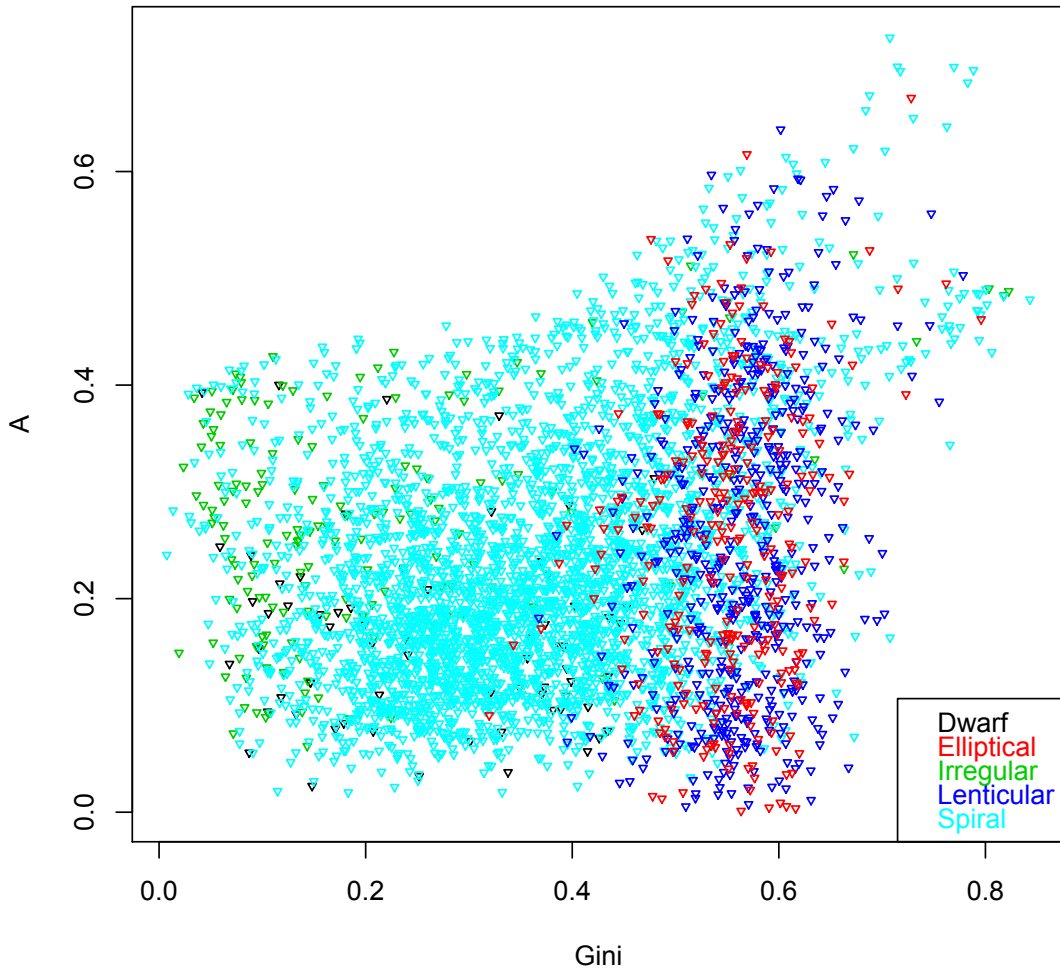


Figure 22: A versus Gini plot of EFIGI data.

galaxies, while the region outside of these bounds is marked as the one containing mainly late-type or “S/Irr” galaxies. We define the “E/S0” region to include galaxies of EHS type 11 and EHS type -6 through 0, while the “S/Irr” region is defined as the region predominantly including galaxies of EHS type 1 through 10.

In our EFIGI sample, 1090 are early-type galaxies (EHS type 11 and EHS type -6 through 0) and 3263 are late-type galaxies (EHS types 1 through 10), out of the

EFIGI data: A vs Gini

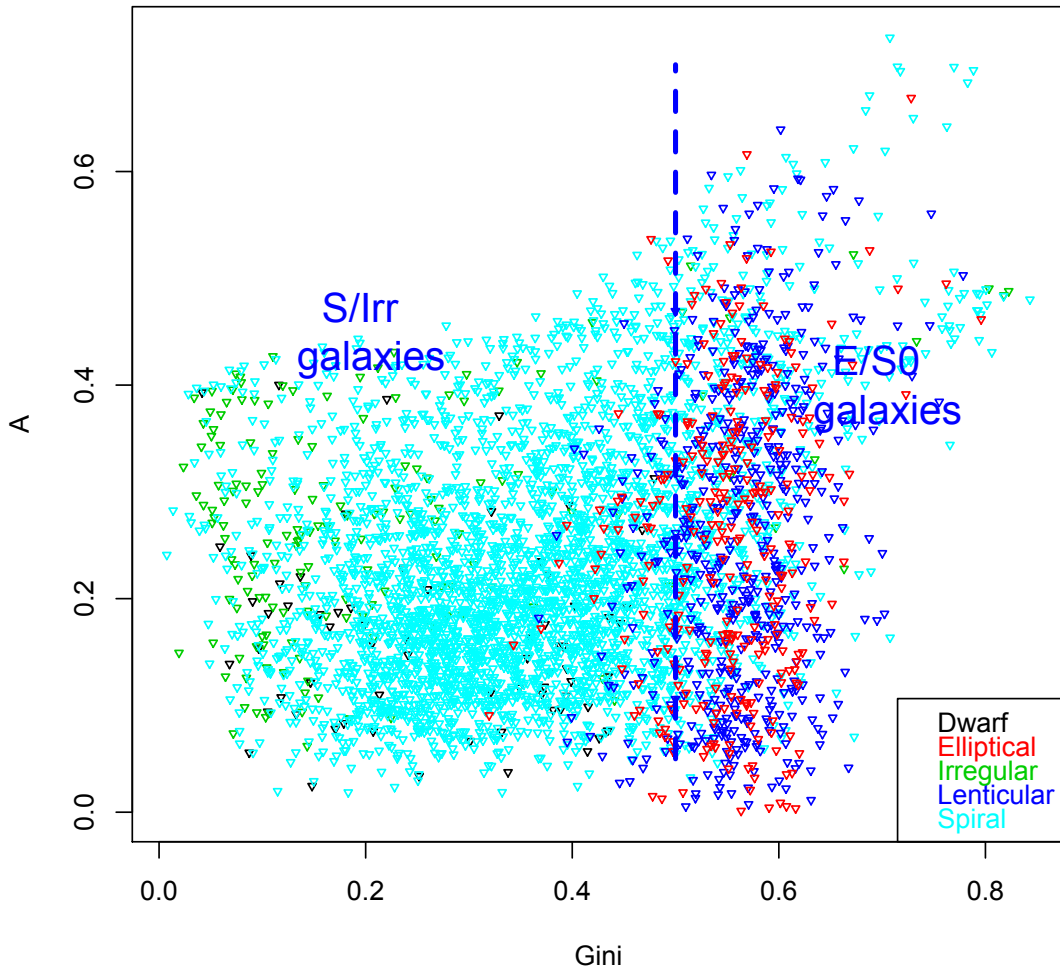


Figure 23: A versus Gini plot of EFIGI data with classification regions.

total 4352 EFIGI catalog galaxies plotted. On the A versus Gini plane, 1709 galaxies are located in the “E/S0” classification region in Figure 23, and approximately 56% of them are early-type galaxies. In other words, the contamination of this region by the late-type galaxies is approximately 44%. On the other hand, out of the 2643 galaxies in the “S/Irr” classification region in Figure 23, about 95% are late-type.

We can also look at these data from a different perspective — out of the 1090

early-type galaxies plotted on the A versus Gini plane, 88% fall into the “E/S0” classification region (see Figure 23). And similarly for the “S/Irr” classification region, approximately 77% of all late-type galaxies (2515 out of 3262) on the plane remain in this region. Thus, we see that from our morphological measurements, a majority of the visually classified early-type EFIGI galaxies occupy the “E/S0” classification region we selected, and similarly, a majority of the visually classified late-type EFIGI galaxies occupy their respective classification region (the “S/Irr” classification area). These results indicate that the classification regions selected on the A versus Gini plane agree with the visual classification of the EFIGI galaxies’ morphology, therefore, by plotting galaxies’ on the A versus Gini plane, it is possible to classify their morphology with a high degree of certainty.

The distribution of the EFIGI galaxies’ A versus Gini values in Figure 23 can be compared to their distribution on the A versus Theil plane in Figure 24. As was mentioned previously, the Theil index is a new statistic we have adopted for morphological measurement. From Figure 24, it can be seen that the spread of the galaxies’ along the Theil axis is more compact when compared to the Gini axis in Figure 23. Otherwise, the early-type and late-type galaxies are located in similar positions on the A versus Theil plane as on the A versus Gini plane.

On the A versus Theil plane, we adopt the region of $\text{Theil} \geq 0.06$ as the region containing mainly the early-type galaxies (the “E/S0” region), while the region outside of these bounds is marked as the one containing mainly late-type (“S/Irr”) galaxies. The A versus Theil plane is similar to A versus Gini, with 1645 galaxies plotted in the “E/S0” classification region and approximately 57% of them being early-type galaxies. Out of the 2707 galaxies in the “S/Irr” classification region on the A versus Theil plane, about 94% are late-type. In this distribution, we also see a majority of galaxies occupying their appropriate classification regions on the plane.

EFIGI data: A vs Theil

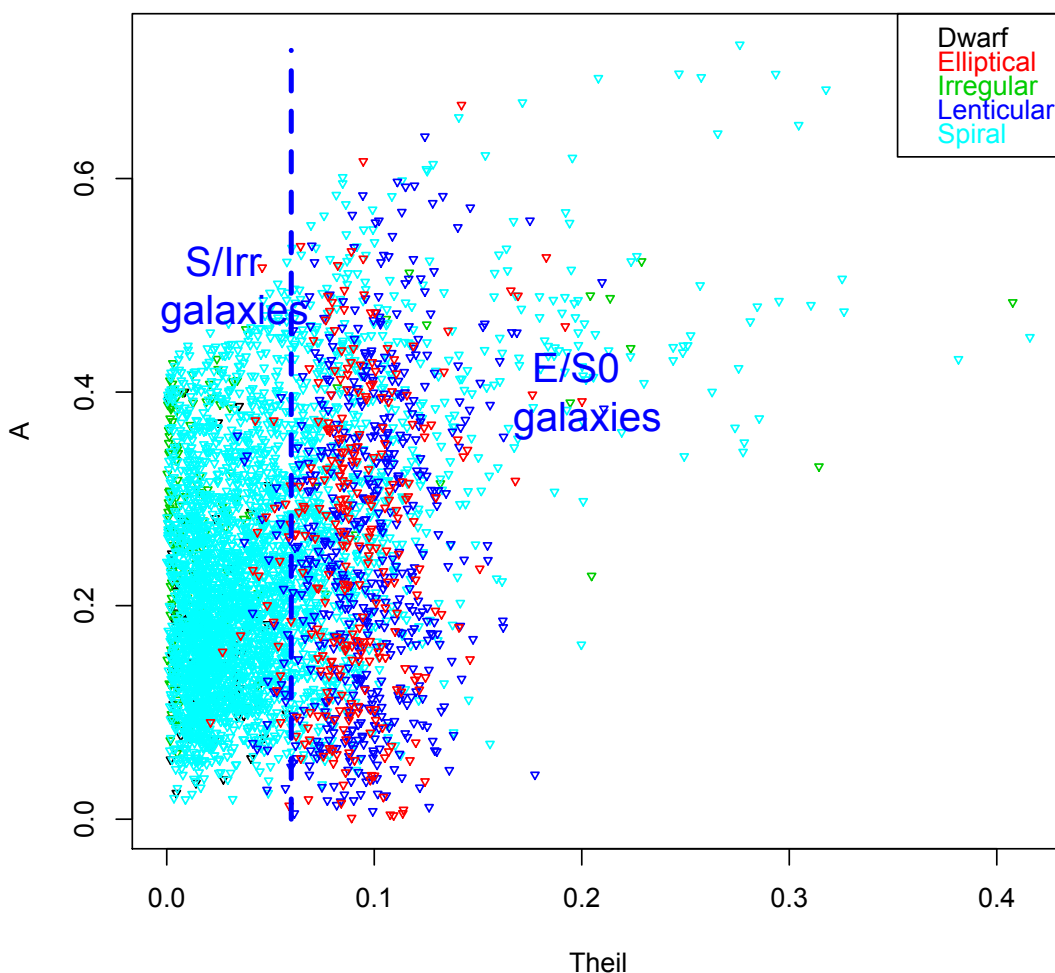


Figure 24: A versus Theil plot of EFIGI data with classification regions.

We also relate the nonparametric quantities measured for each EFIGI galaxy by our software to the galaxies' B/T and B/D ratios measured by GALFIT. Among the published literature, we contrast our results to Cheng *et al.* (2009) and Conselice (2003). Figure 25 is a plot of the relationship between C and B/T published in Cheng *et al.* (2009). We match this plot with our results in Figure 26.

Figure 25 illustrates the distribution of 984 non-star forming galaxies from the

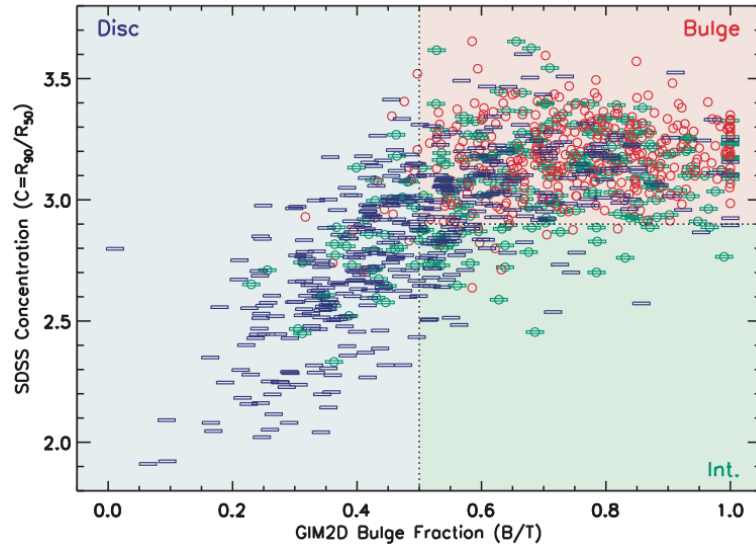


Figure 25: C versus B/T ratio of 984 galaxies from the SDSS red sequence, as published in Cheng *et al.* (2009).

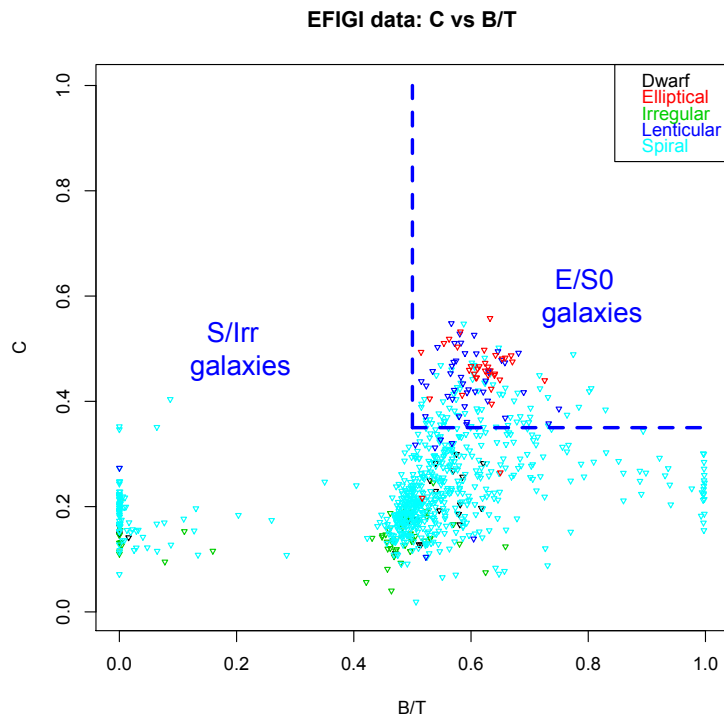


Figure 26: C versus B/T ratio for 815 galaxies from the EFIGI data.

SDSS red sequence. Based on their light profiles, these galaxies were classified by eye into three main groups: bulges, smooth disks, and unsmooth disks. The C parameter was measured for each galaxy using the following definition:

$$C = \frac{R_{90}}{R_{50}}, \quad (3.1)$$

where R_{90} and R_{50} are the radii containing 90% and 50% of the Petrosian flux in the r -band. The Petrosian flux (F_p) is defined as the sum of all the flux within a Petrosian radius R :

$$F_p = 2\pi \int_0^{kR} I(R') R' dR', \quad (3.2)$$

where $I(R)$ is the surface brightness profile, and SDSS has selected $k = 2$ to define the aperture (e.g. Shimasaku *et al.* 2001; Blanton *et al.* 2001). The Petrosian (1976) radius is defined as a ratio of the surface brightness at a radius R to the average surface brightness interior to R that is equal to some fixed value $\eta(R)$:

$$\eta(R) = \frac{I(R)}{\langle I(< R) \rangle}, \quad (3.3)$$

where $\eta(R)$ is typically set to 0.2 (e.g. Shimasaku *et al.* 2001; Yasuda *et al.* 2001; Spinrad 2005; Lotz *et al.* 2008). It is a distance-independent way to describe the radial light profile of a galaxy.

The authors used the IRAF galaxy modeling package GALAXY IMAGE 2D (GIM2D; Simard *et al.* 2002) to measure several quantitative parameters. Similar

to GALFIT, GIM2D models a galaxy image with the use of light profiles — the de Vaucouleurs profile for bulges and an exponential profile for disks. In order to find B/T , GIM2D fits the light of a galaxy as the sum of these two profiles. The definition of B/T used in Cheng *et al.* (2009) is the same as described in Equation 2.16.

Figure 25 depicts selected SDSS galaxies in the C (defined in Equation 3.1) versus B/T plane. The red circles represent systems with visually obvious bulges, the blue bars represent disk systems, and the green circles+bars are intermediate systems. The shaded regions represent the automated classification boundaries defined by the authors. From the figure, it can be seen that the authors define the region bound by $B/T > 0.5$ and $C > 2.9$ as the area where bulge systems are predominantly found (pink-shaded region). Most of the galaxies in the $B/T > 0.5$ and $C \leq 2.9$ region (shaded green) are visual intermediates, and $B/T < 0.5$ (shaded blue) contains mainly disk systems.

Comparing Figure 25 to our results in Figure 26, we see interesting similarities. We classify our sample of EFIGI galaxies into the same five large bins mentioned earlier and use GALFIT to measure their B/T values. From the square root of the variance of the data, we impose a cut of $\chi^2 \leq 10$ and narrow the sample down to 815 galaxies. Our results match quite closely to Cheng *et al.* (2009): we define the region bounded by $B/T > 0.5$ and $C \geq 0.36$ to belong to mainly early-type galaxies, and the region outside of these boundaries contain mainly late-type systems. Due to the different definitions of C , the boundaries of our classification regions vary slightly.

Furthermore, we study the distribution of the EFIGI data across the C versus Gini, C versus Theil, and Gini versus Theil planes. As seen in Figure 27, C and Gini demonstrate a nearly linear relationship. The data is distributed along the line with approximately a unity slope. There is weak curvature near the low- C and low-Gini values at the bottom left end of the distribution, as well as on the high- C and high-

Gini end at the top right. As was already mentioned, the EFIGI data we tested in this thesis is from the r -band, but we can compare Figure 27 to Figure 28, which is from Abraham *et al.* (2003).

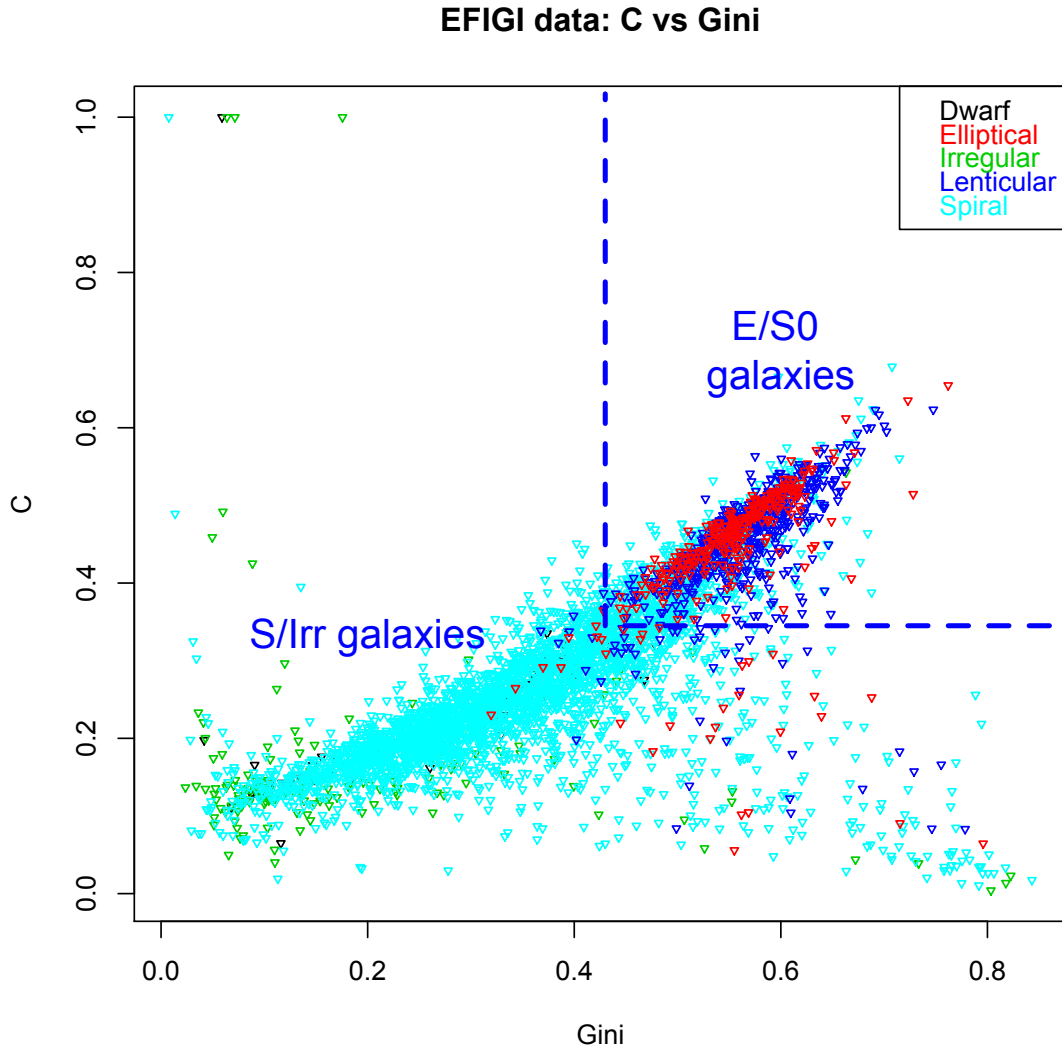


Figure 27: C versus Gini plot of EFIGI data.

In Abraham *et al.* (2003), the authors explored the relationship between Gini and C for approximately 930 galaxies from the SDSS Early Data Release (EDR) taken in the i -band and g -band. The images used in Abraham *et al.* (2003) are

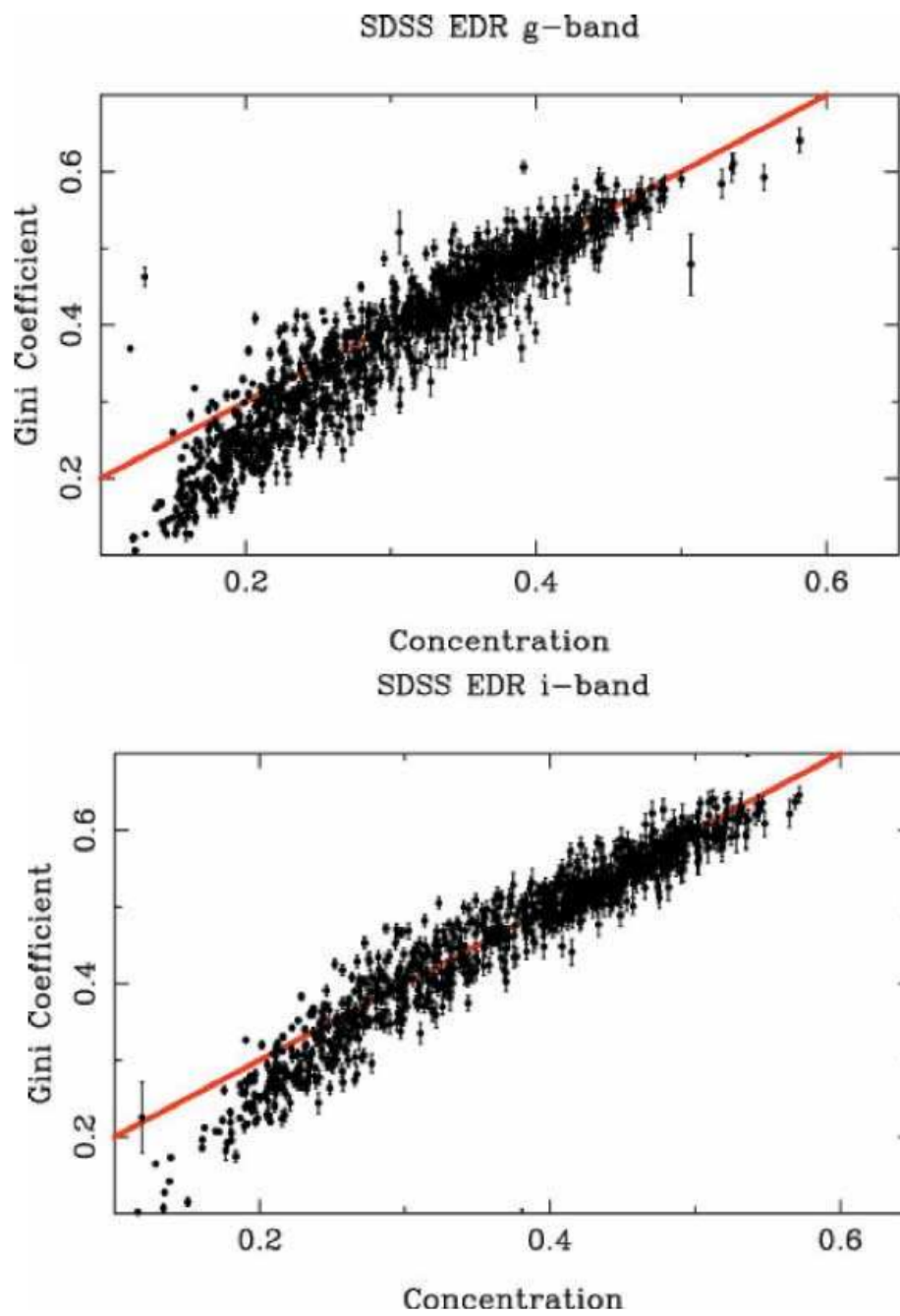


Figure 28: Gini versus C plot of SDSS EDR data using the g and i bands.

small “postage-stamp” images of each galaxy, much like the images in the EFIGI sample used in our study. In Figure 28, the galaxies in the diagram span a broad range of morphologies, but pure disk systems occupy the low- C values and centrally concentration elliptical galaxies occupy the high- C values. The results seen in Figure 27 closely match Abraham *et al.* (2003).

From Figures 27 and 28, it can be seen that due to their linear nature, the Gini coefficient can be used as a substitute for C . The benefit of using Gini rather than the C coefficient is that it relies on fewer assumptions about the shape of the galaxy analyzed. The Gini coefficient can be applicable to galaxies of arbitrary shape or to ones that do not have a single, well-defined center. Due to this, the Gini coefficient may be an important alternative to classifying high-redshift galaxies, which tend to have ambiguous morphologies.

As seen in Figure 28, the results of the distribution of data using two wavebands is relatively the same, except there is a greater scatter of data in the g -band compared to the i -band. It has been found that systematic biases in sample selection or redshift distribution are not the main cause of the scatter in the relationship between C and Gini. Abraham *et al.* (2003) suggest an explanation for the scatter on the C versus Gini plane may be related to the mean surface brightness.

In our study, we mark the region bounded by $C \geq 0.345$ and $\text{Gini} \geq 0.43$ as the region containing mainly early-type “E/S0” galaxies, while the region outside of these bounds contains mainly late-type or “S/Irr” galaxies (see Figure 27). On the C versus Gini plane, 888 out of 1090 (approximately 81%) early-type galaxies in our EFIGI sample are found within the “E/S0” classification region, and 2769 out of 3262 (approximately 85%) late-type galaxies lie in the “S/Irr” classification region. In other words, out of the 1380 late and early type galaxies found in the “E/S0” region, 64% are early-type, and in the “S/Irr” region, out of the 2971 galaxies plotted, 93% are

late-type galaxies.

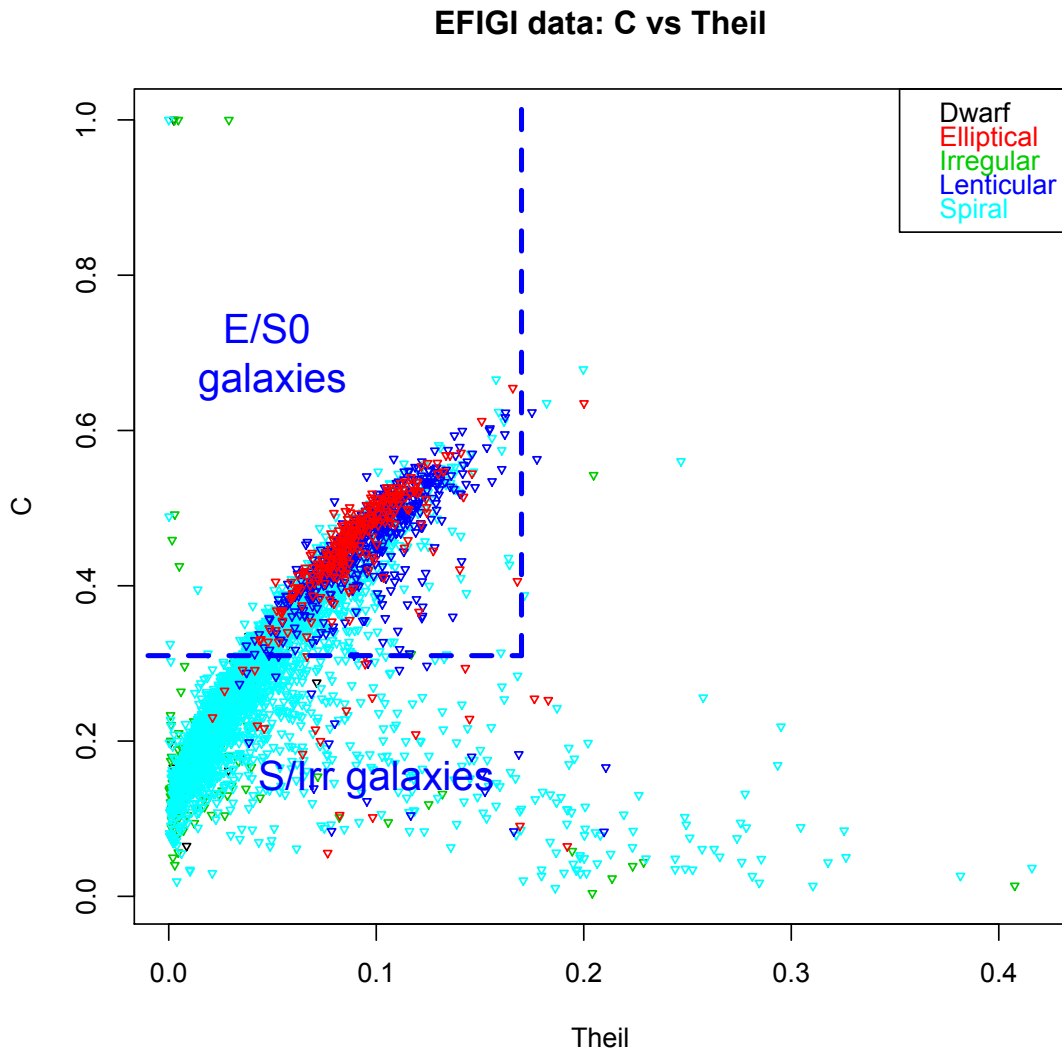


Figure 29: C versus Theil plot of EFIGI data.

Comparing the C versus Gini plane to the C versus Theil plane in Figure 29, we see a denser distribution in the later. The region bounded by $C \geq 0.31$ and $\text{Theil} \leq 0.17$ is defined as the region containing mainly the early-type “E/S0” galaxies. This region contains 951 early-type galaxies out of the total 1090 early-type galaxies on the plot, which means approximately 87% of the early-type galaxies in the sample

are positioned in this region. The “S/Irr” classification region contains approximately 74% late-type galaxies (2419 out of 3262 galaxies). Of the 1788 early and late type galaxies in the “E/S0” region, 53% are early-type, while out of the 2558 galaxies in the “S/Irr” region, 95% are late-type. From these results, we can see that the late-type galaxies greatly contaminate the “E/S0” classification region on this plane than in the C versus Gini plane, which may be due to the high-density of the distribution.

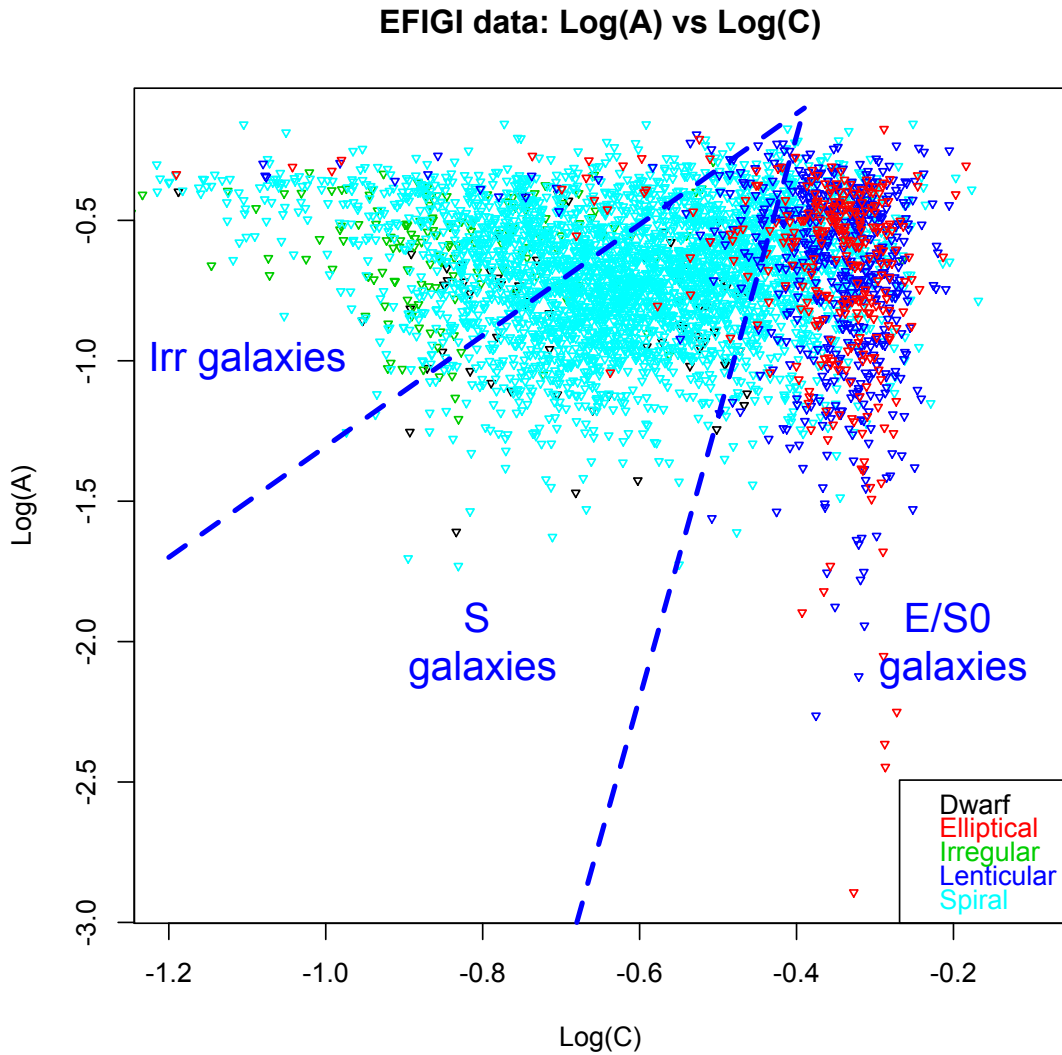


Figure 30: $\text{Log}(A)$ versus $\text{Log}(C)$ plot of EFIGI data.

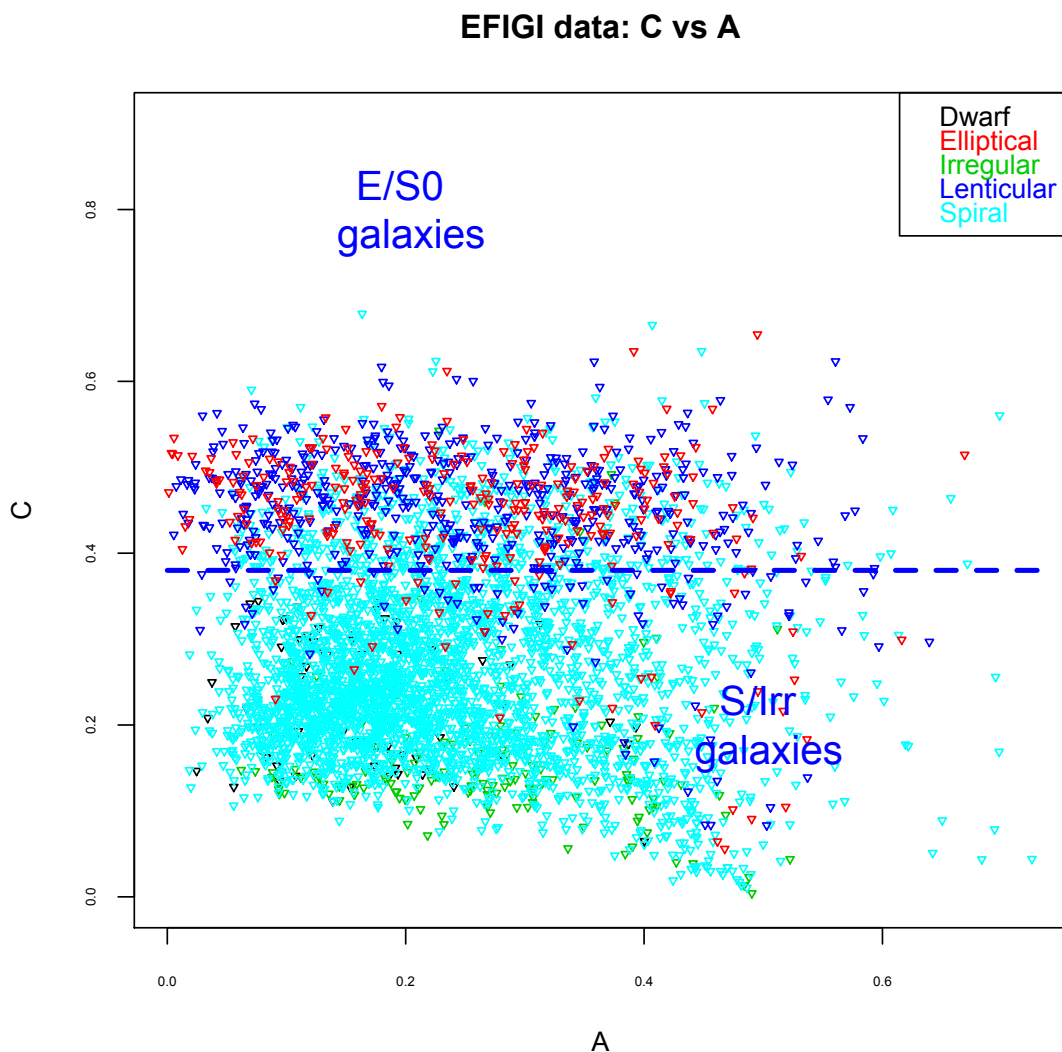


Figure 31: C versus A plot of EFIGI data.

Another test of our morphology software involves comparing the galaxies' C parameter to their A values. Figure 30 is a plot of $\text{Log}(A)$ versus $\text{Log}(C)$ of the 4352 EFIGI galaxies in our analysis. We compare this to Figure 13, from Abraham *et al.* (1996a), which displays the authors' visual classification of approximately 300 galaxies from the Hubble Deep Field (HDF). As was discussed in the previous chapter, Abraham *et al.* (1996a) divided the plot into three sectors according to the visual

classification of the galaxies. Similarly to Abraham *et al.* (1996a), we introduce three sectors to the distribution in Figure 30. It can be seen that the galaxies in the EFIGI sample disperse in a similar manner on the $\text{Log}(A)$ versus $\text{Log}(C)$ parameter space as the ones from the HDF sample.

For our analysis, we focus on the C versus A distribution, as depicted in Figure 31. We define the region $C \geq 0.38$ as the “E/S0” classification region belonging to mainly early-type galaxies. Approximately 69% (829 out of 1210) of the early-type galaxies in the sample are positioned in this region. The “S/Irr” classification region contains approximately 92% late-type galaxies (2874 out of 3135 galaxies). Out of the 1090 early-type galaxies on this plane, 76% are early-type in the “E/S0”, while out of the 3262 total late-type galaxies, 88% are in the “S/Irr” region.

After comparing all relations between each parameter, we investigated the results of our morphological software by consecutively applying a number of parameter plane cuts to the data. In this manner, the parameter planes act like “filters” on the data. Compared to how each individual plane classifies galaxies, applying numerous planes in a certain order can increase the precision of classification. For these tests, we began by arranging the EFIGI galaxies into two broad bins — E and S. The E bin includes EHS type 11 and EHS type -6 through 0, while the “S” bin includes galaxies of EHS type 1 through 10.

One combination of planes studied was between the Gini coefficient, Theil index, C , and A . We applied three parameter-relation planes in the following order: C versus Gini $\rightarrow C$ versus $A \rightarrow$ Gini versus Theil. Since each plane acts like a “filter”, galaxies with appropriate criteria are able to proceed to the next plane and those that do not are labeled as such. The plot of the final distribution in this combination of relations can be seen at the top of Figure 32.

In the distribution of galaxies on the Gini versus Theil plane alone, 51% of galaxies

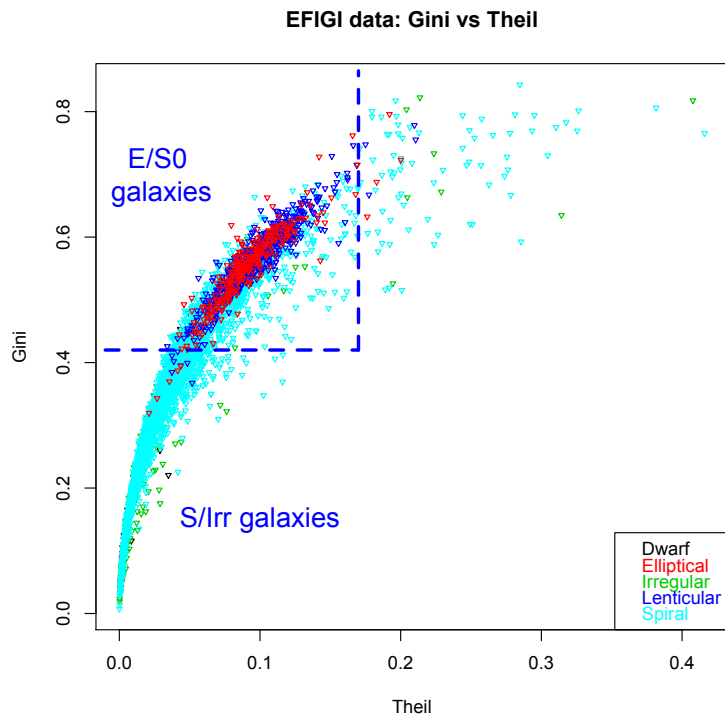
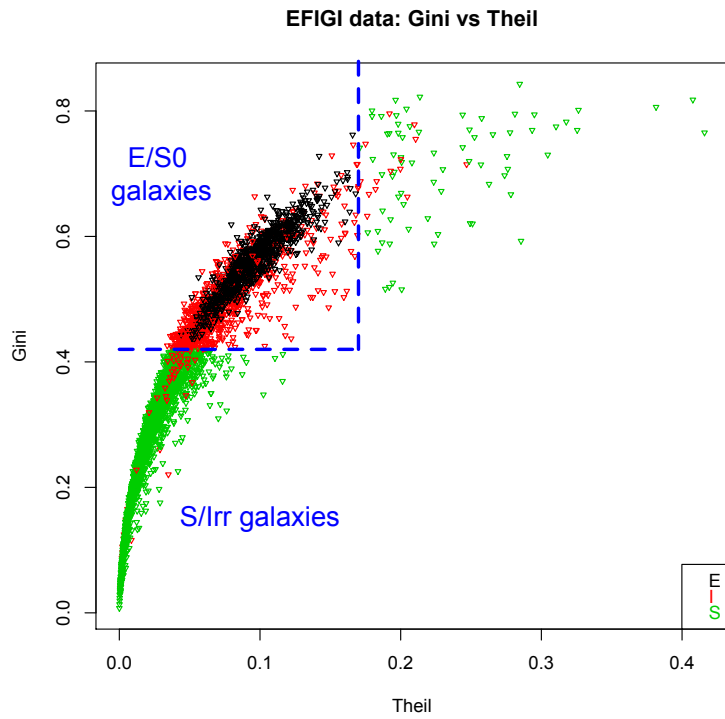


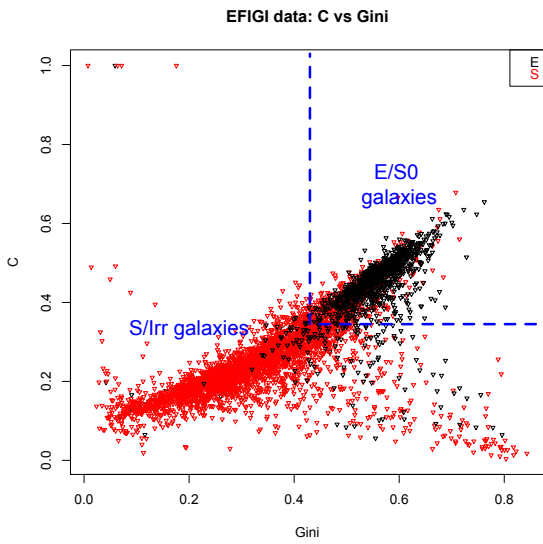
Figure 32: Distribution of EFIGI data in the Gini vs. Theil plane after filtering (top) and without (bottom).

in the designated “E/S0” region are early-type, and 96% of galaxies in the “S/Irr” region are late-type. Here we define the region bound by $Gini \geq 0.42$ and $Theil \leq 0.17$ as “E/S0” as it is primarily occupied by early-type galaxies. The region outside this boundary is classified as “S/Irr” and contains mainly late-type galaxies.

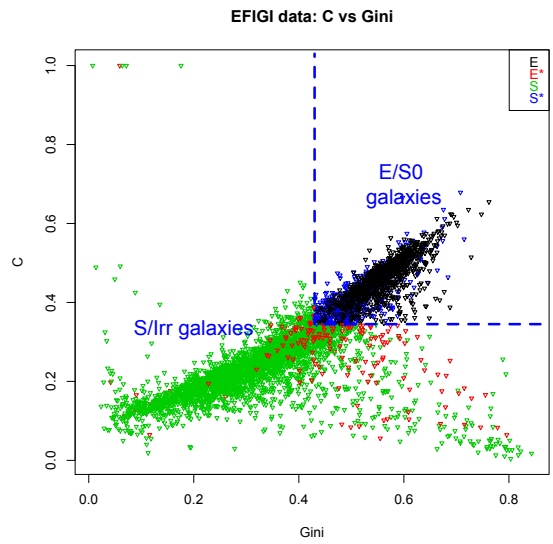
After looking at the C versus Gini $\rightarrow C$ versus $A \rightarrow$ Gini versus Theil planes, we define an intermediate class of galaxies (*i.e* the “T” class), which are galaxies that are classified into the wrong classification region on a parameter plane, meaning, are classified as early-type on one plane and late-type on another. The final plane can be seen at the top of Figure 32.

Applying three or more parameter planes to the data results in a larger number of galaxies being classified as “T”. In the case of Figure 32 (top), 44% are early-type in the “E/S0” region, and 95% of the galaxies are late-type in the “S/Irr” zone. The “T” class is 27%.

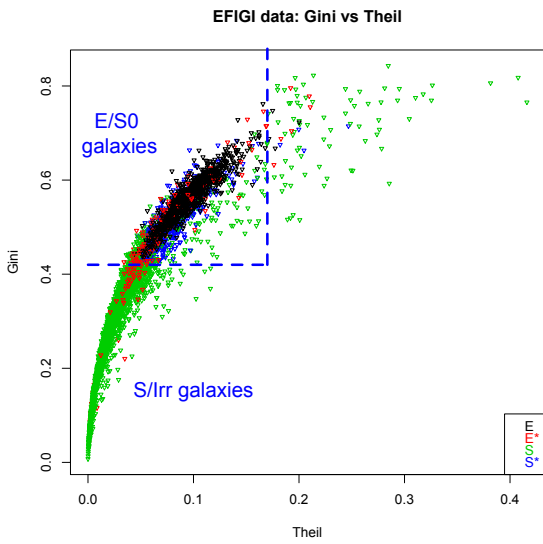
Therefore, we test two combinations of planes on the EFIGI data instead of three or more. Analyzing data through the C versus Gini \rightarrow Gini versus Theil planes, all 4352 EFIGI galaxies in our study are classified into three classes: 884 galaxies are classified as early-type (E), 2371 were late-type (S), and 1096 were classified as “T” (which is approximately 25% of the sample). Figure 33 demonstrates the stages of the analysis. As was mentioned, Figure 33(a.) is the distribution of the EFIGI galaxies on the C versus Gini plane, where the galaxies are sorted into two broad bins — E and S. The E bin includes EHS type 11 and EHS type -6 through 0, and the S bin includes galaxies of EHS type 1 through 10. Early-type galaxies located in the “S/Irr” classification region are then labeled “E*”, and like-wise, late-type galaxies in the “E/S0” region are labeled “S*”, as can be seen in Figure 33(b.). Using these classifications, we plot the distribution of the EFIGI galaxies on the Gini versus Theil plane in Figure 33(c.). Finally, galaxies classified as “E*” or “S*” are grouped together



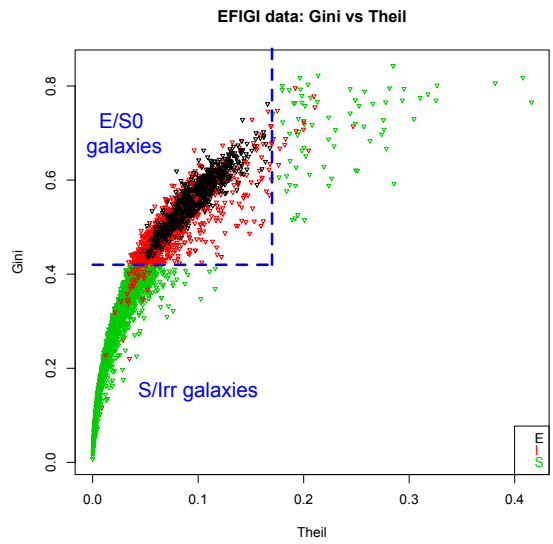
(a.)



(b.)



(c.)



(d.)

Figure 33: Applying two planes — the C versus Gini \rightarrow Gini versus Theil — to the EFIGI data.

into the “T” class seen in Figure 33(d).

We found that A versus Gini $\rightarrow A$ versus Theil produces the best classification results, meaning, most early-type and late-type galaxies are found within their appropriate classification regions and there is a small percentage of galaxies classified as “T” (approximately 22%, in this case). We will focus on this method of classification for the rest of the data sets examined in this thesis.

CHAPTER IV

HIGH-REDSHIFT DATA

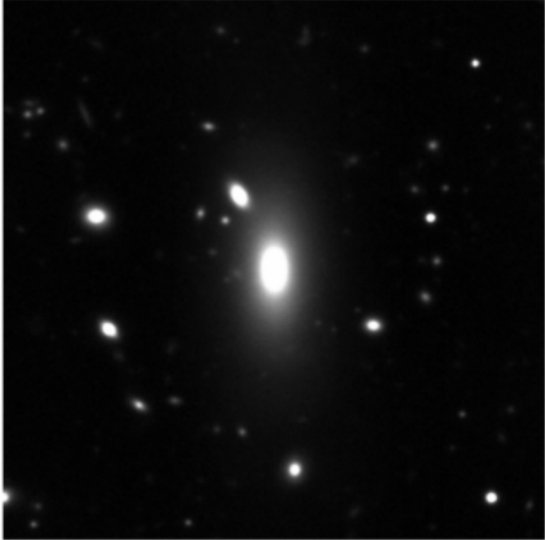
4.1 High-redshift CFHT Clusters

We applied our morphology software on a large number of diverse data sets. Among the data examined were 15 Abell high-redshift galaxy clusters from Rude (2015; *et al.* 2018 in preparation), whose central coordinates for the brightest cluster galaxy (BCG), redshift, r -band exposure time, and the radial coverage, are presented in Table 6. The clusters were observed by the Canada-France-Hawaii Telescope (CFHT) and reduced by Rude (2015). They range in redshift from 0.03 to 0.18.

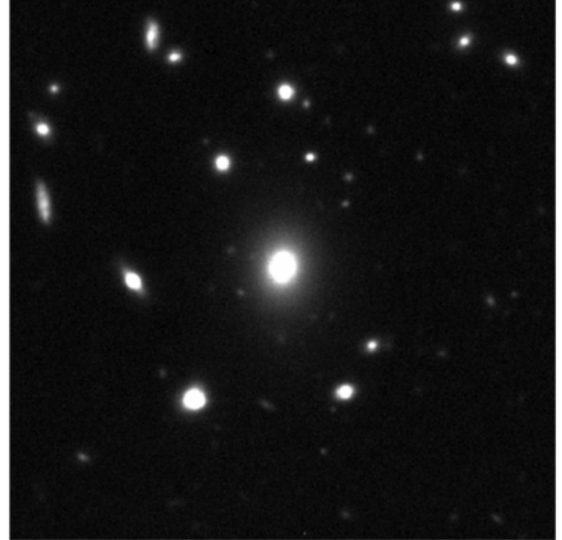
The CFHT is a 3.6 meter telescope located on Mauna Kea, Hawaii. The images were taken with the Megacam CCD mosaic camera, which consists of 36 CCD chips that are 2048 x 4612 pixels in size, and cover a full 1x1 square degree field-of-view with a resolution of 0.185 arcsecond/pixel.

Figure 34 displays several r -band images of galaxies in the CFHT data set. Each postage stamp is 500 x 500 pixels in size. The galaxy analyzed is positioned in the center of each postage stamp, with its right ascension and declination stated above each image. The galaxies are selected from various clusters in the data set. In the next section, we describe the method of preparing the CFHT files for our morphology software and analyze the galaxies' classification results.

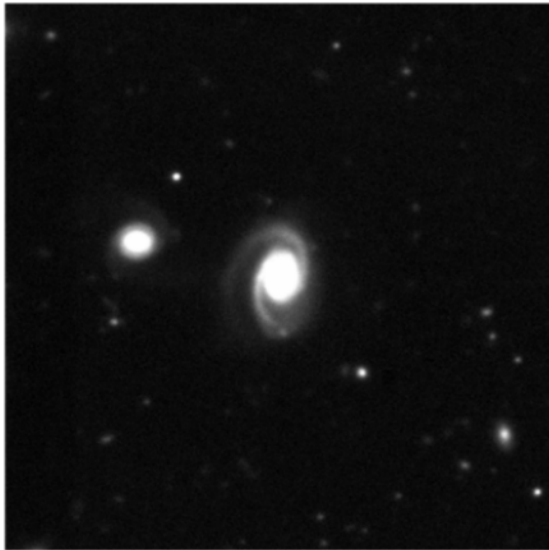
RA: 217.0087
Dec: 55.82212



RA: 0.057172
Dec: 15.8402



RA: 234.1897
Dec: 37.61852



RA: 234.1367
Dec: 37.60778

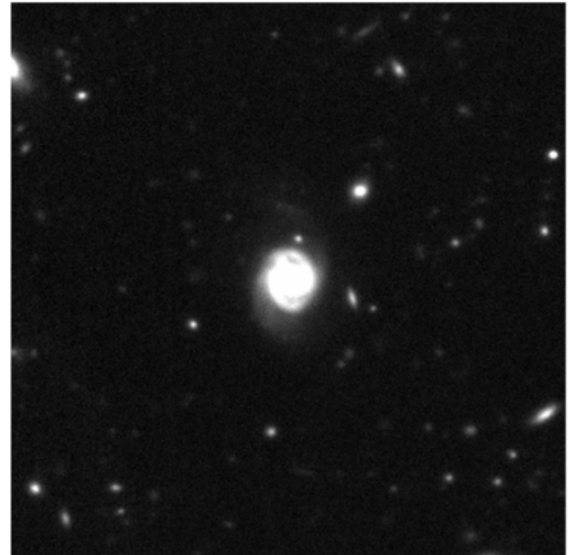


Figure 34: Sample of several r -band postage stamps of galaxies from the CFHT data set. See text for details.

4.1.1 Analysis

In this thesis, we analyze FITS images of galaxies which are already reduced and processed. For each cluster we first assemble a DAT file, as described in Section 2.3.

In order to increase the speed of measuring the parameters for each galaxy studied, we cut out the individual galaxies from their cluster FITS image. We create postage-stamp FITS images of each galaxy using Perl scripts and the *imcopy* command in IRAF. Afterwards, we apply our morphological software to the data and measure the classification parameters for each galaxy. Lastly, we apply GALFIT to the postage-stamp images and measure the B/D and B/T ratios.

In this section, we analyze the data using the methods described in Section 3.2.2. In order to compare the results with the EFIGI data studied in the previous chapter,

Table 6: Fifteen CFHT clusters from Rude (2015; *et al.* 2018 in preparation) studied in this thesis.

| Cluster | RA (degrees) | Dec (degrees) | z | r -band exposure (s) | Radial Coverage (Mpc) |
|---------|-----------------|------------------|-------|------------------------------|-----------------------------|
| A76 | 9.98315 | 6.8486 | 0.041 | 240 | 1.0 |
| A98N | 11.6031 | 20.6218 | 0.104 | 2160 | 3.2 |
| A98S | 11.6221 | 20.4680 | 0.104 | 2160 | 3.2 |
| A350 | 36.2721 | -9.8366 | 0.157 | 2000 | 1.6 |
| A351 | 36.3331 | -4.8827 | 0.111 | 2000 | 1.6 |
| A362 | 37.9215 | -4.8827 | 0.184 | 2500 | 0.7 |
| A655 | 126.3712 | 47.1337 | 0.127 | 2940 | 2.0 |
| A795 | 141.0222 | 14.1727 | 0.136 | 2880 | 1.0 |
| A1920 | 216.8524 | 55.7502 | 0.131 | 4000 | 1.3 |
| A1940 | 218.8686 | 55.1312 | 0.140 | 2000 | 1.3 |
| A2100 | 234.0773 | 37.6438 | 0.153 | 1600 | 0.6 |
| A2107 | 234.9127 | 21.7827 | 0.041 | 600 | 1.0 |
| A2147 | 240.5709 | 15.9747 | 0.035 | 600 | 0.8 |
| A2199 | 247.1594 | 39.5513 | 0.030 | 1600 | 0.7 |
| A2688 | 0.0318 | 15.8342 | 0.151 | 2160 | 0.6 |

421 visually classified CFHT galaxies

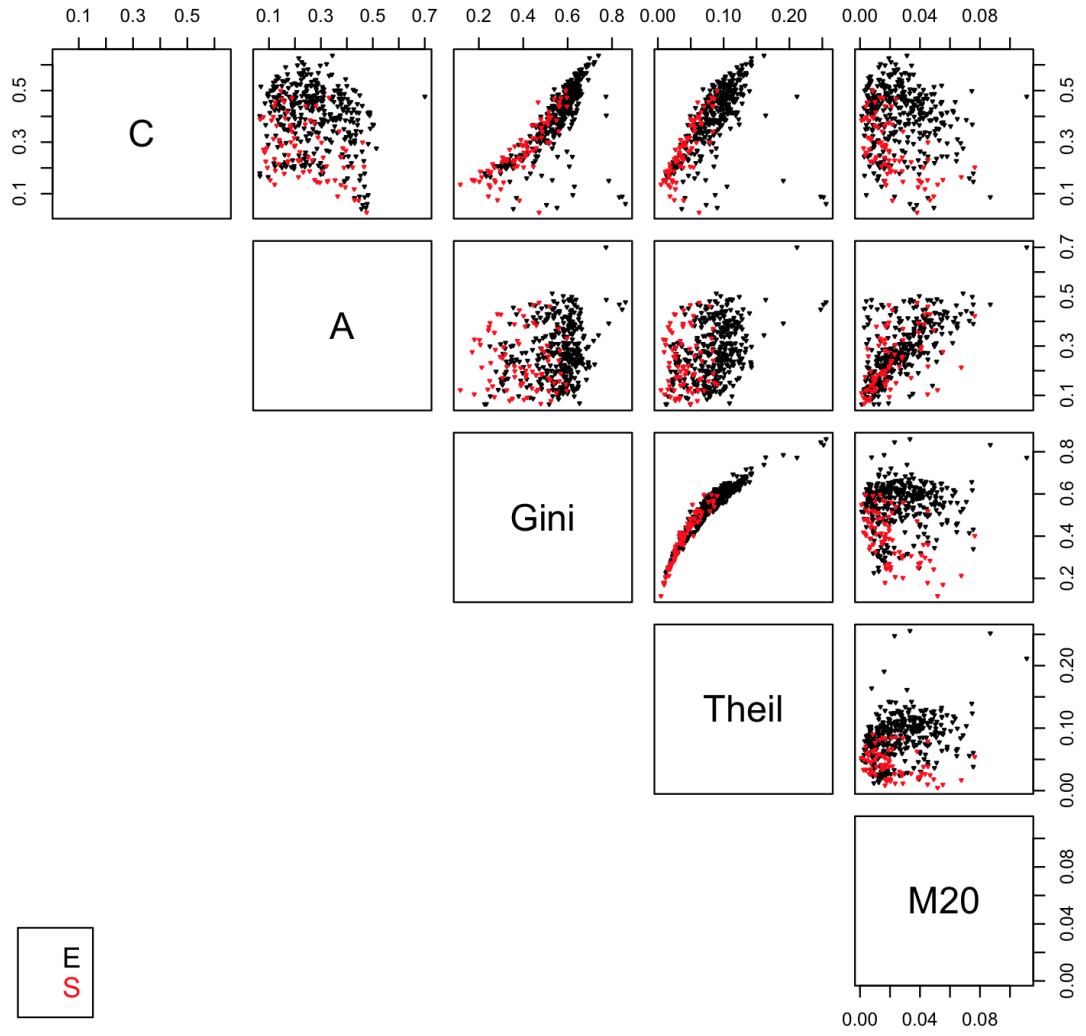


Figure 35: Relations between five parameters for 421 galaxies from 15 CFHT galaxy clusters.

we visually classify 421 bright galaxies from the 15 CFHT galaxy clusters. This sample contains 342 early-type galaxies and 79 late-type systems. As was shown in Figure 21, Figure 35 illustrates the relationship between each nonparametric quantity as a function of the others. In this sample, the 421 galaxies are grouped into two broad classes — “E” for early-type and “S” for late-type galaxies. We apply the regions defined in Chapter III to the CFHT data.

Figures 36 and 37 display histograms for central concentration, asymmetry, Gini coefficients, and Theil indexes measured for the visually classified CFHT data. Early-type galaxies (E) are represented in red and late-types (S) are in blue. The histogram in Figure 36 (a) displays the galaxies in these two Hubble classes as a function of central concentration binned by values of $C = 0.03$. Figure 36 (b) is the histogram of the Hubble classes as a function of asymmetry, also binned by 0.03. Figure 37 (a) shows the Hubble Types as functions of the Gini coefficient, binned by $\text{Gini} = 0.03$. Figure 37 (b) shows the Hubble Types as functions of the Theil index, binned by 0.01.

The distribution of the 421 visually classified CFHT galaxies in the A versus Gini plane is shown in Figure 38, with the regions defined in the previous chapter where $\text{Gini} \geq 0.45$ as the region containing mainly the early-type galaxies, while the region outside of these bounds contains mainly late-type galaxies. In this plane, 290 galaxies are located in the “E/S0” classification region, and approximately 91% of them are visually classified as early-type galaxies. There are 102 galaxies in the “S/Irr” classification region in Figure 38 of which 50% are late-type. In other words, 290 out of 342 (*i.e.* 85%) early-type galaxies are located on the “E/S0” region in the A versus Gini plane, and 51 out of 79 (*i.e.* 65%) late-type galaxies are plotted in the “S/Irr” region. Therefore, just as we see with the EFIGI data, the majority of visually classified CFHT galaxies are also plotted in corresponding classification regions.

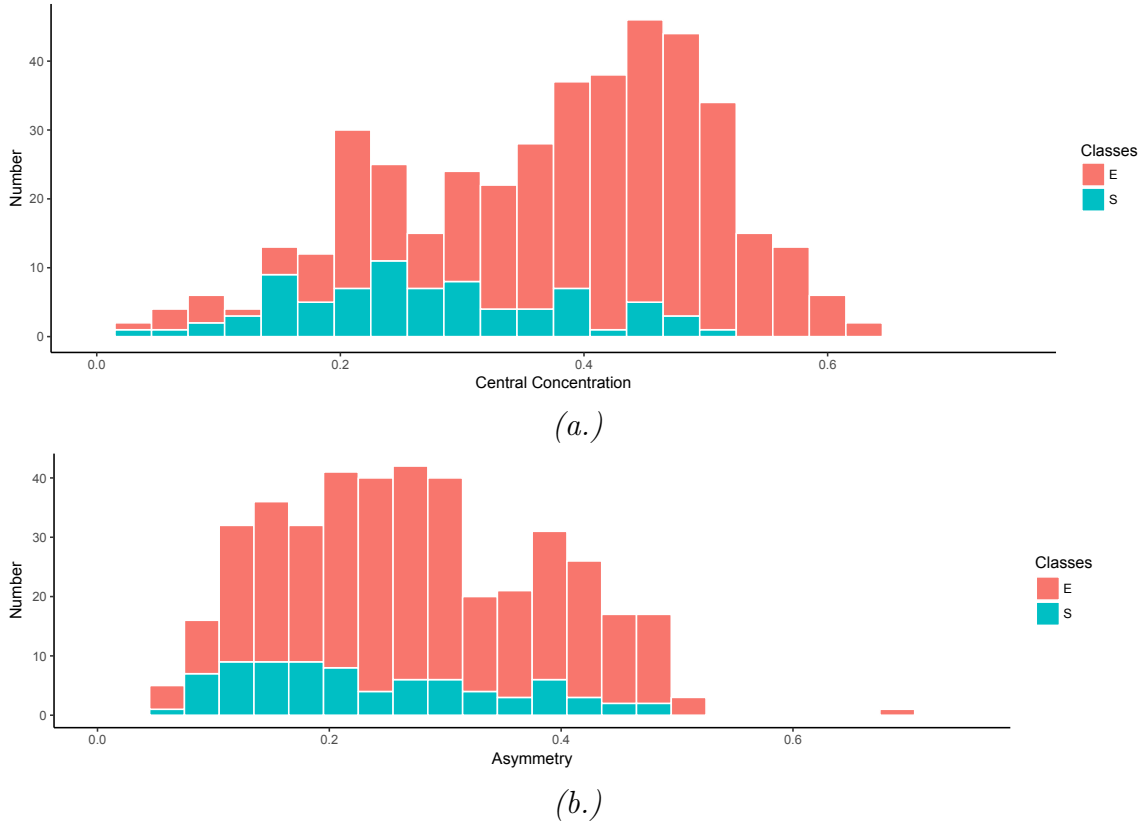


Figure 36: Histogram of Hubble Types as functions of central concentration (top) and asymmetry (bottom). See text for details.

We compare the A versus Gini plane to the A versus Theil plane in Figures 38 (a) and (b). In Figure 38 (b), we define the region bounded by $\text{Theil} \geq 0.05$ as the area containing mainly early-type galaxies (the “E/S0” region), while the region outside of this contains mainly late-type (“S/Irr”) galaxies. Out of the 289 galaxies in the “E/S0” region, 94% are early-type, and out of the 131 in the “S/Irr” region, 47% are late-type. There is a larger fraction of contamination in the “S/Irr” region in the A versus Theil than in the A versus Gini plane, which can be attributed to the smaller spread of data on the former plane than on the latter. Meaning, the greater spread in the data in the A versus Gini plane causes a smaller percentage of contamination of the early-type galaxies in the late-type region. Nonetheless, 271 of 342 (*i.e.* 79%)

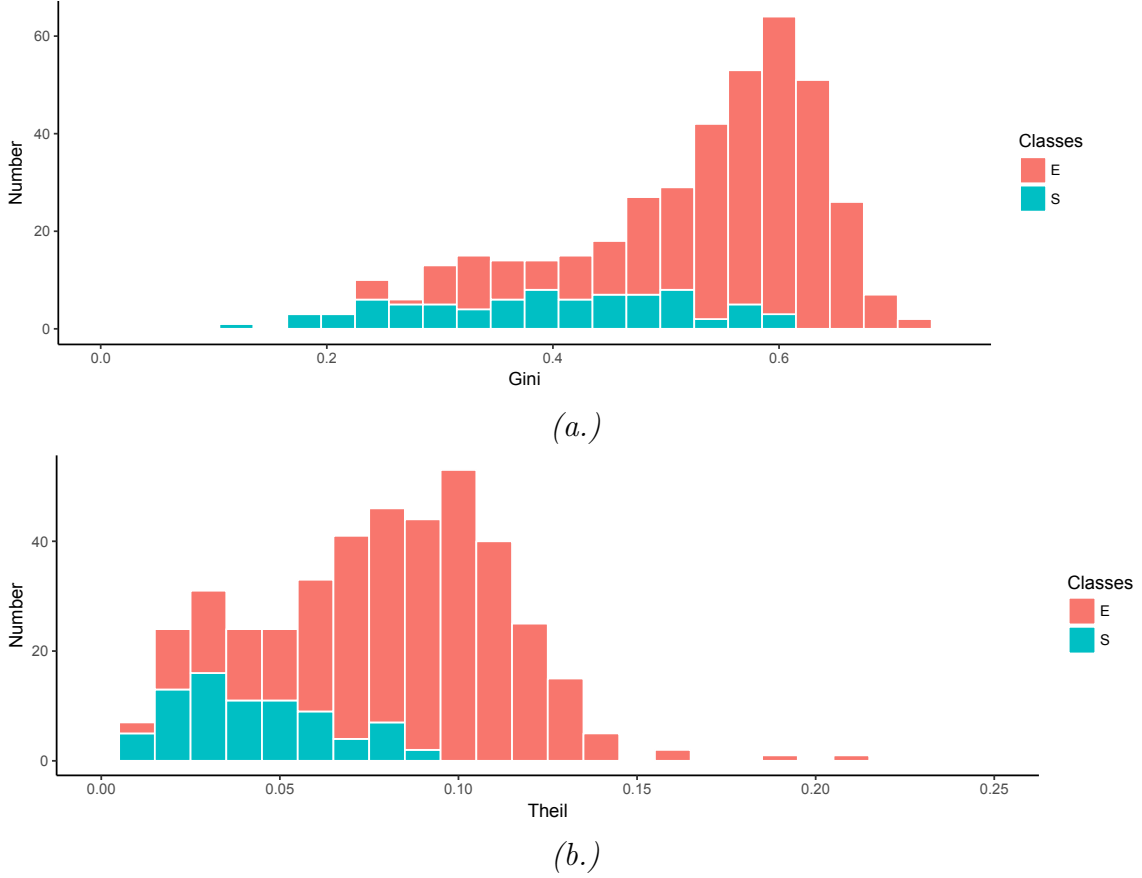


Figure 37: Histogram two Hubble Types as functions of Gini (top) and Theil (bottom). See text for details.

early-type galaxies are plotted in the “E/S0” region, and 61 of 79 (*i.e.* 77%) late-type galaxies are in the “S/Irr” region.

In the case of the C versus A distribution as shown in Figure 39, we again define the region where $C \geq 0.36$ as the “E/S0” classification area belonging mainly to early-type galaxies, and the region below containing predominantly late-type galaxies. In the “E/S0”, 215 out of 231 (*i.e.* 93%) galaxies are early-type and in the “S/Irr” region, 63 of 189 (*i.e.* 33%) galaxies are late-type. The majority of the early-type galaxies are plotted within the “E/S0” region.

For the plot of C versus Gini values, $C \geq 0.345$ and $\text{Gini} \geq 0.43$ define the “E/S0”

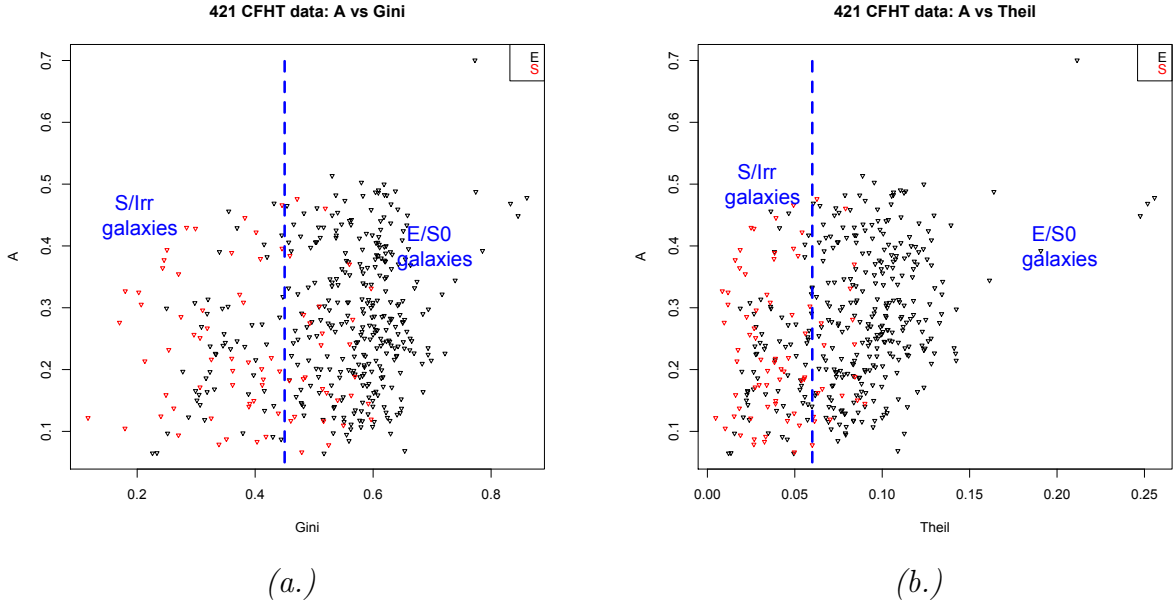


Figure 38: A versus Gini (a.) and A versus Theil (b.) plots of 421 CFHT galaxies.

region, and the outside area is the “S/Irr” zone. Just as previously seen with the EFIGI data, this distribution contains a majority of early-type galaxies in the “E/S0” region, but there is a higher rate of contamination of early-type galaxies in the “S/Irr” region. About 71% of 342 early-type galaxies are found in the “E/S0” region, and about 73% of the 79 late-type galaxies are in the “S/Irr” region.

For the plot of C versus Theil values, $C \geq 0.31$ and $\text{Theil} \leq 0.17$ define the “E/S0” region. We find a higher rate of contamination of early-type galaxies in the “S/Irr” region than for the C versus Gini plot, since the data are not as widely distributed as in the C versus Gini plane. We find 263 of 342 (*i.e.* 77%) early-type galaxies in the “E/S0” region and 53 out of 79 (*i.e.* 67%) of late-type galaxies in the “S/Irr” region.

In order to classify all galaxies from the 15 CFHT cluster data, we apply the technique developed in the previous chapter — applying two planes to the data: A versus Gini $\rightarrow A$ versus Theil. Figure 42 illustrates the stages of this analysis.

Figure 42(a.) depicts the distribution of the 421 visually classified CFHT galaxies

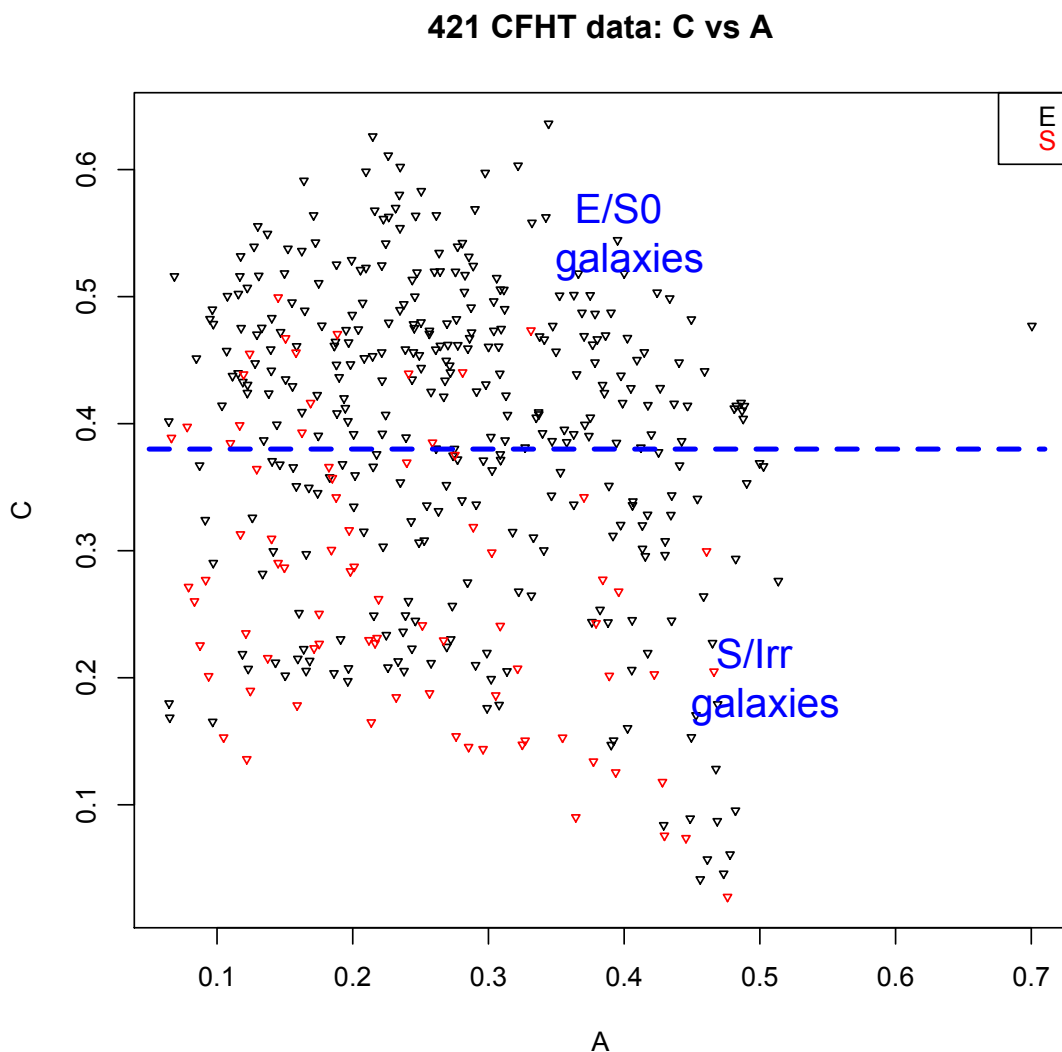


Figure 39: C versus A plot of 421 CFHT galaxies.

on the A versus Gini plane, where the galaxies are sorted into two broad bins — E and S. In Figure 42(b.), we label the early-type galaxies in the “S/Irr” as “E*” and late-type galaxies found in the “E/S0” region as “S*”. Next, we apply the A versus Theil selection method to the data in Figure 42(b.). The result is shown in Figure 42(c.). The final step in this classification process is depicted in Figure 42(d.), where the classified data are plotted on the A versus Theil plane. Galaxies classified as “E*”

421 CFHT galaxies: C vs Gini

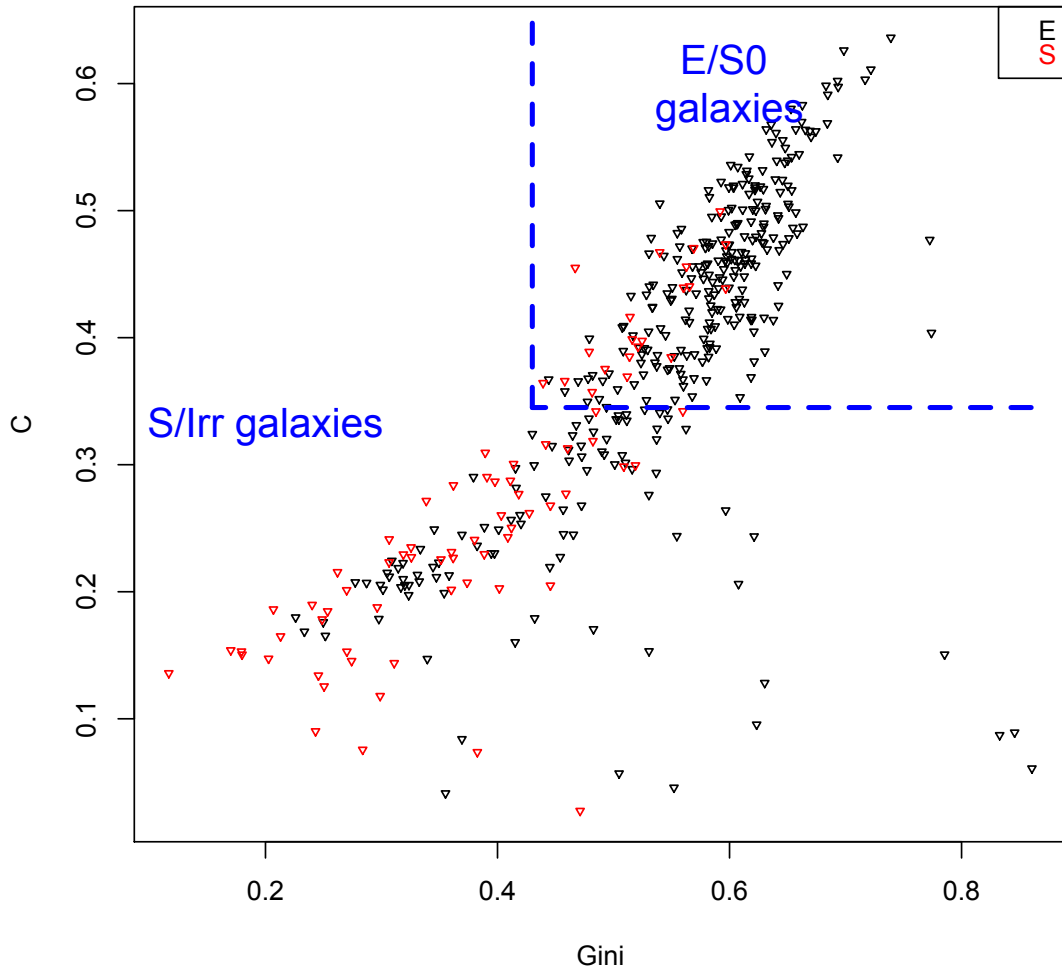


Figure 40: C versus Gini plot of 421 CFHT galaxies.

or “S*” are grouped together into the “T” class, while early-type galaxies that remain “early-type” through the two filters are “E”. Likewise, galaxies that were consistently flagged as late-type are labeled “S”, as shown in Figure 42(d).

Of the 342 visually classified early-type galaxies and 79 visually classified late-type galaxies, we find that 269 are classified as “E”, 51 as “S”, and 100 as the mixed class “T”. Unlike the other plane combinations tested (which were discussed in the previous

421 CFHT galaxies: C vs Theil

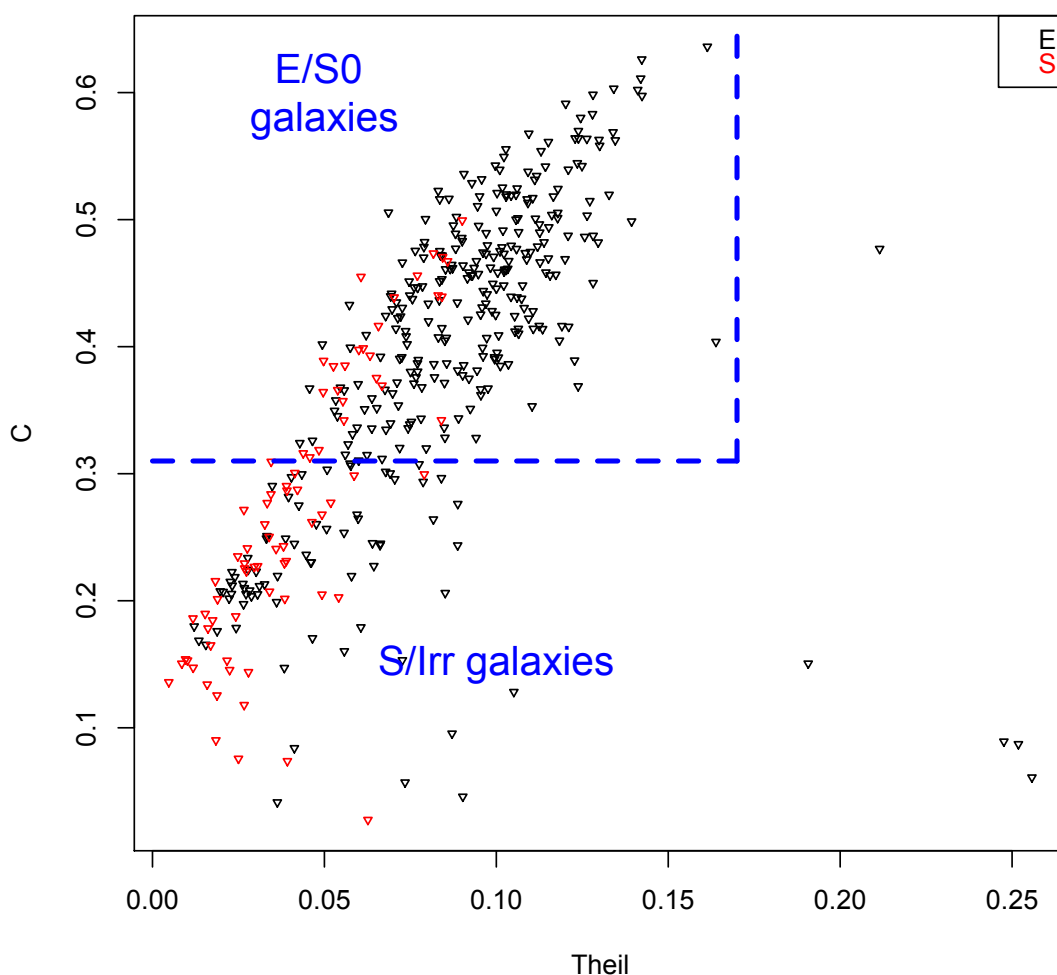


Figure 41: C versus Theil plot of 421 CFHT galaxies.

chapter), applying the A versus Gini \rightarrow A versus Theil planes to the data proves to classify the majority of visually classified early and late type galaxies into their respective categories of “E” and “S” galaxies, while producing a smaller mixed class of galaxies. This mixed class “I” is approximately 24% of the sample, while for the other combinations of planes, the population of galaxies classified as “I” is greater.

We apply this method to classify all galaxies in the 15 CFHT cluster sample. A

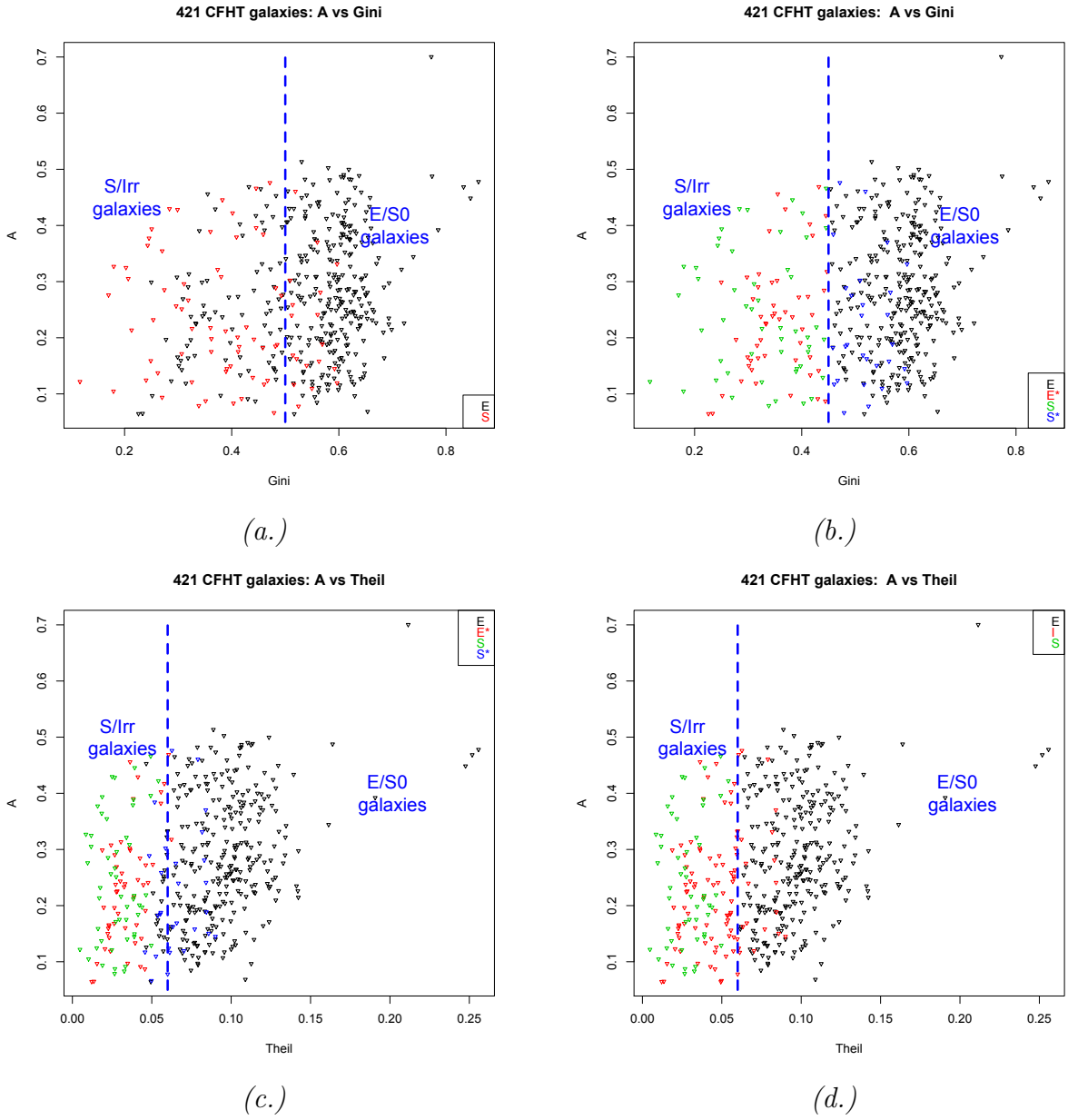


Figure 42: Applying two planes — the A versus Gini \rightarrow A versus Sheil — to 421 galaxies from the CFHT sample.

sample of the results is presented in Table 7. This table contains eighteen columns with the following information:

- Column 1: The cluster name to which the galaxy belongs.
- Column 2: The center of the galaxy along the x-axis (in pixels).
- Column 3: The center of the galaxy along the y-axis (in pixels).
- Column 4: Right Ascension (J2000) in degrees.
- Column 5: Declination (J2000) in degrees.
- Column 6: The count of how many non-zero pixels are in the galaxy image (AREA).
- Column 7: Isophotal magnitude, defined as: $\text{IMAG} = -2.5 * \text{LOG}_{10}(\text{FLUX} / \text{ZEROMAG})$.
- Column 8: Circular Aperture Magnitude: $\text{CAMAG} = -2.5 * \text{LOG}_{10}(\text{CIRCULARFLUX} / \text{ZEROMAG})$.
- Column 9: Mean surface brightness in units of magnitude/arcseconds² and is defined as: $\text{MSB} = -2.5 * \text{LOG}_{10}(\text{FLUX} / (\text{AREA} * (\text{Pixelscale})^2) / \text{ZEROMAG})$.
- Column 10: Central surface brightness (magnitude/arcseconds²) : $\text{CSB} = -2.5 * \text{LOG}_{10}(\text{CENTRALFLUX} / (\text{CENTRALAREA} * \text{Pixelscale})^2) / \text{ZEROMAG}$.
- Column 11: Central Concentration (C).
- Column 12: Asymmetry (A).
- Column 13: Gini coefficient.
- Column 14: Theil index.
- Column 15: Second-order moment of the brightest 20% region of a galaxy (M_{20}).
- Column 16: Bulge-to-disk (B/D) ratio.

- Column 17: Bulge-to-total (B/T) ratio.
- Column 18: Morphological software classification.

A total of 35,914 galaxies were studied in this sample. As with any current morphological software, there are limits to the features it can analyze. For example, objects that are too faint will not have accurate classifications through either the visual or automatic methods. Therefore, we restrict our sample by first calculating the surface brightness threshold (μ_T) for each cluster. The calculation of this value is described in Section 2.3.

To improve classification further, we separate galaxies into three categories — bright (B), dim (D), and not classified (N). We classify galaxies as B, D, or N by measuring the full width at half maximum (FWHM) for stars in each cluster. We use the “r” feature of *imexamine* command in IRAF to measure FWHM from the radial profile of stars in cluster images. The radial profile plot displays the brightness of pixels as a function of radius. We then average the FWHM values and define a minimum diameter to be $3 \times \text{FWHM}$. After averaging the FWHM values for a certain number of stars in the cluster image, we calculate the area of the brightest region of each galaxy (in units of pixels²). After calculating the total ADU inside this square area, we convert this value into ADU/1”.

Our morphological software measures the isophotal area of each galaxy. We compare the isophotal area of each galaxy to the isophotal limit we calculate using FWHM. Galaxies that are smaller than the isophotal area limit are labeled dim (D) and are considered too small for accurate morphological measurements to be performed. Galaxies that have isophotal area greater than this limit are labeled bright (B).

We find that 5,361 galaxies are classified as “B” in the CFHT sample. After applying the A versus Gini \rightarrow A versus Theil planes to these data, we find that of

Table 7: Sample of the morphological classification and GALFIT analysis of 15 CFHT clusters.

| Cluster C | x A | y Gini | RA Theil | Dec M20 | AREA B/D | IMAG B/T | CAMAG Class. | MSB | CSB |
|--------------|---------|-----------|-------------|------------|-------------|-------------|-----------------|-------|-------|
| A76 | 12074.6 | 8599.8 | 9.86 | 6.73 | 31580 | 13.02 | 14.35 | 20.60 | 17.50 |
| A76 | 0.11 | 0.60 | 0.80 | 0.00 | 1.50 | 0.60 | E | | |
| A76 | 7037.4 | 7730.6 | 10.12 | 6.69 | 649 | 17.88 | 17.80 | 21.24 | 19.98 |
| A76 | 0.25 | 0.38 | 0.04 | 0.02 | 1.04 | 0.51 | S | | |
| A98N | 10841.6 | 12097.8 | 11.56 | 20.58 | 16728 | 14.55 | 15.42 | 21.45 | 18.18 |
| A98N | 0.15 | 0.66 | 0.11 | 0.01 | NA | NA | E | | |
| A98N | 9414 | 13419.1 | 11.64 | 20.65 | 498 | 19.32 | 19.22 | 22.40 | 21.41 |
| A98N | 0.20 | 0.33 | 0.03 | 0.02 | NA | NA | S | | |
| A98 | 9651.7 | 9905.7 | 11.62 | 20.47 | 17823 | 14.54 | 15.80 | 21.50 | 18.68 |
| A98 | 0.26 | 0.64 | 0.10 | 0.01 | NA | NA | E | | |
| A350 | 13862.3 | 10509.3 | 36.22 | -9.82 | 4472 | 14.50 | 16.55 | 19.96 | 18.97 |
| A350 | 0.09 | 0.85 | 0.25 | 0.02 | NA | NA | E | | |
| A350 | 11775.6 | 10114.6 | 36.33 | -9.84 | 619 | 17.78 | 17.76 | 21.09 | 19.27 |
| A350 | 0.43 | 0.57 | 0.09 | 0.04 | NA | NA | E | | |
| A351 | 12009.6 | 14224.6 | 36.32 | -8.69 | 16723 | 14.27 | 15.29 | 21.16 | 18.21 |
| A351 | 0.13 | 0.62 | 0.09 | 0.00 | 1.31 | 0.57 | E | | |
| A351 | 11863.2 | 13771.6 | 36.33 | -8.72 | 17133 | 14.36 | 17.02 | 21.28 | 20.18 |
| A351 | 0.42 | 0.62 | 0.09 | 0.00 | NA | NA | E | | |
| A362 | 18047 | 16110 | 37.92 | -4.88 | 5396 | 15.50 | 16.30 | 21.17 | 18.66 |
| A362 | 0.47 | 0.53 | 0.07 | 0.01 | NA | NA | E | | |
| A362 | 17739.9 | 16395.1 | 37.94 | -4.87 | 213 | 18.92 | 18.77 | 21.08 | 19.94 |
| A362 | 0.46 | 0.48 | 0.07 | 0.07 | NA | NA | E | | |
| A655 | 12624 | 12128 | 126.37 | 47.13 | 30944 | 13.79 | 15.15 | 21.35 | 18.35 |
| A655 | 0.27 | 0.63 | 0.09 | 0.00 | NA | NA | E | | |
| A655 | 13403.4 | 10833.8 | 126.31 | 47.07 | 3927 | 15.83 | 16.18 | 21.15 | 18.05 |
| A655 | 0.26 | 0.66 | 0.12 | 0.02 | 1.26 | 0.56 | E | | |
| A795 | 11296.3 | 12723.6 | 141.08 | 14.24 | 4652 | 16.85 | 17.48 | 22.36 | 19.50 |
| A795 | 0.46 | 0.50 | 0.07 | 0.01 | 1.67 | 0.63 | E | | |
| A795 | 14177.4 | 10332.4 | 140.93 | 14.11 | 474 | 18.38 | 18.33 | 21.40 | 19.57 |
| A795 | 0.41 | 0.59 | 0.10 | 0.04 | 0.98 | 0.49 | E | | |
| A1920 | 19613 | 16895.1 | 216.85 | 55.75 | 17507 | 14.24 | 15.93 | 21.19 | 19.18 |
| A1920 | 0.33 | 0.48 | 0.05 | 0.00 | NA | NA | I | | |
| A1920 | 18727.8 | 18769.2 | 216.93 | 55.85 | 3218 | 15.81 | 16.37 | 20.91 | 18.66 |
| A1920 | 0.36 | 0.48 | 0.06 | 0.01 | 1.15 | 0.53 | I | | |
| A1940 | 14549.3 | 4871.3 | 218.87 | 55.13 | 20854 | 14.57 | 15.92 | 21.70 | 18.82 |
| A1940 | 0.24 | 0.63 | 0.12 | 0.01 | NA | NA | E | | |
| A1940 | 11693.4 | 4164.8 | 219.13 | 55.10 | 2579 | 16.53 | 16.89 | 21.40 | 18.85 |
| A1940 | 0.44 | 0.56 | 0.08 | 0.02 | NA | NA | E | | |
| A2100 | 8530.8 | 10247.3 | 234.19 | 37.62 | 5858 | 14.74 | 15.52 | 20.50 | 18.00 |
| A2100 | 0.40 | 0.52 | 0.06 | 0.01 | NA | NA | E | | |
| A2100 | 10603.6 | 11787.2 | 234.05 | 37.70 | 2404 | 15.66 | 16.04 | 20.45 | 18.13 |
| A2100 | 0.39 | 0.51 | 0.06 | 0.01 | NA | NA | E | | |
| A2107 | 8634.9 | 6046 | 234.99 | 21.47 | 32338 | 12.81 | 14.82 | 20.42 | 18.19 |
| A2107 | 0.38 | 0.55 | 0.05 | 0.00 | 1.12 | 0.53 | I | | |
| A2107 | 13693 | 12985 | 234.71 | 21.83 | 8797 | 14.57 | 15.71 | 20.77 | 18.50 |
| A2107 | 0.35 | 0.53 | 0.06 | 0.00 | NA | NA | E | | |
| A2107 | 11255.4 | 3150.4 | 234.85 | 21.32 | 770 | 18.25 | 18.24 | 21.80 | 20.96 |
| A2107 | 0.18 | 0.23 | 0.01 | 0.01 | NA | NA | S | | |
| A2147 | 9536 | 13020 | 240.57 | 15.97 | 37783 | 13.16 | 15.07 | 20.94 | 18.48 |
| A2147 | 0.40 | 0.51 | 0.05 | 0.00 | NA | NA | I | | |
| A2147 | 1709.6 | 13532.7 | 240.99 | 16.00 | 19042 | 14.08 | 15.50 | 21.11 | 18.51 |
| A2147 | 0.39 | 0.48 | 0.05 | 0.00 | 1.74 | 0.63 | I | | |
| A2199 | 9962 | 11151 | 247.16 | 39.55 | 86515 | 11.68 | 14.32 | 20.36 | 17.96 |
| A2199 | 0.42 | 0.61 | 0.07 | 0.00 | 0.69 | 0.41 | E | | |
| A2199 | 7356.6 | 8784.8 | 247.33 | 39.43 | 28498 | 13.15 | 14.31 | 20.62 | 17.38 |
| A2199 | 0.14 | 0.65 | 0.10 | 0.01 | NA | NA | E | | |
| A2688 | 10404 | 10572 | 0.03 | 15.83 | 4022 | 15.96 | 16.51 | 21.31 | 18.89 |
| A2688 | 0.27 | 0.53 | 0.07 | 0.01 | NA | NA | E | | |

the bright galaxies, 2,073 are classified as “E”, 261 are classified as “T”, and 3,026 are “S”. For the “D” galaxies and the ones that were initially excluded from the sample, we assign “N” for “Not Classified.” The morphological classifications in Column 18 of Table 7 are based on this method.

4.1.2 *Nucleated vs. Non-nucleated Dwarf Galaxies*

Unlike luminous, high mass galaxies — which are classified by the Hubble (or a modified Hubble) system — dwarf galaxies are faint, small, low mass galaxies that do not fit on the Hubble scheme. However, dwarf galaxies are the most abundant galaxies in the Universe. It is believed that they may be the building blocks of much larger stellar systems, therefore studying their structure may offer clues about the formation and evolution of normal galaxies (Oh & Lin 2000). Also, because of their low mass, dwarf galaxies can be used to study the dense environment of galaxy clusters since as dwarf galaxies enter the cluster environment, they experience ram pressure and galaxy harassment (Rude 2015). These mechanisms can have an effect on star formation and the morphology of galaxies. For this thesis, dwarf galaxies are defined to have an absolute magnitude in the r -band of $-19.5 \leq M_r \leq -17.0$ (Rude 2015; Barkhouse 2009). Based on their photometric appearance and gas content, there are currently three main classes of dwarf galaxies (Binggeli & Cameron 1991; Grebel 1998; Oh & Lin 2000):

- **Dwarf Elliptical galaxies (dE):** are observed to have many similar properties of normal elliptical galaxies, but are small in size. In other words, they can be defined as having a flatter brightness profile than giant elliptical galaxies (Sandage & Binggeli 1984). dE galaxies appear to have little or no gas, and have no evidence of recent star formation. These galaxies can be further divided into

nucleated and non-nucleated, meaning there is a presence of a central nucleus or no central nucleus, respectively.

- **Dwarf Spheroidal galaxies (dSph):** fainter than most dwarf elliptical galaxies, less massive, and are more spheroidal in shape rather than elliptical. Similar to dE galaxies, dSphs contain little or no gas and no recent star formation. dSph galaxies are observed to have no pronounced nucleus and little central concentration.
- **Dwarf Irregular galaxies (dIrr):** are gas-rich, but low-mass irregular shaped galaxies. They are observed to have recent or ongoing star formation.

There are also certain types of dwarf galaxies in-between these three main classes, such as dS0s. It is observed that dwarf elliptical galaxies near the centers of galaxy clusters are mostly nucleated, while those in the outskirts of clusters are non-nucleated (Binggeli & Cameron 1991; Oh & Lin 2000; Mistani *et al.* 2015). Observations also suggest that the two types of dwarf galaxies may have similar origins (Conselice *et al.* 2001).

There are various theoretical methods that explain the formation of nuclei in dwarf galaxies (*e.g.* Oh & Lin 2000; Conselice *et al.* 2001; Grant *et al.* 2005; Lisker *et al.* 2006). Since the majority of nucleated dwarf galaxies are found in cluster environments, it may be that gravitation effects of the cluster core may induce nuclei formation. Due to their low mass, dwarf galaxies are more likely susceptible to gravitational effects. Star formation may be induced in the center of dwarf galaxies located in very dense environments, which develops into a nucleus of such galaxies. Another possibility is that through the process of harassment, nearby spiral galaxies may transfer material onto dwarf elliptical galaxies in the dense cluster environment, which would produce bursts of star formation at the centers of those dwarf galaxies.

Similarly, material may be transferred to dwarf ellipticals in the cluster core through ram-pressure stripping of irregular galaxies. It could also be that the nuclei of certain dwarf galaxies may be produced from globular clusters that have sunk down to the center due to dynamical friction.

In this section of the thesis, we study the morphologies of nucleated versus non-nucleated dwarf galaxies in cluster environments. Using the Λ CDM model of the Universe, in which $H_0 = 70$ km/s/Mpc, $\Omega_\Lambda = 0.7$, and $\Omega_m = 0.3$, we find the absolute magnitude of the approximately 36,000 galaxies in our CFHT sample using the distance modulus ($\mu = 5 \cdot \text{Log}(d) - 5$, where d is in units of parsecs) and the k-corrections:

$$M = m - \mu - \text{k-correction.} \quad (4.1)$$

Since galaxies are located at different redshifts, in order to accurately measure their magnitude, k-corrections must be applied. K-corrections are applied when an astronomical measurement is performed through a single filter (*i.e.* light from an object is measured for only certain wavelengths), therefore only a portion of the total light is seen, and this light is redshifted relative to the rest frame of the observer. As described in Rude (2015), we use the following equation to estimate k-corrections:

$$\sum_{i=0}^5 \sum_{j=0}^3 a_{ij} z^i c^j, \quad (4.2)$$

where a_{ij} is the coefficient, z is the redshift, and c is the $u-r$ color. From the bright (B) galaxies in our CFHT sample, we select galaxies with faint absolute magnitudes

of the range $-19.5 \leq M_r \leq -15.0$, which we define as dwarf galaxies. The histogram as a function of central concentration for these 2,967 galaxies is shown in Figure 43.

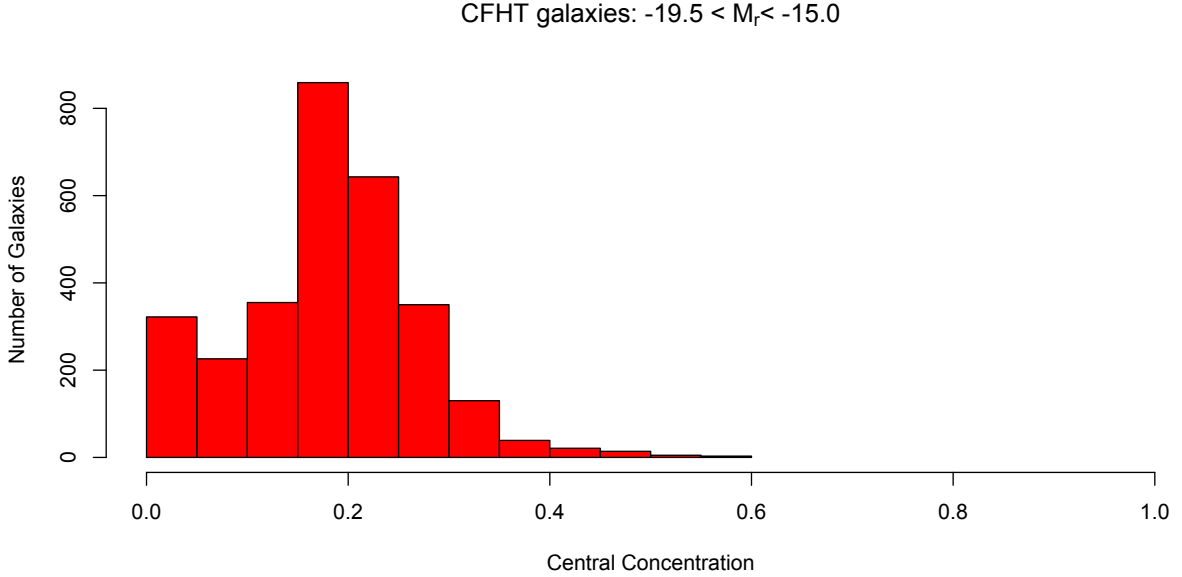


Figure 43: Histogram of dwarf CFHT galaxies as a function of central concentration.

Dwarf galaxies with large values of C are nucleated and those with small values of C are non-nucleated. We separate the dwarf galaxies into two, approximately equal samples: ones with high central concentration of $C \geq 0.25$ and ones with low central concentration ranging from $0 \leq C \leq 0.1$.

The distance to each galaxy (r) relative to the cluster center is found. The central coordinates of each cluster are found by locating the brightest cluster galaxy (BCG) (Rude 2015). X-ray data was also used to confirm that the choice of BCG in a cluster was correct by locating the nearest E-galaxy to the X-ray emission centroid. X-ray data can narrow down a BCG in a cluster in cases where there are two potential centers, such as in Abell 98. The pixel scale of the image can be used to convert the distance to megaparsecs (Mpc) using the angular diameter distance.

We find the distance from the center of the cluster as a fraction of r_{200} (*i.e.* r/r_{200}),

where r_{200} is defined as a radius within which the average density is 200 times the critical density of the universe ($\rho_c(z)$) at the cluster's redshift. The critical density, that which gives rise to a flat Universe, is defined as:

$$\rho_c(z) = \frac{3H^2(z)}{8\pi G}, \quad (4.3)$$

where H and G are the Hubble and gravitational constants, respectively. Following Rude (2015), the r_{200} values for each cluster are found from the velocity dispersion, σ_v , which are available from literature:

$$r_{200} = \frac{\sqrt{3}\sigma_v}{10H(z)}, \quad (4.4)$$

where the Hubble parameter at redshift z is:

$$H(z) = H_0 \sqrt{\Omega_m(1+z)^3 + \Omega_\Lambda}. \quad (4.5)$$

We probe the ratio of high- C dwarf galaxies versus low- C dwarf galaxies as a function of (r/r_{200}) binned by values of $(r/r_{200}) = 0.2$. Because the counts of high- C and low- C dwarf galaxies in every bin is random, the error is estimated by the Poisson statistical process as $N \pm \sqrt{N}$, where N is the number of galaxies and \sqrt{N} is the uncertainty in N .

Since we cannot assume the number of high- C dwarf galaxies is correlated with the number of low- C dwarf galaxies, *i.e.* these are independent variables, the error on the ratios of high- C vs. low- C in each of the five (r/r_{200}) bins is found in quadrature,

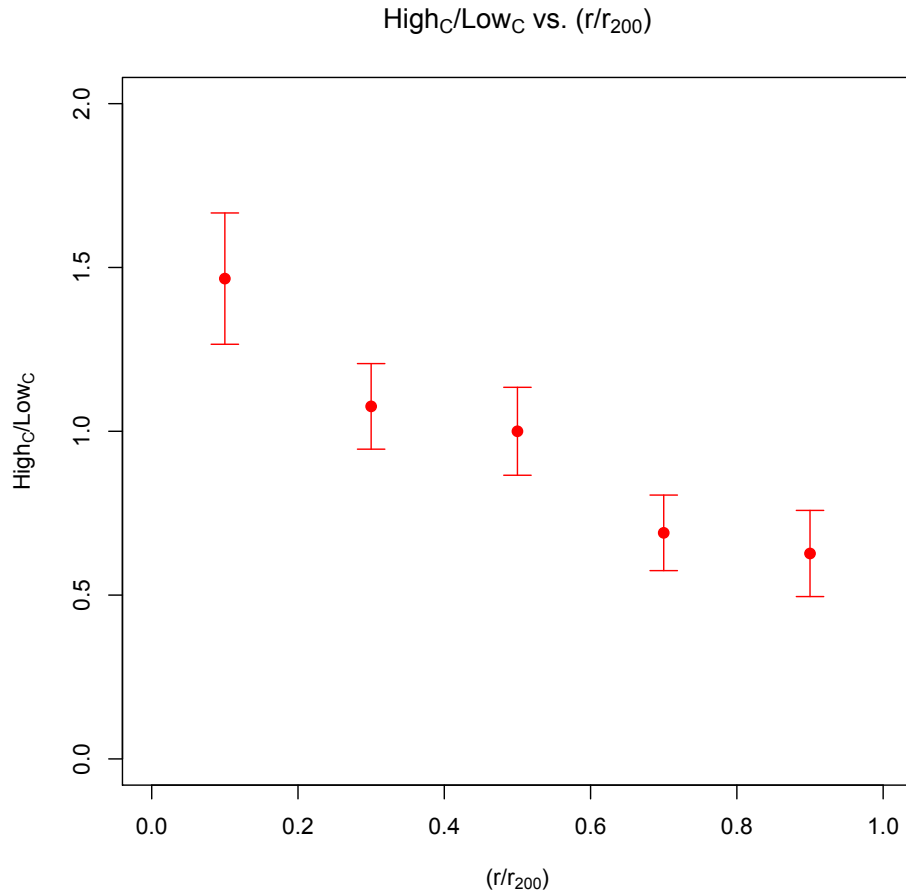


Figure 44: Ratio of high- C versus low- C dwarf galaxies from the CFHT sample as a function of (r/r_{200}) .

which is the square root of the sum of the squares. We find the ratio (R) of high- C vs. low- C in each (r/r_{200}) bin as:

$$R = \frac{N_{HC}}{N_{LC}}, \quad (4.6)$$

where N_{HC} is the number of high- C and N_{LC} is the number of low- C galaxies in an

(r/r_{200}) bin. The error is estimated as:

$$\frac{\delta R}{R} = \sqrt{\left(\frac{\sqrt{N_{HC}}}{N_{HC}}\right)^2 + \left(\frac{\sqrt{N_{LC}}}{N_{LC}}\right)^2}. \quad (4.7)$$

Figure 44 represents the plot of the ratio of the number of high- C versus low- C dwarf galaxies relative to the center of each cluster as a fraction of (r/r_{200}) . There is a large significant difference of 3.5σ between the inner most and outer most (r/r_{200}) bins. It can be seen that dwarf galaxies in the core of clusters are predominantly nucleated, high- C dE galaxies, whereas the outer regions of clusters contain greater numbers of non-nucleated, low- C dwarf galaxies. We further discuss these results in Section 7.3.

CHAPTER V

LOW-REDSHIFT DATA

5.1 Low-Redshift Abell Clusters

We analyze the properties of 57 low-redshift Abell galaxy clusters from Barkhouse *et al.* (2007). The clusters span redshifts of $0.04 \leq z \leq 0.20$. They were selected from a compilation of bright X-ray clusters from Einstein's IPC (Jones & Forman 1999). The determining criteria is outlined in Barkhouse (2007). The sample includes 47 clusters observed in B and Kron-Cousins R_C and I taken from the Kitt Peak National Observatory (KPNO) in Arizona, U.S.A., with the 0.9m telescope using the 2048 x 2048 pixel T2KA CCD. The field of view covered is $23.'2 \times 23.'2$ with a scale of $0.68''$ pixel⁻¹ by Lopez-Cruz (1997), Lopez-Cruz *et al* (1997), and Lopez-Cruz (2001). Using the same selection criteria, two galaxy clusters observed in B and R_C from Brown (1997) are also included with this sample. The images were observed using the same instrumental setup.

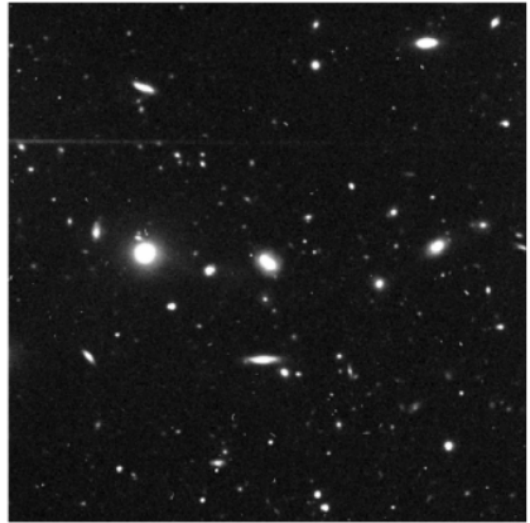
Eight clusters from Barkhouse (2003) are also included with this sample of 49 clusters. These eight clusters have a low-redshift range of $0.02 \leq z \leq 0.04$ and were obtained at KPNO with the 0.9m telescope using the 8K MOSAIC camera (8192 x 8192 pixels). The field of view covered is 1 deg.^2 with a pixel scale of $0.423''$ pixel⁻¹.

Figure 45 displays several R -band images of galaxies in the KPNO data set. Each postage stamp is 500 x 500 pixels in size. The galaxy analyzed is positioned at the center of each postage stamp, with its x and y image coordinates stated above each

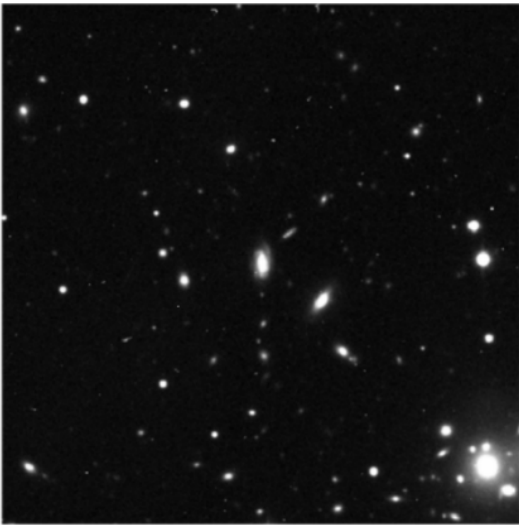
X: 1441
Y: 1522.4



X: 348.37
Y: 885.21



X: 878.4
Y: 1231.4



X: 3548
Y: 3341

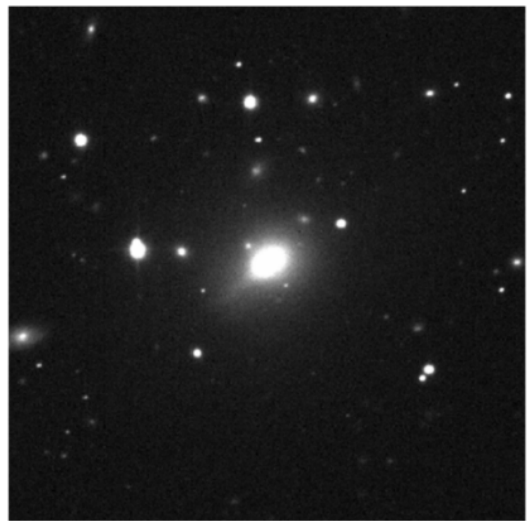


Figure 45: Sample of several r -band postage stamps of galaxies from the KPNO data set. See text for details.

259 visually classified KPNO galaxies

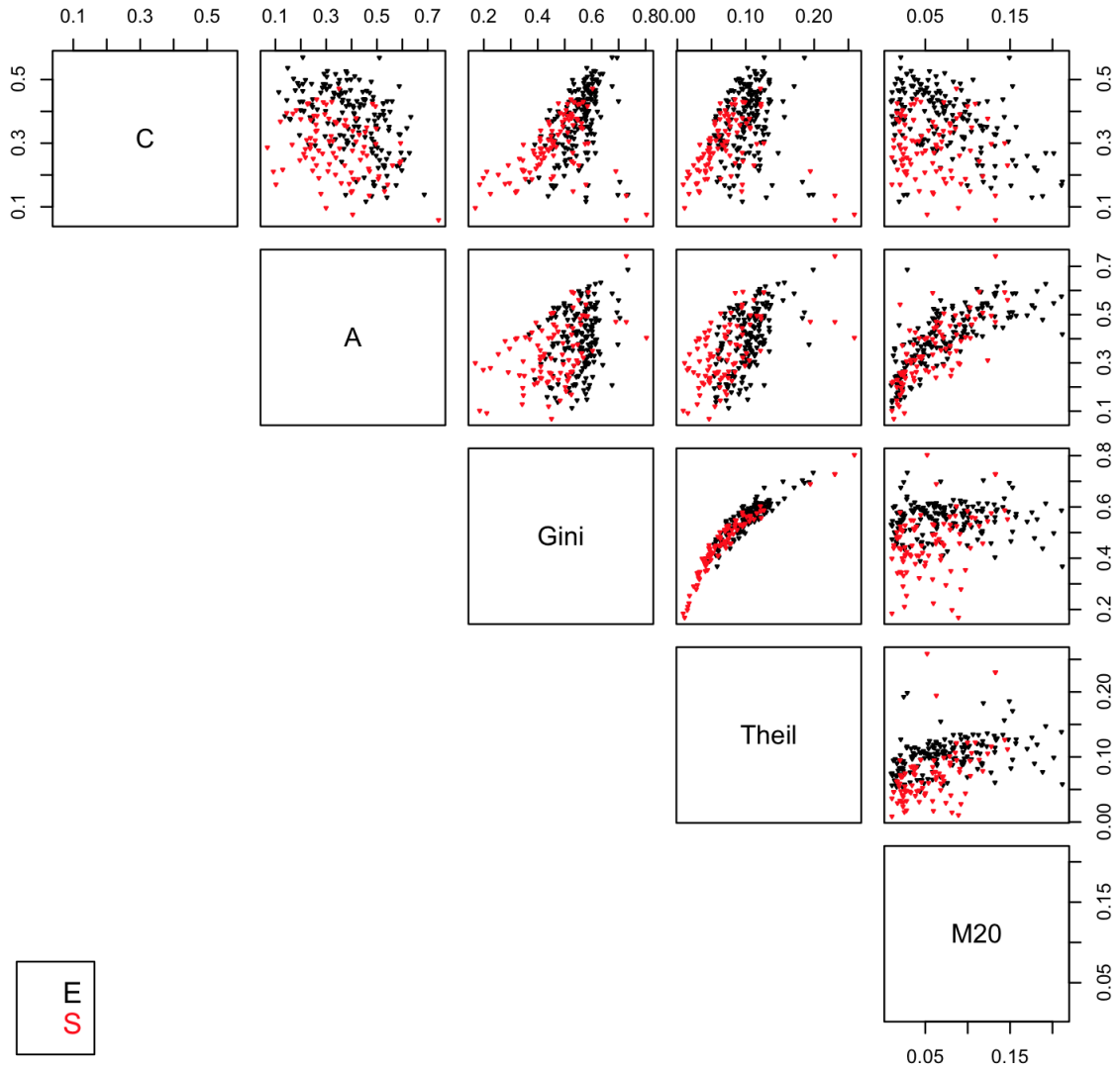


Figure 46: Relations between five parameters in the 259 galaxies from 57 KPNO galaxy clusters.

image in pixels. The galaxies are from various clusters in the data set. In the next section, we describe the analysis of this data.

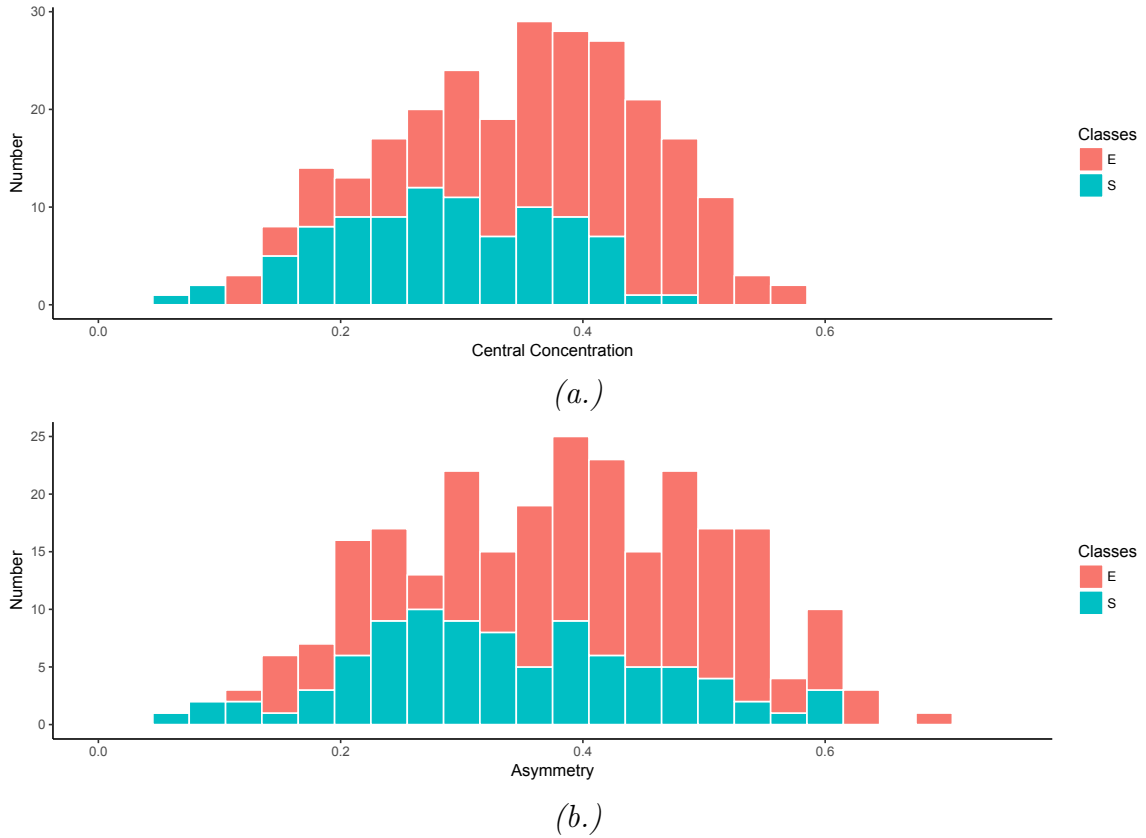


Figure 47: Histogram of two Hubble Types as a function of central concentration (top) and asymmetry (bottom) for KPNO. See text for details.

5.1.1 Analysis

We apply the methods described in Section 3.2.2 and the previous chapter to analyze the data from this sample. Out of the approximately 9,500 galaxies in this sample, we visually classify 259 bright galaxies with distinguishable structure. This sample contains 167 early-type galaxies and 92 late-type. Figure 46 illustrates the relationship between each parameter as a function of the others. The early-type galaxies are labeled as “E” and late-type as “S”.

Figures 47 and 48 display histograms for central concentration, asymmetry, Gini coefficients, and Theil indexes measured for the visually classified KPNO data. Early-type galaxies (E) are represented in red and late-type (S) are in blue. The histogram in Figure 47 (a) displays the galaxies in these two Hubble classes as a function of central concentration binned by values of $C = 0.03$. Figure 47 (b) is the histogram of the Hubble classes as a function of asymmetry, also binned by 0.03. Figure 48 (a) shows the Hubble Types as functions of the Gini coefficient, binned by $\text{Gini} = 0.03$. Figure 48 (b) shows the Hubble Types as functions of the Theil index, binned by 0.01.

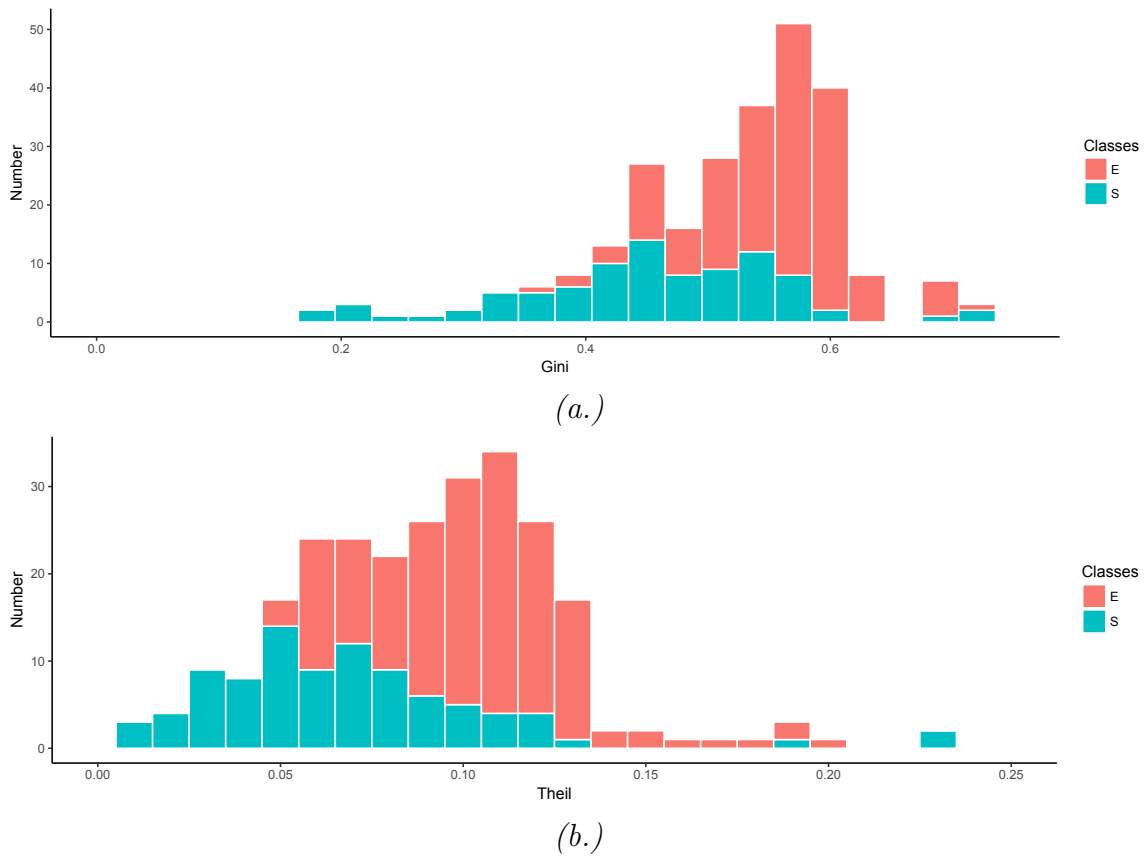


Figure 48: Histogram of two Hubble Types as functions of Gini (top) and Theil (bottom) for KPNO. See text for details.

The distribution of the 259 visually classified galaxies from the KPNO sample are
123

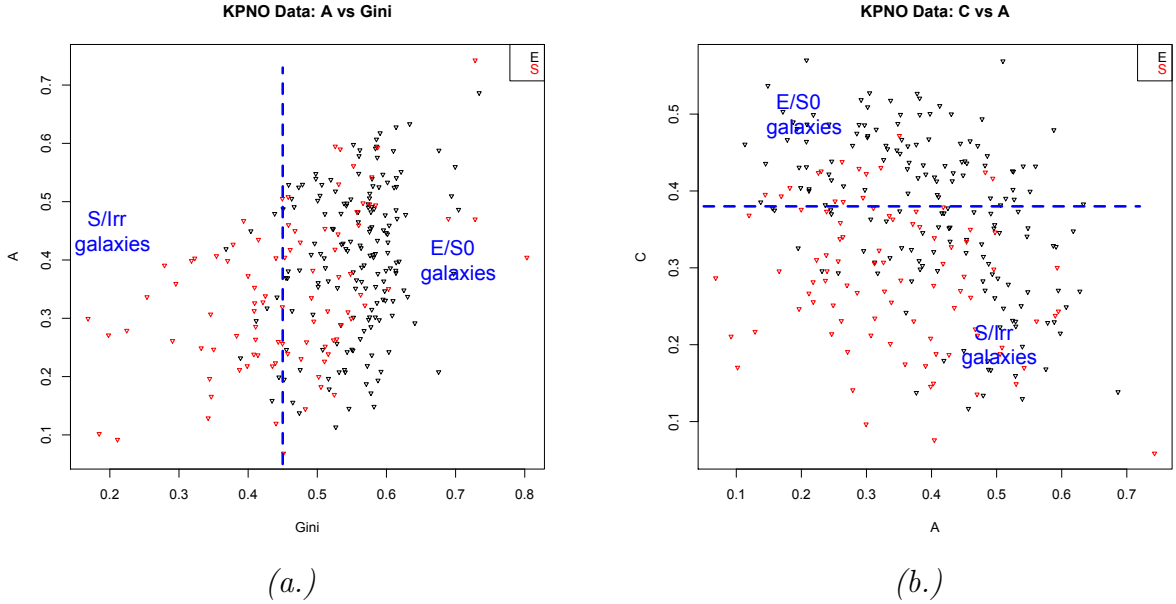


Figure 49: A versus Gini (a) and C versus A (b) plot of 259 KPNO galaxies.

plotted on the A versus Gini plane is shown in Figure 49 (a), with the regions defined in the previous chapter where $\text{Gini} \geq 0.45$ as the area containing mainly early-type galaxies, while the region outside of these bounds contains mainly late-type galaxies.

We find the majority of visually classified KPNO galaxies to be plotted in the defined classification regions as seen in Figure 49 (a). Approximately 76% of the galaxies plotted in the “E/S0” region are early-type (158 out of 207). Similarly, about 83% of the galaxies in the “S/Irr” region are late-type (43 out of 52).

On the C versus A plane, as depicted in Figure 49 (b), we define the rectangular region bound by $C \geq 0.38$ as the “E/S0” classification region belonging to mainly early-type galaxies. In the “E/S0” area, 242 of 263 (*i.e.* 92%) galaxies are early-type and in the “S/Irr” region, 58 out of 157 (*i.e.* 37%) galaxies are late-type. Due to the sparsity of the sample, there is greater contamination of early-type galaxies in the “S/Irr” region.

Figure 50 shows the final result of applying the A versus Gini \rightarrow A versus Theil

KPNO Data: A vs Theil

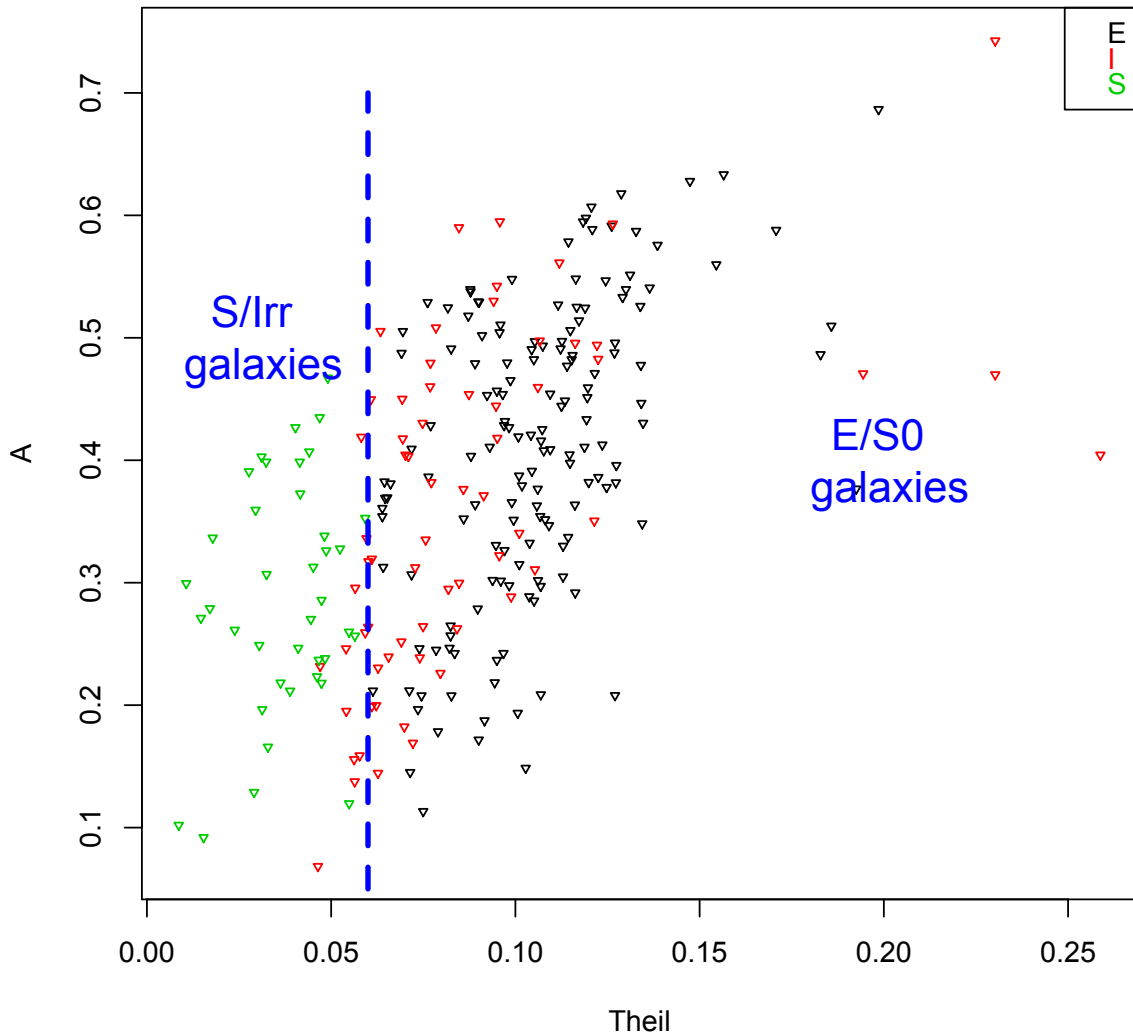


Figure 50: The final result of applying two cuts — A versus Gini \rightarrow A versus Theil — to 259 galaxies from the KPNO sample.

planes to the 259 KPNO galaxies. The method and labels of the graph are described in previous sections: Section 3.2.2 and Section 4.1.1.

Out of the 167 visually classified early-type galaxies and 92 visually classified late-type galaxies, we find that 154 are classified as “E”, 39 as “S”, and 66 as the mixed

class “T”. This mixed class “T” is approximately 25% of the sample. This sample is smaller than the visually classified CFHT galaxies and the EFIGI catalog we have studied previously.

As described in Section 4.1.1, we also classify galaxies into three categories — bright (B), dim (D), and not classified (N) — by calculating the area of the brightest region of each galaxy (in units of pixels²) and comparing this value to 3*FWHM of the cluster image. Out of the 9,487 KPNO galaxies in our study, we find that 4,200 galaxies are classified as bright.

5.2 WINGS

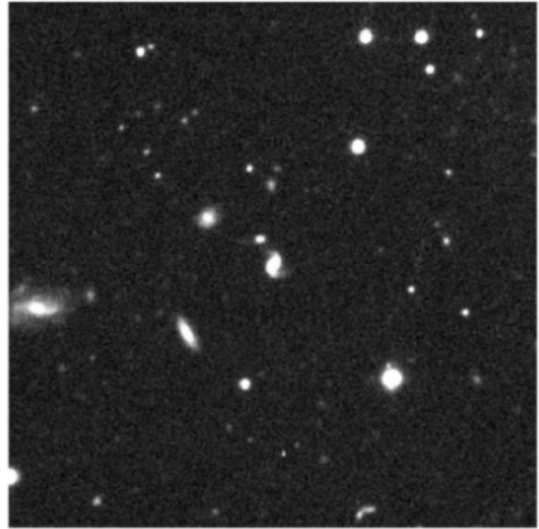
The Wide-field Nearby Galaxy-clusters Survey (WINGS: Fasano *et al.* 2003; Fasano *et al.* 2006; Fasano *et al.* 2012) is a two-band (the optical B and V) wide-field imaging survey of a complete, all-sky (galactic latitude of $|b| > 20^\circ$) X-ray selected sample of 77 clusters (41 in the Southern hemisphere and 36 in the Northern hemisphere) in the redshift range $0.04 < z < 0.07$. The upper redshift limit ensures adequate spatial resolution ($1'' = 1.3$ kpc at $z = 0.07$, $H_0 = 70$ km/s/Mpc). The central area is 1.5 Mpc² at $z = 0.04$. The clusters in the WINGS project have been selected from three X-ray flux limited samples. In the Northern hemisphere, the data is compiled from ROSAT All-Sky Survey data: the ROSAT Brightest Cluster Sample (Ebeling *et al.* 1998), and its extension (Ebeling *et al.* 2000). In the Southern hemisphere, the X-Ray-Brightest Abell-type Cluster sample (Ebeling *et al.* 1996) is used.

The goal of the WINGS project is to systematically study correlations between cluster properties and cluster galaxy populations. Therefore, a well-defined, large cluster sample is required, with available X-ray data and covering a wide range of optical and X-ray properties. In the Northern hemisphere, the images were acquired

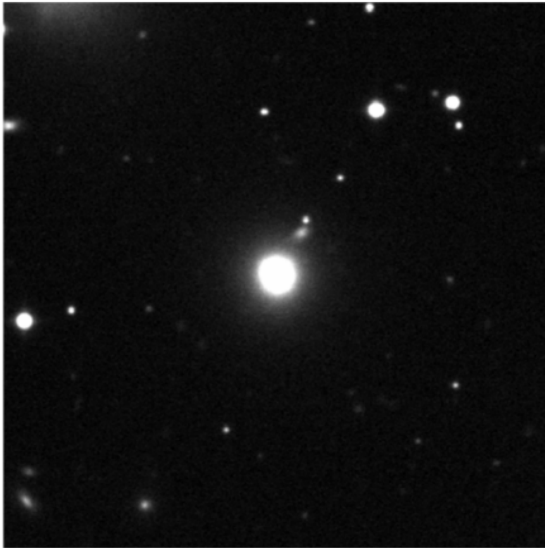
X: 5467.4434
Y: 5465.0521



X: 2522.5959
Y: 5029.3761



X: 7662.3749
Y: 1752.5964



X: 3548
Y: 3341

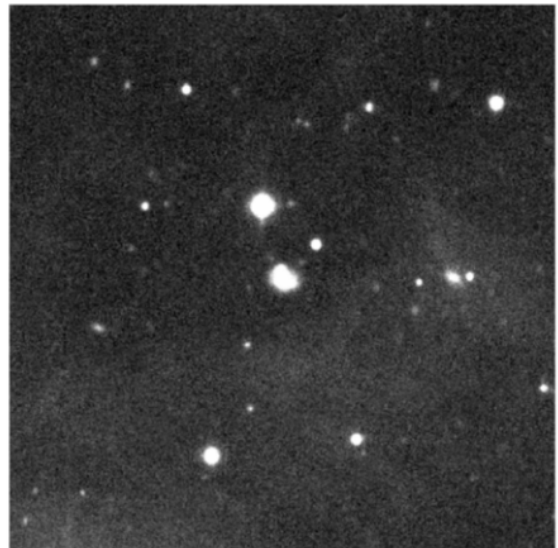


Figure 51: Sample of several r -band postage stamps of galaxies from the WINGS data set. See text for details.

using the Wide Field Camera (WFC), with a pixel scale of $0.33''/\text{pixel}$ and field of view of $34' \times 34'$, mounted on the INT-2.5 m telescope in La Palma (Canary Islands, Spain). In the Southern hemisphere, the Wide Field Imager (WFI), with a pixel scale of $0.238''/\text{pixel}$ and field of view of $34' \times 33'$, mounted on the MPG/ESO-2.2 m telescope in La Silla (Chile) was used. The WINGS catalog contains approximately 40,000 galaxies.

Figure 51 displays several V -band images from a sample of galaxies in the WINGS data set. Each postage stamp is 500×500 pixels in size. The galaxy analyzed is positioned in the center of each postage stamp, with its right ascension and declination stated above each image. The galaxies are selected from various clusters in the data set. In the next section, we describe the method of preparing the WINGS files for our morphology software and analyze the galaxies' classification results.

5.2.1 Analysis

Applying the methods described in Section 3.2.2 and the previous chapter, we analyze the data from the WINGS catalog. Out of the approximately 40,000 galaxies in this sample, we visually classify 608 bright galaxies possessing distinguishable structure. This sample contains 438 early-type galaxies and 170 late-type systems. Figure 52 illustrates the relationship between each parameter as a function of others. The early-type galaxies are labeled as "E" and late-type as "S".

Figures 53 and 54 display histograms of central concentration, asymmetry, Gini coefficients, and Theil indexes measured for the visually classified WINGS data. Early-type galaxies (E) are represented in red and late-type (S) are in blue. As seen previously, the histogram in Figure 53 (a) displays the galaxies in these two Hubble classes as a function of central concentration, also binned by values of $C = 0.03$. Figure 53 (b) is the histogram of the Hubble classes as a function of asymmetry, binned by 0.03.

608 visually classified WINGS galaxies

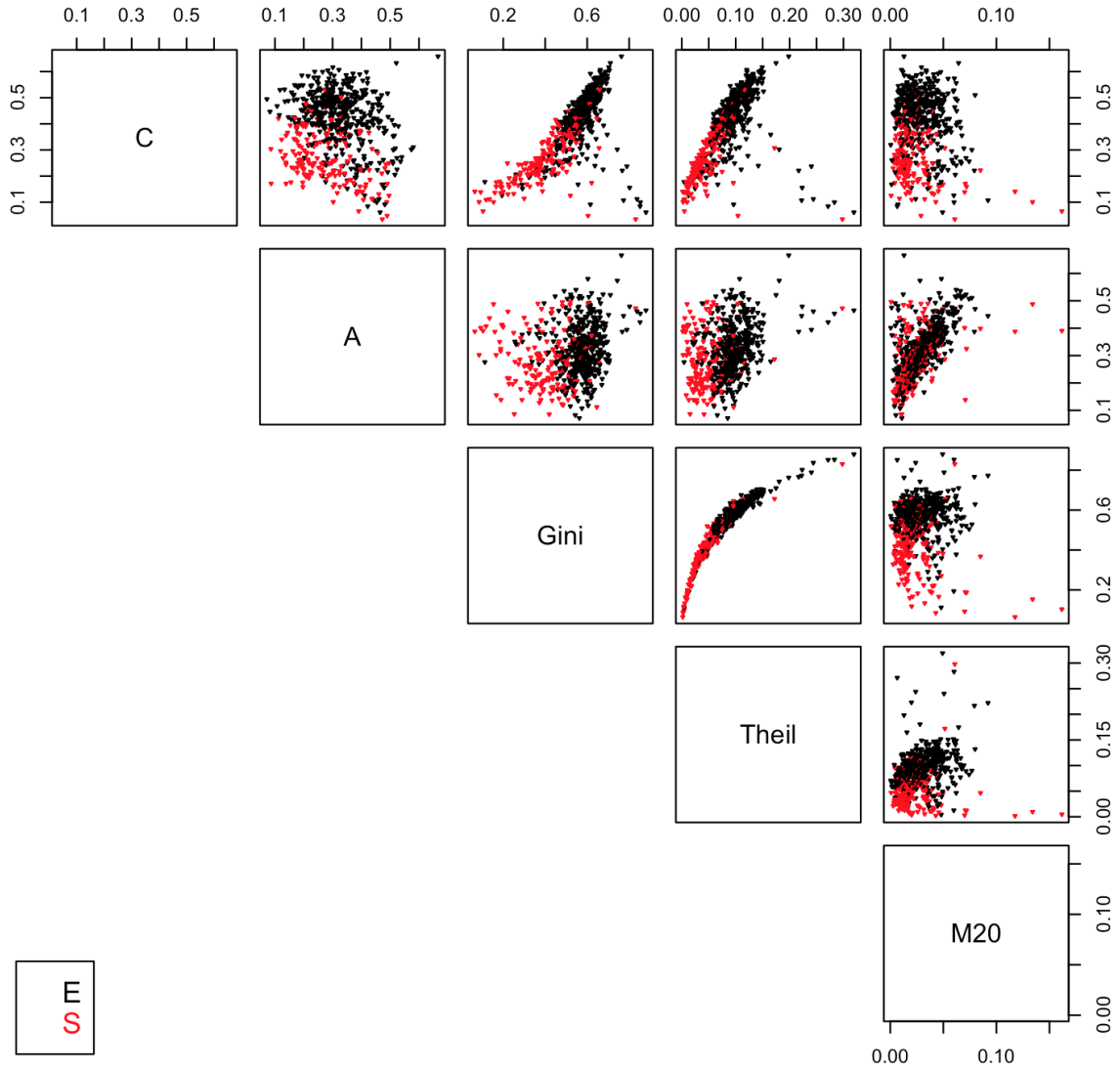


Figure 52: Relations between five parameters for the 608 visually classified galaxies from the WINGS galaxy clusters.

Figure 54 (a) shows the Hubble Types as functions of the Gini coefficient, binned by $Gini = 0.03$. Figure 54 (b) shows the Hubble Types as functions of the Theil index, binned by 0.01.

The distribution of 608 visually classified galaxies from the WINGS sample is

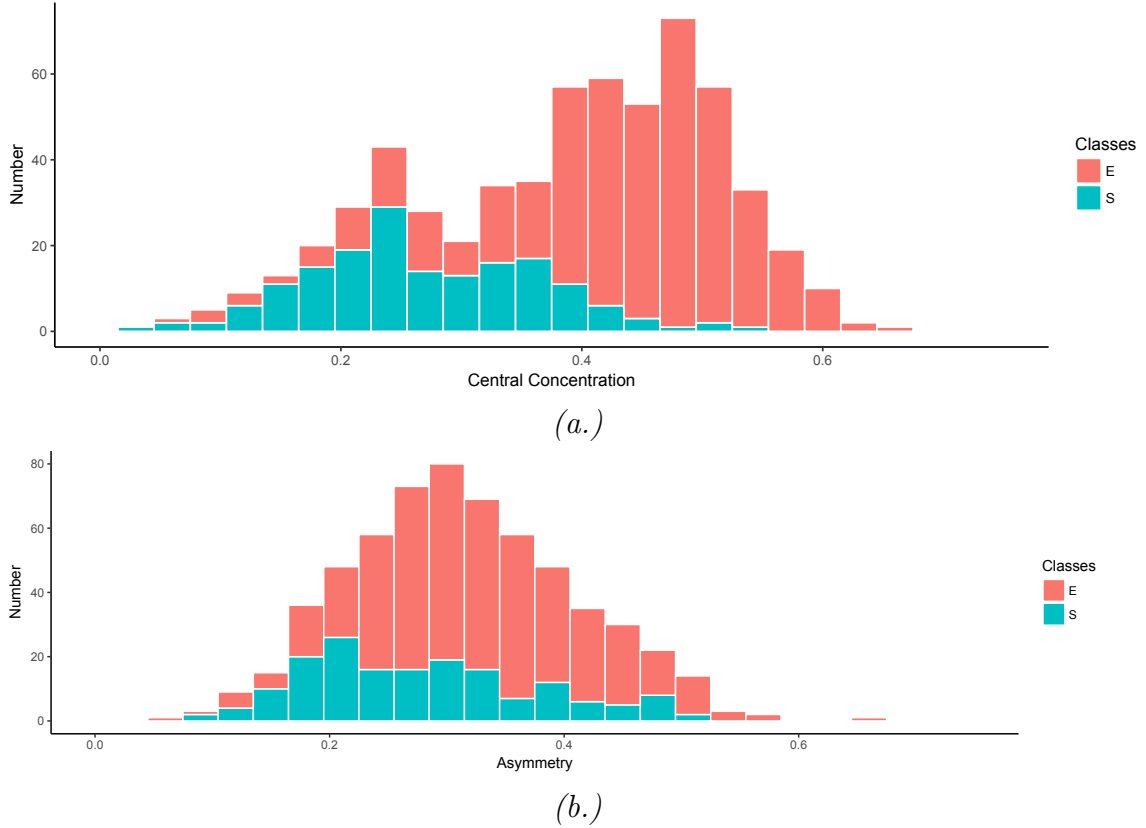


Figure 53: Histogram of two Hubble Types as functions of central concentration (top) and asymmetry (bottom) for the WINGS data. See text for details.

plotted on the A versus Gini plane as shown in Figure 55 (a), where $\text{Gini} \geq 0.45$ is defined as the region containing mainly early-type galaxies, while the region outside of these bounds contains mainly late-type galaxies.

We find the majority of visually classified WINGS galaxies to be plotted in the defined classification regions as shown in Figure 55 (a). Approximately 87% of the galaxies plotted in the “E/S0” region are early-type (410 out of 469). Similarly, about 81% of the galaxies in the “S/Irr” region are late-type systems (110 out of 136).

On the C versus A plane, as depicted in Figure 55 (b), we define the rectangular region bound by $C \geq 0.38$ as the “E/S0” classification region belonging to mainly early-type galaxies. In this “E/S0” region, 337 of 360 (*i.e.* 94%) galaxies are early-

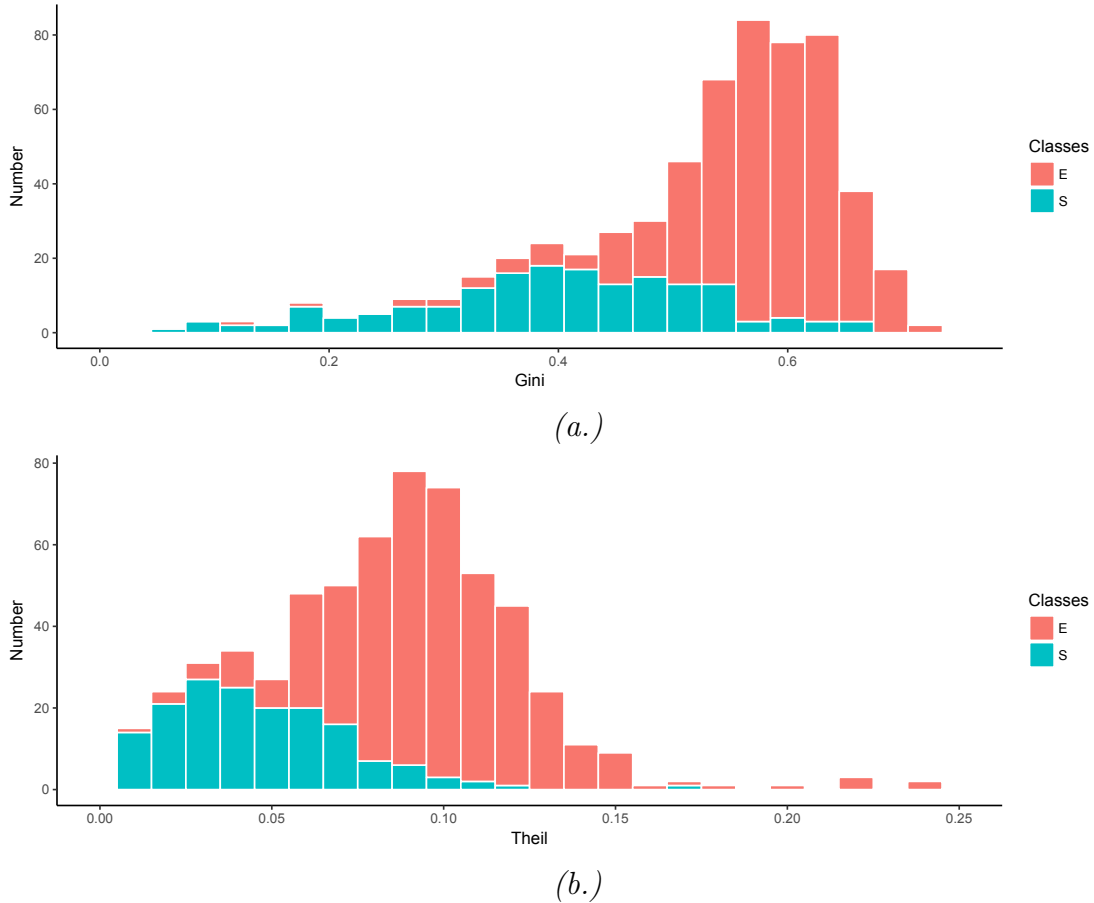


Figure 54: Histogram two Hubble Types as functions of Gini (top) and Theil (bottom) for WINGS data. See text for details.

type, and in the “S/Irr” area, 146 out of 245 (*i.e.* 60%) galaxies are late-type. Unlike our smaller visually classified KPNO sample, the sample of visually classified WINGS galaxies is larger and therefore more reliable.

Figure 56 shows the final result of applying the A versus Gini $\rightarrow A$ versus Theil cuts to the 608 visually classified WINGS galaxies. The method and labels of the graph are described in previous sections: Section 3.2.2 and Section 4.1.1.

Out of the 438 visually classified early-type galaxies and 170 visually classified late-type galaxies, we find that 393 are classified as “E”, 108 as “S”, and 104 as the mixed class “I”. This mixed class “I” is approximately 17% of the sample, again suggesting

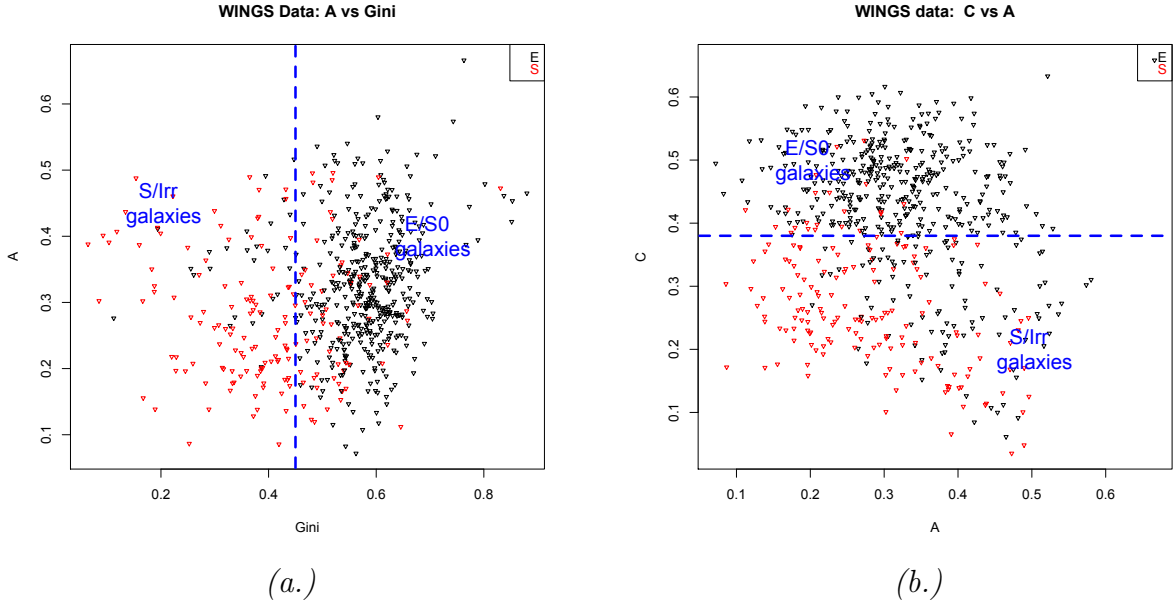


Figure 55: A versus Gini (a) and C versus A (b) plot of approximately 600 WINGS galaxies.

that the application of the two parameter planes to the data may be a reliable method of classification.

Furthermore, as described in Section 4.1.1, we classify galaxies into three categories — bright (B), dim (D), and not classified (N) — by calculating the area of the brightest region in each galaxy (in units of pixels²) and comparing this value to $3 \times \text{FWHM}$ of the cluster image. Out of the 37,357 WINGS galaxies in our study, we find that 15,206 galaxies are classified as bright.

WINGS Data: A vs Theil

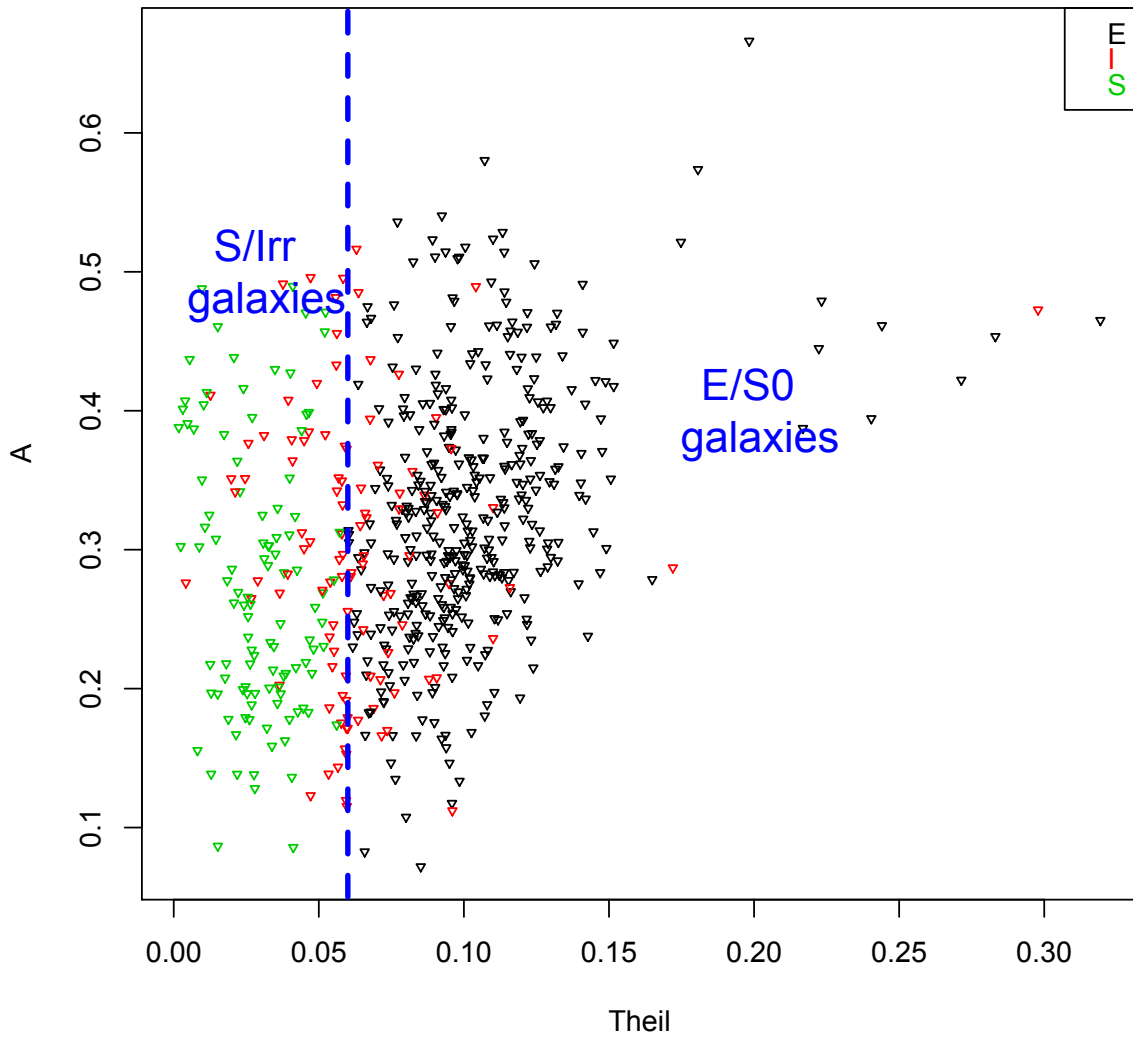


Figure 56: The final result of applying two cuts — the A versus Gini \rightarrow A versus Theil — to 608 galaxies from the WINGS sample.

CHAPTER VI

PRINCIPAL COMPONENT ANALYSIS

6.1 Introduction

Principal component analysis (PCA) is a statistical technique for reducing the dimensionality of data by removing redundancy and noise in the data to identify correlated variables. PCA is used in many branches of science, such as physics, statistics, biology, finance, chemistry, etc. It was first suggested by Pearson (1901) and further developed by Hotelling (1933), who introduced the term “principal component”. In his 1933 paper “Analysis of a complex of statistical variables into principal components”, Hotelling describes how to transform a set of possibly correlated variables into “some more fundamental set of independent variables, perhaps fewer in number than the [original variables], which determine the values the [original variables] will take.”

PCA simplifies multivariable data by finding new linear combinations of the data, which would allow trends, groupings, and outliers in the data to be observed, and allow for better visualization of the data. PCA does not reduce features in the data, meaning, it does not change the original data. It is an orthogonal linear transformation of the original data to a new coordinate system.

Variance of data looks at the spread of the data analyzed. High variance in the data represents areas of greatest “signal” (*i.e.* less noise), meaning, important phenomena in the data can be studied. Additionally, if the variables in the data are highly correlated it means they most likely represent a related phenomena. Correlation indicates redundancy, meaning the variables can be combined to a single measurement.

Redundancy can be used to reduce the original variables into a smaller number of new variables that still explain most of the variance in the original data. PCA transforms the original variables into a new, smaller set of variables, without losing the information contained in the original variables. The new variables are a linear combination of the original and are called the principal components. They are uncorrelated with one another (meaning, they are orthogonal to each other in the original dimension space) and include as much of the original variance in the data as possible.

In this thesis, we perform PCA using the built-in functions in R¹, which is a free software environment for statistical computing and visualization (as well as a programming language), and Minitab 18. In the next two sections, we describe the theory of PCA, as well as the method used to apply this technique in R to the EFIGI data set.

6.1.1 Theory

Suppose we have an $n \times m$ matrix of data, X , with n samples and m measurements. When m is larger than two or three it becomes difficult to visualize the data on a plot. PCA can reduce this dimensionality. In this technique, we wish to perform eigen-decomposition (*i.e.* spectral decomposition) of the square, symmetric $m \times m$ matrix $X^T X$ in order to find its eigenvectors (W) and eigenvalues (λ), where X^T is the transpose of X . These eigenvectors and eigenvalues can be then used to describe the data X by finding the following: $T = XW$, where T is an $n \times m$ matrix whose values are called the “scores”, and the eigenvector columns of the $m \times m$ matrix, W , contain the “loadings”. Each column of W is a principal component.

The steps of PCA are as follows:

1. **Scale the values in the original data set.** Since one of the goals of PCA

¹<https://www.r-project.org>

is to capture the total variance in a set of variables, the variables need to have similar scales of measurement. Scaling is achieved by dividing each variable by its standard deviation. It prevents features with large numeric range from dominating over other features in the data set.

- 2. Calculate the correlation matrix or the covariance matrix between every pair of variables in the centered and scaled data.** Correlation coefficient is used to measure the similarity between two dimensional data points x and y . A correlation matrix is a table of correlation coefficients between sets of dimensions (*i.e.* variables) in a multivariable data set. We use the Pearson definition of the correlation coefficient, which measures the strength and direction of the linear relationship between two variables. It is defined as:

$$\text{corr}(x, y) = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sigma(x)\sigma(y)}, \quad (6.1)$$

where the sum is over all data points in data sample of size N ; \bar{x} and \bar{y} are the means of x and y , respectively; and $\sigma(x)$ and $\sigma(y)$ are the standard deviations of x and y , respectively. Standard deviation is defined as:

$$\sigma(x) = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N - 1}}. \quad (6.2)$$

The correlation matrix is used when variables in a data set have different scales (*i.e.* the correlation matrix standardizes the data), and a covariance matrix is used when the variable scales are similar or if we do not wish to scale the data. By their definition, using either the correlation matrix or the covariance matrix centers the data by subtracting the mean from each variable. This produces a

data set whose mean is zero.

As was previously mentioned, variance is the measure of the deviation from the mean. It is defined as:

$$\text{var}(x) = \frac{\sum_{i=1}^N (x_i - \bar{x})(x_i - \bar{x})}{N - 1} = \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N - 1}, \quad (6.3)$$

where \bar{x} is the mean of x and N is the size of the sample. The mean is the average of all data values in the set X . It is found by dividing the sum of all data by the number of data points, N . Variance can also be defined as the square of the standard deviation.

Covariance is the measure of how much each of the dimensions vary from the mean with respect to each other in order to see if there is a relationship between the two dimensions studied. The covariance between one dimension and itself is the variance. The covariance between two variables, x and y , is as follows:

$$\text{cov}(x, y) = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{N - 1}, \quad (6.4)$$

where \bar{x} and \bar{y} are the means of x and y , respectively, and N is the size of the sample. For example, for a three-dimensional data set (x, y, z) , the covariance between each of the dimensions — x and y dimensions, y and z dimensions, and

x and z dimensions — is given by the covariance matrix, as follows:

$$C = \begin{pmatrix} \text{cov}(x, x) & \text{cov}(x, y) & \text{cov}(x, z) \\ \text{cov}(y, x) & \text{cov}(y, y) & \text{cov}(y, z) \\ \text{cov}(z, x) & \text{cov}(z, y) & \text{cov}(z, z) \end{pmatrix}.$$

The covariance between x and x, or y and y, or z and z is the variance of the x, y, and z dimensions, respectively, meaning $\text{cov}(x,x) = \text{var}(x)$, $\text{cov}(y,y) = \text{var}(y)$, and $\text{cov}(z,z) = \text{var}(z)$. Covariance is symmetric under the interchange of its arguments, *i.e.* $\text{cov}(x,y) = \text{cov}(y,x)$, therefore the covariance matrix is symmetric, *i.e.* $C = C^T$. A positive value of covariance indicates the dimensions increase or decrease together; negative value of covariance indicates that if one dimension increases, another decreases; and a covariance of zero indicates that the dimensions are independent of each other.

Another way to write the covariance matrix of an $m \times n$ matrix, X is:

$$C = \frac{1}{N-1} X X^T. \tag{6.5}$$

3. Compute the eigenvectors and eigenvalues of the covariance matrix.

The main idea of PCA is that the principal components are the eigenvectors of the covariance matrix of the data set. The eigenvectors with the largest eigenvalues correspond to the dimensions that have the strongest variation in the data set. A set of eigenvalues of a matrix can also be called the spectrum of that matrix. Any symmetric matrix has a spectrum decomposition, which can be stated as follows:

Theorem VI.1. : *(The Spectral Theorem) An $m \times m$ symmetric matrix A (meaning $A^T = A$) has the following properties: A has i real eigenvalues; A is orthogonally diagonalizable; the eigenvectors corresponding to the different eigenvalues are orthogonal (Lay 2012).*

Suppose $A = PDP^T$, where D is diagonal and P is orthogonal, meaning $P^{-1} = P^T$ or $P^T P = P P^T = 1$. The columns of P are orthonormal eigenvectors (v_1, v_2, \dots, v_i) and the corresponding eigenvalues ($\lambda_1, \lambda_2, \dots, \lambda_i$) are in the diagonal matrix D . Each eigenvalue has an associated eigenvector, meaning, the i^{th} column vector in P is the eigenvector associated with the eigenvalue λ_i in D .

Multiplying $A = PDP^T$ on the right by P gives: $AP = PDP^T P = PD$. Since the columns of P are the eigenvectors and D contains the eigenvalues, then this can be written as: $Av_i = \lambda_i v_i$.

We can solve for the eigenvalues of the covariance matrix (C) by taking the determinant (det): $\det(C - \lambda I) = 0$, where I is the identity matrix. We can also use Singular Value Decomposition (SVD) to compute the eigenvalues and eigenvectors.

- 4. Sort the eigenvectors by their eigenvalues from the highest eigenvalue being the first to the lowest being the last.** The number of chosen eigenvectors will be the number of dimensions in the new data set. Each eigenvector is a principal component, meaning, the numeric values in each eigenvector are the coefficients of each principal component. The new principal components define a new coordinate system.

A graphical example of PCA can be seen in Figure 57, where the data is represented in the X-Y coordinate system. The arrows show the directions of the new coordinate axes (principal components: PC1 and PC2). In PCA, the arrows (*i.e.*

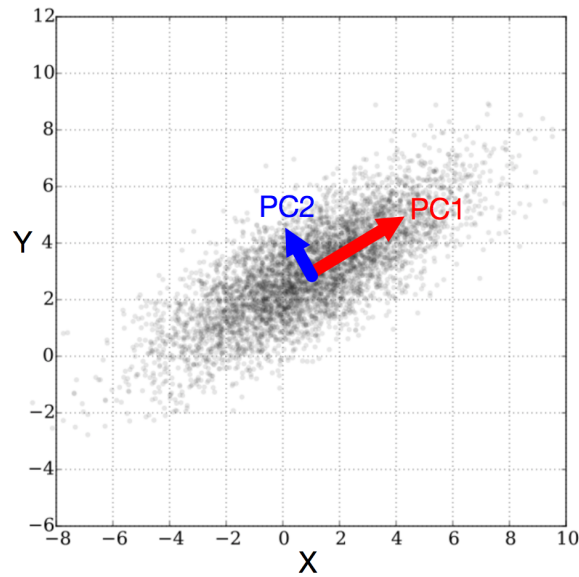


Figure 57: Graphical representation of PCA. The vectors are the eigenvectors (*i.e.* the principal components, PC1 and PC2) of the scaled covariance matrix of the data. Image Source: Nicoguardo, modified by M. Sultanova.

directions) with the largest variance are the most important in describing the data, in which case, PC1 shows the largest variation in Figure 57 and is therefore the first principal component. PC2 would be the second most important direction, *i.e.* the second principal component. PC2 is orthogonal to PC1. The eigenvectors found in PCA lie in the directions of the principal axes, while the eigenvalues give the relative length of the corresponding principal axes (Janert 2010). The two-dimensional data in Figure 57 can be reduced to one-dimension by projecting it onto the first principal component, PC1.

6.1.2 When to use PCA

PCA is “just a mathematical transformation that can be applied to any symmetric matrix” (Janert 2010). There are other techniques one could use to study a set of data, such as Factor Analysis (FA), Independent component analysis, Non-negative matrix

factorization, etc. PCA is a popular technique for studying data since it is rigorous and “transforms the original data in a precisely prescribed way, without ambiguity and without making further assumptions” (Jarnet 2010), and with appropriate tools it can be easy to perform. PCA is different from other data analysis methods. Independent component analysis (ICA), for example, is a method for separating a multivariate signal into additive subcomponents that are independent. It is typically not used for reducing dimensionality but for separating superimposed signals. Non-negative matrix factorization (NMF) is a dimension reduction method which, unlike PCA, uses only non-negative elements in the matrices.

While the main goal of PCA is to find new combinations of variables which describe a set of data, one of the main goals of FA is to find latent variables which affect a set of data. These hidden variables cannot be measured directly, but instead are studied through the way they influence the original variables in a set of data. Both PCA and FA are methods of data reduction. Deciding which method to use to study one’s data depends on the goal of one’s project. PCA is a better method to reduce the dimensionality of a data set, while FA would be more appropriate to use in order to hypothesize latent variables in the data.

6.1.3 PCA using R for the EFIGI

In this thesis, we apply PCA to the data sets using Minitab 18 and the build-in PCA functions in R. There are two methods of performing PCA in R: through spectral decomposition or through SVD. Both approaches are similar, however, whereas SVD is a more general technique that can be applied to any $m \times n$ matrix (rectangular or square), spectral decomposition can only be applied to diagonalizable (square) matrices. The build-in function `princomp()` uses the spectral decomposition method, while functions, such as `prcomp()`, use SVD. In this thesis, we use the build-in function

`prcomp()`, therefore we will focus on describing this method.

The format for this function is: `prcomp(x, scale=TRUE)`, where the arguments are:

- `x`: the data matrix.
- `scale`: logical value; indicates whether the variables should be scaled before the analysis takes place.

The data matrix needs to be prepared before applying the `prcomp()` function. The rows (n) are the observations and the columns (m) are the variables. In order to perform PCA on the data, the rows of data studied should only contains numeric values. We omit any data cells with missing values by using `na.omit`. The outputs of the `prcomp()` function are:

- `sdev`: the standard deviation of the principal components.
- `rotation`: the matrix of the variable loadings, where the columns are eigenvectors.
- `scale`: the variable standard deviations.
- `center`: the variable means that were subtracted.
- `x`: the scores, *i.e.* the coordinates of the data on the principal components.

We focus on five numeric variables (meaning, five dimensions) in this thesis: central concentration, asymmetry, Gini coefficient, Theil index, and M20. To perform PCA on the EFIGI data set in R, we start with the following (Coghlan 2014):

```
>data = read.csv("EFIGI_data.csv")
>datapca <- prcomp(na.omit(data[, c("C", "A", "Gini", "Theil", "M20")] ),
                  scale=TRUE)
```

where the results are stored in `datapca`. We can examine the results afterwards, for example, by viewing the “scores”. This is a sample of eleven rows from the full list:

```
> datapca$x
      PC1          PC2          PC3          PC4          PC5
1  0.568598925 -1.1575876532  0.3074781987  0.1875576249  1.146566e-01
2  0.742526700  0.9274673325 -0.9396301441 -0.4415587000  1.625116e-01
3  2.347170294 -0.4243466606  0.4512366516  0.4151235320 -2.197608e-01
4 -1.570650130  7.0648914932  7.9564144195 -0.6183151895  1.517048e+00
5  2.927946845 -0.2717919021  0.5858836828  0.5651658858 -4.569879e-01
6  2.251576937  0.1545386361 -0.0411144847  0.1392082171 -2.298173e-01
7  1.871746647  0.7072904199 -0.4454131435 -0.2933527761 -2.230390e-01
8  2.460309361  0.3702110004  0.0393612186 -0.1694993185 -5.490040e-01
9  0.426230286 -0.4718104839 -0.1794252725  0.0675522653  2.148525e-01
10 2.683782027 -0.6282951969  0.6204321746  0.7009945818 -2.464509e-01
11 2.048019141  0.3631917384  0.2948469984 -0.0270897340 -2.807761e-01
.....
```

We can also view the “loadings” of our final matrix through the specification below.

```
> datapca$rotation
      PC1          PC2          PC3          PC4          PC5
C  -0.4455053 -0.48555108  0.21781984 -0.68181261 -0.2311953
A  -0.2679412  0.66842656 -0.52099757 -0.45725528 -0.0298760
Gini -0.5812988 -0.20524104 -0.19541733  0.21545419  0.7316825
Theil -0.5788162  0.05559264 -0.05945288  0.52841115 -0.6157340
M20  -0.2383192  0.52175824  0.79962083 -0.02025652  0.1765460
```

As explained in Phan (2016), the procedure for PCA can be demonstrated in Figure 58 for the EFIGI data. The original data matrix containing the measurements of the five parameters for the 4,352 EFIGI galaxies is scaled, and a sample of seventeen rows is displayed in the table in Figure 58 (a). After performing PCA via `prcomp()` on R, we can see the “loadings” in the table in Figure 58 (b). The columns are the principal component vectors. Since we have five dimensions, PCA will produce five principal components with five floating coefficient values. For example, the first principal component is a linear combination of the variables: $-0.44*Z_1 - 0.26*Z_2 -$

| Number | C | A | Gini | Theil | M20 |
|--------|-------------|-----------|-----------|------------|------------|
| 1 | 0.22134174 | -1.18E+00 | -2.70E-02 | -0.3824689 | -0.4153694 |
| 2 | -0.72231407 | 1.11E+00 | -4.12E-01 | -0.6540195 | -0.3574426 |
| 3 | -0.97357533 | -1.33E+00 | -1.43E+00 | -1.0511897 | -0.4071362 |
| 4 | -0.92672703 | 1.24E+00 | -1.11E+00 | -0.431039 | 8.78092874 |
| 5 | -1.32451368 | -1.52E+00 | -1.97E+00 | -1.1611346 | -0.4039017 |
| 6 | -1.12886265 | -5.35E-01 | -1.47E+00 | -1.0739293 | -0.4606523 |
| 7 | -1.02274131 | 3.45E-01 | -1.37E+00 | -1.0322033 | -0.4066951 |
| 8 | -1.02476964 | -3.38E-01 | -1.95E+00 | -1.1538491 | -0.3972857 |
| 9 | -0.09561302 | -3.73E-01 | 5.75E-02 | -0.3581839 | -0.3968446 |
| 10 | -1.17639593 | -1.78E+00 | -1.58E+00 | -1.0997597 | -0.4580059 |
| 11 | -0.94114377 | -4.39E-01 | -1.53E+00 | -1.0211646 | -0.1148557 |
| 12 | -1.3233769 | -7.05E-01 | -1.91E+00 | -1.1511998 | -0.3872881 |
| 13 | -0.77441623 | -1.35E+00 | -1.20E+00 | -0.9586859 | -0.3712627 |
| 14 | -1.02901544 | -9.70E-01 | -1.59E+00 | -1.0750332 | -0.3690574 |
| 15 | -1.12789658 | -1.22E+00 | -1.68E+00 | -1.1057206 | -0.3693514 |
| 16 | 0.36957827 | -1.40E+00 | 2.10E-01 | -0.2235125 | -0.3530319 |
| 17 | -0.80906978 | -1.07E+00 | -1.16E+00 | -0.9454396 | -0.3515617 |

(a.)

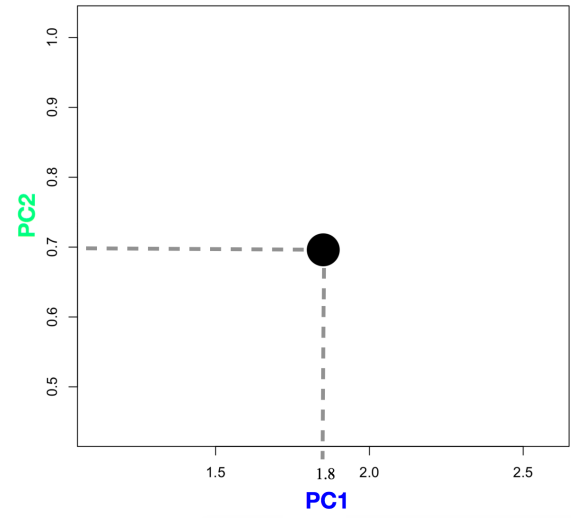
| Variable | PC1 | PC2 | PC3 | PC4 | PC5 |
|----------|------------|------------|------------|------------|------------|
| C | -0.4455053 | -0.4855511 | 0.21781984 | -0.6818126 | -0.2311953 |
| A | -0.2679412 | 0.66842656 | -0.5209976 | -0.4572553 | -0.029876 |
| Gini | -0.5812988 | -0.205241 | -0.1954173 | 0.21545419 | 0.7316825 |
| Theil | -0.5788162 | 0.05559264 | -0.0594529 | 0.52841115 | -0.615734 |
| M20 | -0.2383192 | 0.52175824 | 0.79962083 | -0.0202565 | 0.176546 |

(b.)

$$PC1: -1.0 * -0.44 + 0.34 * -0.26 + -1.37 * -0.58 + -1.0 * -0.57 + -0.41 * -0.23 = 1.8$$

$$PC2: -1.0 * -0.49 + 0.34 * 0.66 + -1.37 * -0.21 + -1.0 * -0.06 + -0.41 * -0.52 = 0.7$$

(c.)



(d.)

Figure 58: The mechanism of performing principal component analysis. See text for details.

$0.58*Z_3 - 0.57 *Z_4 - 0.23*Z_5$, where Z_1, Z_2, Z_3, Z_4 , and Z_5 are the scaled, original data values.

Each principal component is a newly defined axis, and the data can be projected onto these new axis by computing the dot product between the original scaled data and the coefficients in the principal components, which are shown in Figure 58 (c). The red values are the original scaled data and the blue and green represent the coefficients in PC1 and PC2, respectively, which result in scalar values of 1.8 on the PC1 axis and 0.7 on the PC2 axis. The principal components are orthogonal to each other. We can see the projection of the data onto the new axis in Figure 58 (d). A summary of the principal components using the `summary()` function on the output is:

```
> summary(datapca)
Importance of components:
      PC1      PC2      PC3      PC4      PC5
Standard deviation  1.6116 1.0613 0.9226 0.61553 0.21555
Proportion of Variance 0.5194 0.2253 0.1703 0.07577 0.00929
Cumulative Proportion 0.5194 0.7447 0.9149 0.99071 1.00000
```

From here we see that the first three components (PC1, PC2, and PC3) bring our cumulative proportion of variance to 0.91, meaning, they describe about 91% of the variance in the data.

The standard deviation of each component is:

```
> datapca$sdev
[1] 1.6115573 1.0612822 0.9226210 0.6155263 0.2155488
```

6.1.4 How Many Principal Components to Keep?

The standard deviation and summary of the PCA results can tell us how many principal components best describe our data. However, a better way to see how many

principal components should be kept and the other's discarded is through a scree plot:

```
> screeplot(datapca, type="lines")
```

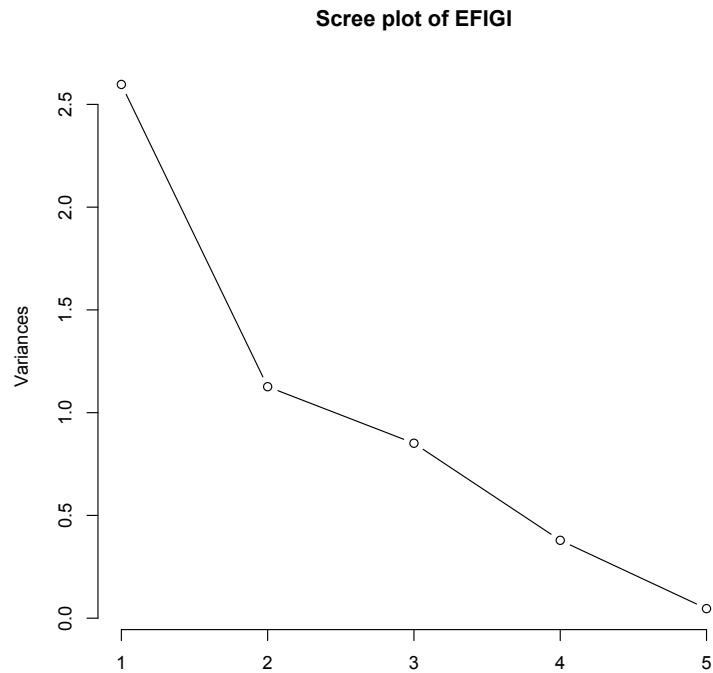


Figure 59: Scree plot of principal components of the EFIGI data.

This command displays the plot in Figure 59. The scree plot is a way to visualize the eigenvalues. It orders the magnitude of variances explained by each principal component from largest to smallest. Since the eigenvector with the largest eigenvalue is the direction along which the data has the maximum variance, the eigenvalues of the correlation matrix essentially equal the variances of the principal components. The first principal component clearly dominates over others in the scree plot in Figure 59. At times, a scree plot may have an obvious change of slope, which looks like an “elbow” on the plot. Based on this, it could be argued that from the scree plot in Figure 59, the first two components should be kept and the rest discarded. However,

other times the scree plot may not possess an obvious change of slope, therefore in this thesis we will decide how many components to retain using Kaiser's criterion (Kaiser 1960), which states that only the principal components that have a variance greater than one should be retained.

We can find the variances displayed on the scree plot:

```
> variances <- datapca$sdev^2
> variances
[1] 2.59711688 1.12631983 0.85122943 0.37887258 0.04646128
```

We see that the variance is greater than one for principal components PC1 and PC2 (which have variances of 2.597 and 1.126, respectively). Therefore, using Kaiser's criterion, we would retain the first two principal components for the EFIGI data. Using the Quality Control Charts (qcc) package on R (Scrucca 2004), we also create a pareto chart for these data. A pareto chart shows the frequency of each principal component:

```
> library (qcc)
> variances <- datapca$sdev^2
> pareto.chart (variances, ylab="Variances")
```

```
Pareto chart analysis for variances
      Frequency    Cum.Freq.  Percentage Cum.Percent.
A    2.59711688    2.59711688   51.94233765   51.94233765
B    1.12631983    3.72343671   22.52639664   74.46873429
C    0.85122943    4.57466615   17.02458867   91.49332296
D    0.37887258    4.95353872    7.57745153   99.07077449
E    0.04646128    5.00000000    0.92922551  100.00000000
```

The pareto chart is shown in Figure 60. The bars represent the principal components, which are ordered by decreasing magnitude based on their variance. The line above the bars represents the cumulative total. The left vertical axis is the frequency of occurrence, *i.e.* the variation accounted for by each principal component. The right

vertical axis is the cumulative frequency expressed as a percentage. In Figure 60, it can be seen that the first two principal components account for almost 75% of the variability in the data.

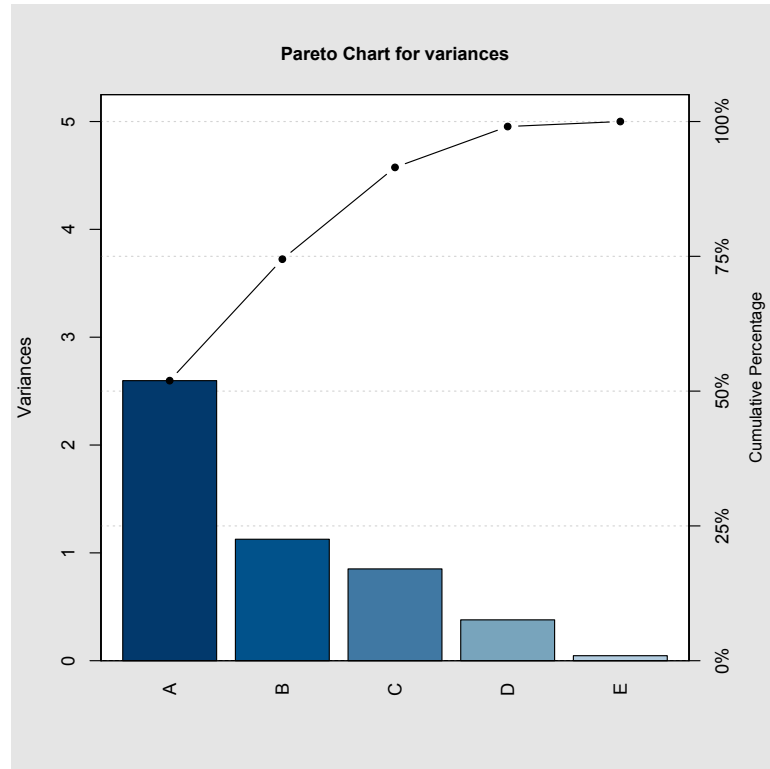


Figure 60: Pareto chart of principal components of the EFIGI data.

Since we have now reduced our five dimensional data to two dimensions, we can plot the two principal components that represent these dimensions in a biplot:

```
> biplot(datapca, xlabs=rep("x", nrow(na.omit(data[, c("C", "A",
  "Gini", "Theil", "M20")]))), choices=c(1,2), scale=0)
```

which creates the biplot seen in Figure 61. A biplot is a generalized scatterplot. It uses both points and vectors to display information. However, the ‘bi’ in biplot refers to the fact that two sets of points (*i.e.*, the rows and columns of the target matrix) are visualized by scalar products. It does not stand for the fact that the plot is two-

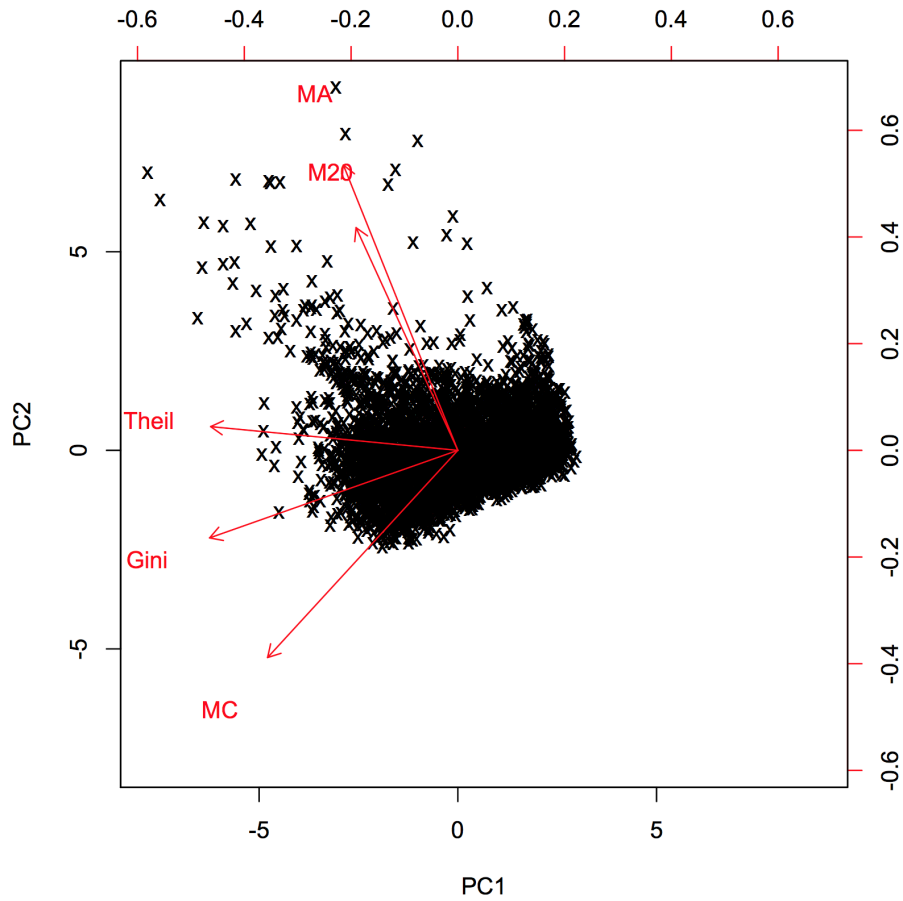


Figure 61: Biplot diagram of the translated data onto the first two principal components. EFIGI data is analyzed. See text for details.

dimensional (Greenacre 2012). The left vertical and the bottom horizontal axis on the biplot are the “scores” of the two principal components. In Figure 61, we plot the two principal components PC1 and PC2, which we found to be the two that best describe the data. The other principal components are discarded. The right vertical and the top horizontal axis represent the “loadings”. In Figure 61, MC is used to represent central concentration and MA is asymmetry.

The points on the biplot are the “scores” of observations on the principal components, and the vectors are the “loadings”. The length of each eigenvector indicates

how strongly the data is being pulled in that direction, *i.e.*, how strongly each variable contributes to the principal component (Greenacre 2010).

Points on the biplot that are close together represent observations that have similar scores. The vectors are cast in the direction of the principal component they are greatly correlated with. Vectors that point in the same direction can be interpreted as similar to each other.

We can interpret the biplot in Figure 61 from the loadings of the principal components PC1 and PC2, as follows:

```
> datapca$rotation[,1:2]
      PC1      PC2
C    -0.4455053 -0.48555108
A    -0.2679412  0.66842656
Gini -0.5812988 -0.20524104
Theil -0.5788162  0.05559264
M20  -0.2383192  0.52175824
```

From the values as well as from the biplot, it can be seen that the Gini coefficient exerts the greatest weight on PC1, followed by the Theil index. The prominent parameter in PC2 is the asymmetry, followed by M20. In PC2, asymmetry, M20, and the Theil index have an inverse relationship with the central coefficient and Gini, meaning when asymmetry, M20, and the Theil index increase, the central concentration and the Gini coefficient decrease.

We can see that the Theil index, Gini coefficient, and central concentration are in close proximity to one another on the biplot, which indicates a correlation between the parameters, and likewise for asymmetry and M20.

6.2 PCA for CFHT Data Set

We follow the procedure described in Sections 6.1.3 and 6.1.4 to perform PCA for the CFHT data. We analyze the 5,361 galaxies we calculated and defined as bright (B) from measurements of the FWHM for stars in each cluster, described earlier. The summary of the results and the standard deviation are as follows:

```
> summary(datapca)
Importance of components:
              PC1    PC2    PC3    PC4    PC5
Standard deviation  1.4878 1.2976 0.8766 0.54441 0.19527
Proportion of Variance 0.4427 0.3367 0.1537 0.05928 0.00763
Cumulative Proportion 0.4427 0.7794 0.9331 0.99237 1.00000

> datapca$sdev
[1] 1.4877647 1.2975515 0.8765887 0.5444064 0.1952693
```

We also display the scree plot and pareto chart in Figures 62 (a) and (b). The results from the pareto chart are:

```
> library (qcc)
> variances <- datapca$sdev^2
> pareto.chart (variances, ylab="Variances")

Pareto chart analysis for variances
      Frequency    Cum.Freq.  Percentage Cum.Percent.
A  2.21344386    2.21344386  44.26887717  44.26887717
B  1.68364000    3.89708386  33.67280006  77.94167722
C  0.76840770    4.66549156  15.36815402  93.30983124
D  0.29637834    4.96186991   5.92756687  99.23739811
E  0.03813009    5.00000000   0.76260189 100.00000000
```

From the frequency (*i.e.* the variances of each principal component), it can be seen that the first two principal components are greater than one. Following Kaiser's criteria, we focus on these two components, and discard the others. From the results

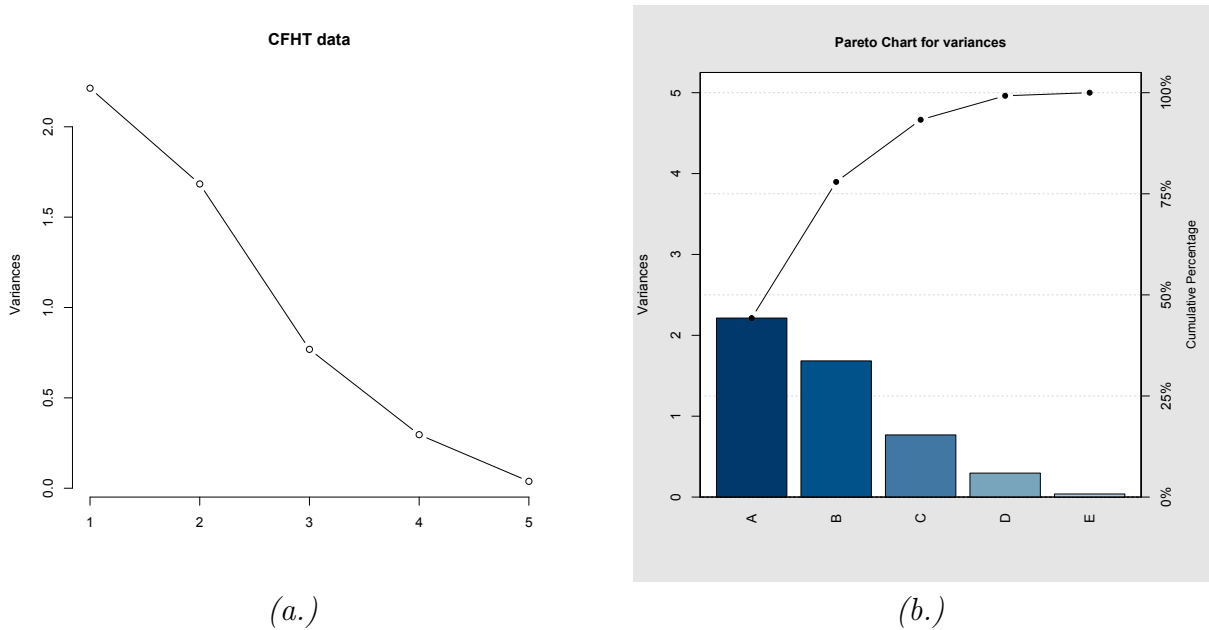


Figure 62: A scree plot (a) and pareto chart (b) of the principal components of the CFHT data.

of the pareto chart, we find that the first two components also describe almost 78% of the data.

The loadings are as follows:

```
> datapca$rotation
      PC1      PC2      PC3      PC4      PC5
C    -0.19539075 -0.59357799 -0.56004290  0.52063570  0.1574104
A    -0.32496131  0.59547264  0.17224184  0.70130594  0.1353337
Gini -0.65849207 -0.09655839  0.04023861 -0.09025724 -0.7397967
Theil -0.64600002  0.03826289 -0.04926564 -0.43955174  0.6209566
M20   0.07276513  0.53130964 -0.80786041 -0.18909347 -0.1549855
```

Similarly to the results with the EFIGI data, we find that the Gini coefficient exhibits a stronger emphasis on PC1, followed by the Theil index. The prominent parameter in PC2 is the asymmetry, followed by M20. Just as with the EFIGI data, asymmetry, M20, and the Theil index have an inverse relationship with the central coefficient and Gini in PC2. Central concentration is nearly equal in weight to asymmetry but in the opposite direction.

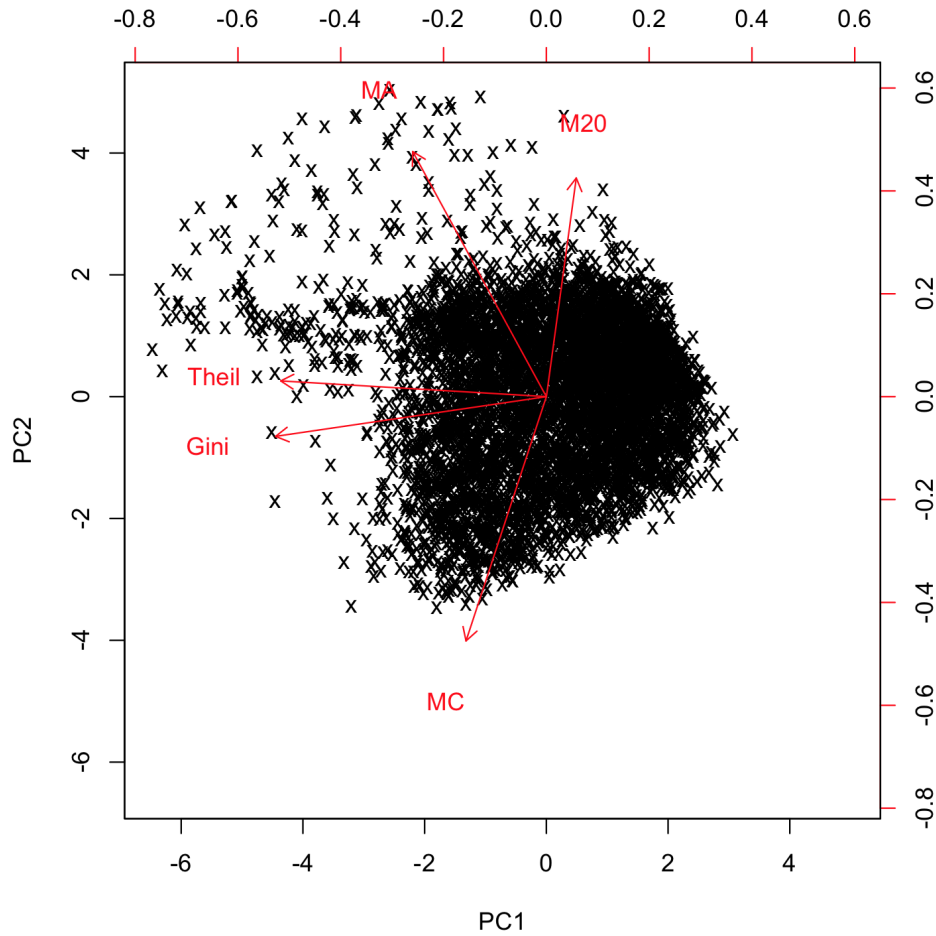


Figure 63: Biplot diagram of the translated data onto the first two principal components of the bright galaxies in the CFHT data. See text for details.

We also find that the Theil index, Gini coefficient, and central concentration are in close proximity to one another on the biplot in Figure 63, which again indicates a correlation between the parameters, and likewise for asymmetry and M20.

6.3 PCA for KPNO Data Set

Similar to Section 6.2, we follow the procedures described in Sections 6.1.3 and 6.1.4 to perform PCA for the 4200 bright galaxies (we define as B) in the KPNO data. A summary of the results and the standard deviation are as follows:

```

> summary(datapca)
Importance of components:
              PC1    PC2    PC3    PC4    PC5
Standard deviation  1.664 1.0302 0.9370 0.50103 0.19949
Proportion of Variance 0.554 0.2122 0.1756 0.05021 0.00796
Cumulative Proportion 0.554 0.7662 0.9418 0.99204 1.00000

> datapca$sdev
[1] 1.6643176 1.0301648 0.9370070 0.5010293 0.1994869

```

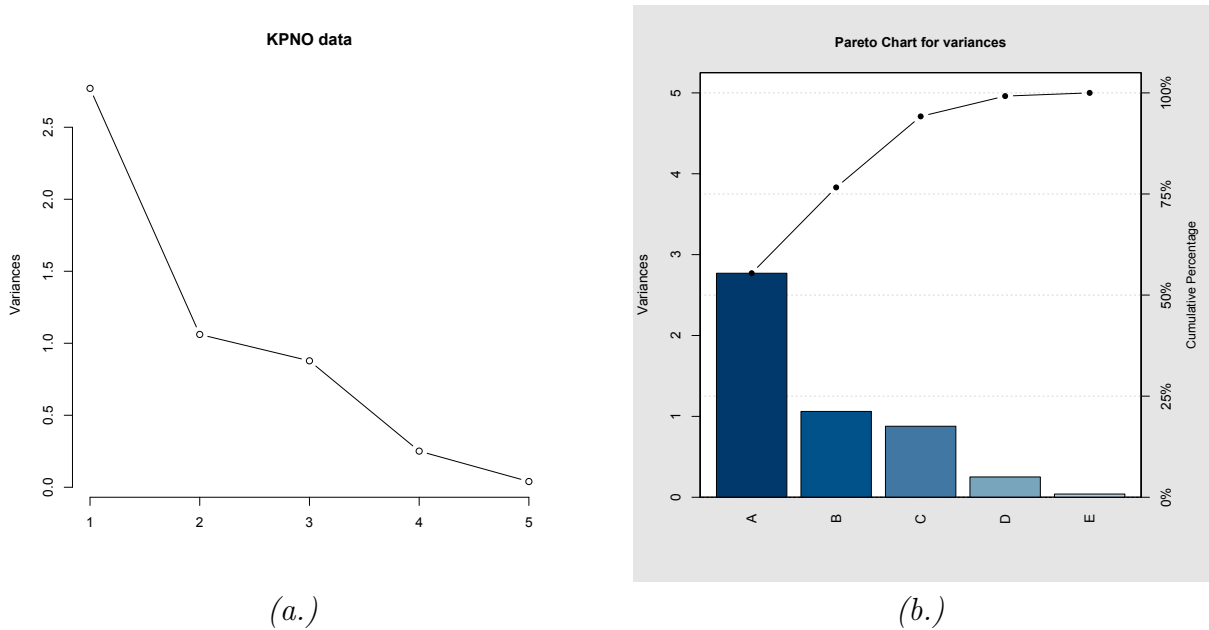


Figure 64: A scree plot (a) and pareto chart (b) of the principal components of the KPNO data.

We also display the scree plot and pareto chart in Figures 64 (a) and (b). The results of the pareto chart are:

```

> library (qcc)
> variances <- datapca$sdev^2
> pareto.chart (variances, ylab="Variances")

```

```

Pareto chart analysis for variances
      Frequency  Cum.Freq.  Percentage Cum.Percent.
A  2.76995319  2.76995319  55.39906382  55.39906382
B  1.06123943  3.83119262  21.22478864  76.62385246

```

| | | | | |
|---|------------|------------|-------------|--------------|
| C | 0.87798203 | 4.70917465 | 17.55964058 | 94.18349304 |
| D | 0.25103034 | 4.96020499 | 5.02060680 | 99.20409985 |
| E | 0.03979501 | 5.00000000 | 0.79590015 | 100.00000000 |

From the pareto chart analysis for variances, it can be seen that the first two principal components are greater than one, therefore we focus on these two components and discard the others as per Kaiser’s criteria. From the results of the pareto chart, we find that the first two components also describe almost 77% of the data.

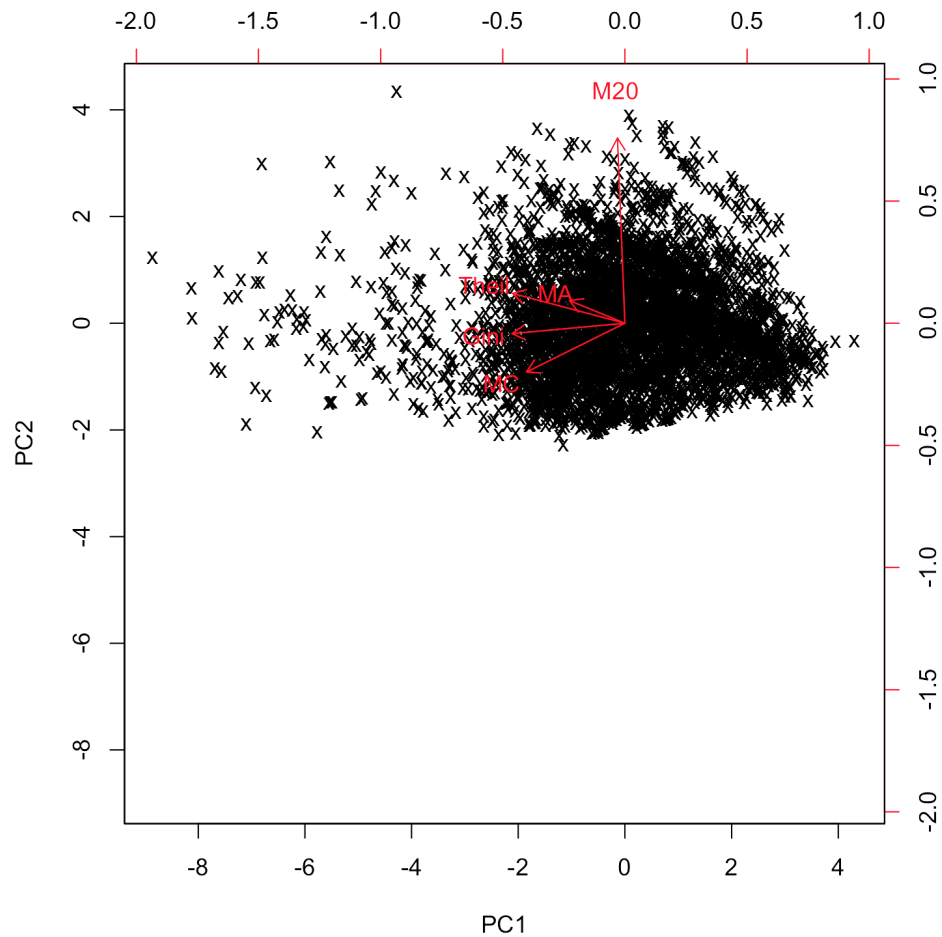


Figure 65: Biplot plot of the translated data onto the first two principal components of the bright galaxies in the KPNO data. See text for details.

The loadings are as follows:

```
> datapca$rotation
      PC1      PC2      PC3      PC4      PC5
C    -0.50403782 -0.24978873 -0.30550042  0.7681373 -0.01364079
A    -0.28323574  0.11555585  0.92506854  0.2204467  0.04556096
Gini -0.57744096 -0.05313616 -0.10277934 -0.4248535  0.68751321
Theil -0.57518590  0.15266847 -0.07223855 -0.3691230 -0.71019947
M20  -0.03803307  0.94769263 -0.18744516  0.2112251  0.14380689
```

The trends seen in these results and Figure 65 are similar to what was found for the EFIGI and CFHT data. The Gini coefficient exhibits a stronger emphasis on PC1, followed closely by the Theil index and the central concentration. M20 dominates the second principal component, followed by the Theil index and asymmetry parameter. However, just as with the EFIGI and CFHT data, asymmetry, M20, and the Theil index have an inverse relationship with the central coefficient and Gini in PC2.

6.4 PCA for WINGS Data Set

We once more follow the procedures described earlier to perform PCA for the 15,206 bright galaxies in the WINGS data. The summary of the results and the standard deviation are as follows:

```
> summary(datapca)
Importance of components:
      PC1      PC2      PC3      PC4      PC5
Standard deviation    1.5271  1.2972  0.7894  0.57125  0.18950
Proportion of Variance 0.4664  0.3365  0.1246  0.06527  0.00718
Cumulative Proportion 0.4664  0.8029  0.9275  0.99282  1.00000

> datapca$sdev
[1] 1.5270649 1.2971901 0.7893877 0.5712518 0.1894970
```

We also display the scree plot and pareto chart in Figures 66 (a) and (b).

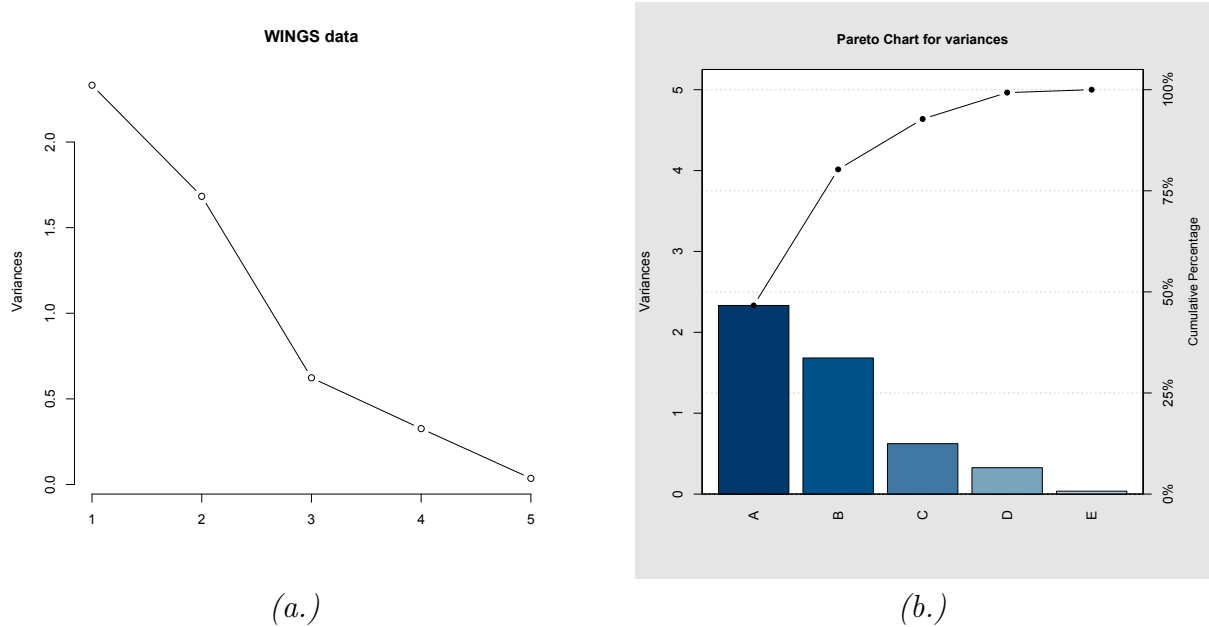


Figure 66: A scree plot (a) and pareto chart (b) of the principal components of the WINGS data.

The results of the pareto chart are:

```
> library (qcc)
> variances <- datapca$sdev^2
> pareto.chart (variances, ylab="Variances")
```

```
Pareto chart analysis for variances
      Frequency  Cum.Freq.  Percentage  Cum.Percent.
A    2.3319273  2.3319273  46.6385464  46.6385464
B    1.6827021  4.0146294  33.6540414  80.2925878
C    0.6231329  4.6377623  12.4626586  92.7552464
D    0.3263286  4.9640909   6.5265717  99.2818181
E    0.0359091  5.0000000   0.7181819 100.0000000
```

From the pareto chart analysis for variances, it can be seen that the first two principal components are greater than one, therefore we focus on these two components and discard the others as per Kaiser's criteria. From the results of the pareto chart, we find that the first two components also describe 80% of the data.

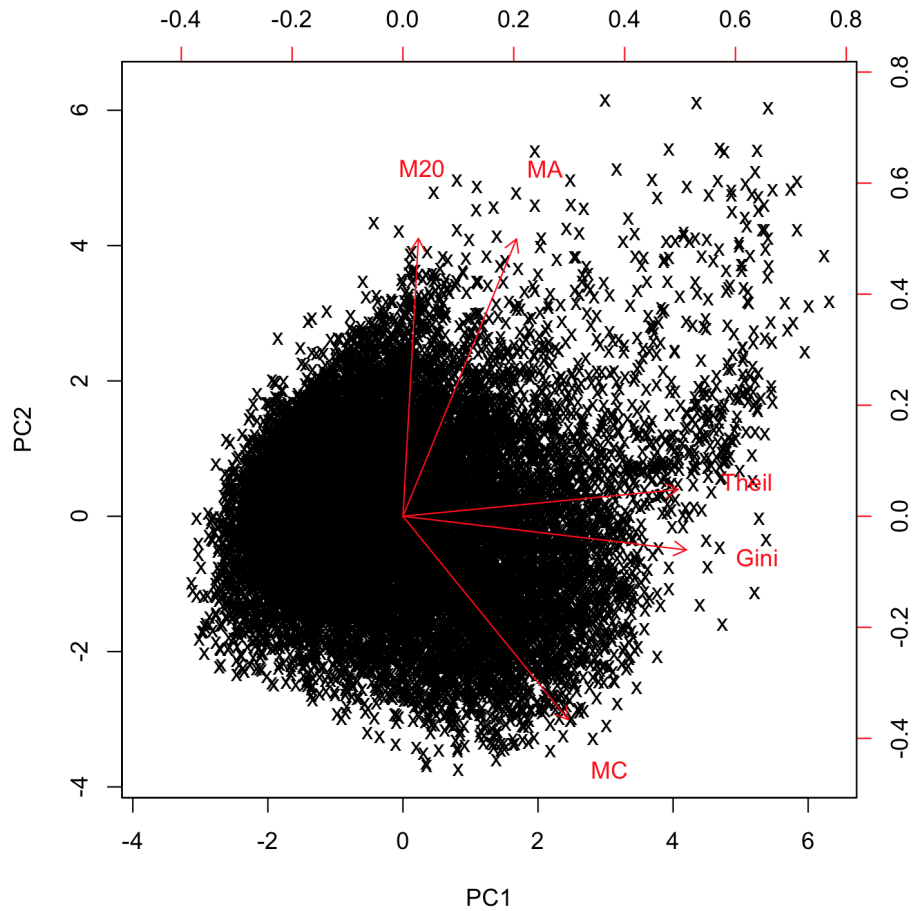


Figure 67: Biplot diagram of the translated data onto the first two principal components of the bright galaxies in the WINGS data. See text for details.

The loadings are as follows:

```
> datapca$rotation
      PC1      PC2      PC3      PC4      PC5
C      0.37268143 -0.45788794 -0.6347873  0.4668624 -0.1747333
A      0.25621275  0.62427625  0.2374152  0.6905266 -0.1069635
Gini   0.63932751 -0.07587292  0.1485694 -0.1045647  0.7433014
Theil  0.62089258  0.06022316  0.1693868 -0.4396556 -0.6235997
M20    0.03487574  0.62548722 -0.6999388 -0.3177761  0.1290484
```

The trends seen in these results and Figure 67 are similar to what was found for EFIGI, CFHT, and KPNO data. The Gini coefficient again exhibits a strong

emphasis on PC1, followed closely by the Theil index. M20 dominates the second principal component, followed by asymmetry. Just as with the EFIGI, CFHT, and KPNO data, asymmetry, M20, and the Theil index have an inverse relationship with the central coefficient and Gini in PC2.

6.5 Minitab: Principle Component Analysis

We perform principal component analysis with the use of Minitab. Minitab is a statistics package for data analysis, developed at the Pennsylvania State University. We use Minitab version 18. Figure 68 displays the “loadings” of the CFHT, KPNO, and WINGS data sets, respectively from (a) through (c). In Figure 68 (a)-(c), MC is used to represent central concentration and MA is asymmetry. The results of PCA analysis from R are in agreement with results from Minitab.

It can be seen that the signs on the “loadings” are different. However, the signs are arbitrary notations. Changing the signs of the loadings does not change the variance that is contained in the principal components and has no influence on the interpretation of the results. The relation between the scores and the loadings stays the same.

Eigenanalysis of the Correlation Matrix

| | | | | | |
|------------|--------|--------|--------|--------|--------|
| Eigenvalue | 2.2134 | 1.6836 | 0.7684 | 0.2964 | 0.0381 |
| Proportion | 0.443 | 0.337 | 0.154 | 0.059 | 0.008 |
| Cumulative | 0.443 | 0.779 | 0.933 | 0.992 | 1.000 |

5360 cases used, 1 cases contain missing values

Eigenvectors

| Variable | PC1 | PC2 | PC3 | PC4 | PC5 |
|----------|--------|--------|--------|--------|--------|
| MC | 0.195 | -0.594 | -0.560 | -0.521 | 0.157 |
| MA | 0.325 | 0.595 | 0.172 | -0.701 | 0.135 |
| Gini | 0.658 | -0.097 | 0.040 | 0.090 | -0.740 |
| Theil | 0.646 | 0.038 | -0.049 | 0.440 | 0.621 |
| M20 | -0.073 | 0.531 | -0.808 | 0.189 | -0.155 |

(a.)

Eigenanalysis of the Correlation Matrix

| | | | | | |
|------------|--------|--------|--------|--------|--------|
| Eigenvalue | 2.7700 | 1.0612 | 0.8780 | 0.2510 | 0.0398 |
| Proportion | 0.554 | 0.212 | 0.176 | 0.050 | 0.008 |
| Cumulative | 0.554 | 0.766 | 0.942 | 0.992 | 1.000 |

Eigenvectors

| Variable | PC1 | PC2 | PC3 | PC4 | PC5 |
|----------|-------|--------|--------|--------|--------|
| MC | 0.504 | 0.250 | -0.306 | 0.768 | -0.014 |
| MA | 0.283 | -0.116 | 0.925 | 0.220 | 0.046 |
| Gini | 0.577 | 0.053 | -0.103 | -0.425 | 0.688 |
| Theil | 0.575 | -0.153 | -0.072 | -0.369 | -0.710 |
| M20 | 0.038 | -0.948 | -0.187 | 0.211 | 0.144 |

(b.)

Eigenanalysis of the Correlation Matrix

| | | | | | |
|------------|--------|--------|--------|--------|--------|
| Eigenvalue | 2.3319 | 1.6827 | 0.6231 | 0.3263 | 0.0359 |
| Proportion | 0.466 | 0.337 | 0.125 | 0.065 | 0.007 |
| Cumulative | 0.466 | 0.803 | 0.928 | 0.993 | 1.000 |

Eigenvectors

| Variable | PC1 | PC2 | PC3 | PC4 | PC5 |
|----------|-------|--------|--------|--------|--------|
| MC | 0.373 | -0.458 | 0.635 | 0.467 | 0.175 |
| MA | 0.256 | 0.624 | -0.237 | 0.691 | 0.107 |
| Gini | 0.639 | -0.076 | -0.149 | -0.105 | -0.743 |
| Theil | 0.621 | 0.060 | -0.169 | -0.440 | 0.624 |
| M20 | 0.035 | 0.625 | 0.700 | -0.318 | -0.129 |

(c.)

Figure 68: PCA analysis for CFHT (a), KPNO (b), and WINGS (c) data sets.

CHAPTER VII

DISCUSSION

7.1 Success of Morphology Software

From Figures 18, 36, 47, it can be seen that nonparameteric values measured for galaxies with the morphology software developed in this thesis correspond to appropriate visual classifications. Generally, we have found that galaxies with values of $C \geq 0.30$, $Gini \geq 0.40$, $A \leq 0.20$, and $Theil \leq 0.17$ are visually classified as early-type. From the analysis of a number of parameter planes, we find that a majority ($> 60\%$) of visually classified early-type galaxies and late-type galaxies are found in the proper regions. For example, on the C versus A plane, $> 63\%$ of the visually classified early-type galaxies are plotted in the “E/S0” region, and $> 80\%$ of the visually classified late-type galaxies are plotted in the the “S/Irr” region.

In this research, we focus on two main parameter planes in order to classify galaxies: A versus Gini and A versus Theil. On the A versus Gini and A versus Theil planes, $> 85\%$ of the visually classified early-type galaxies are plotted in the “E/S0” region, and $> 65\%$ of the visually classified late-type galaxies are plotted in the the “S/Irr” region. We conclude that our software is capable of classifying objects. Therefore, we apply the software to classify galaxies in data sets which do not have visual classifications.

Additional parameters can be integrated into the software, such as the smoothness parameter S or multi-mode M . The smoothness parameter (also known as clumpiness; Conselice 2003; Lotz *et al.* 2004) measures irregularities in a distribution of

light. It is the ratio of the amount of light contained in patchy regions to the total amount of light in the galaxy. To find S , the galaxy image is blurred by a boxcar filter and then subtracted from the original image. The result is then divided by the total light in the galaxy. For elliptical galaxies, S is generally very close to zero (Conselice 2003).

Multi-mode (Freeman *et al.* 2013; Peth 2016) is an area ratio of the two brightest regions of a galaxy, where the brightest regions are determined by a threshold method. This statistic can be used to detect double nuclei (Freeman *et al.* 2013). However, this statistic does not take into account pixel intensities.

The morphology software used in this thesis is limited by the seeing and resolution of a galaxy image. The limit of accurate classification is empirically found to be $3 \times \text{FWHM}$. Objects that are smaller than this value are found too small for their classifications to be considered accurate.

Also, the automatic classification method we have developed uses broad classification bins. Currently, through the application of A versus Gini \rightarrow A versus Theil cuts, galaxies are classified into one of three categories: early-types (E), late-types (S), and an in-between class (I). Lenticular galaxies are not clearly segregated from elliptical galaxies. Additional parameters may need to be tested in order to study finer detail in galaxy structure.

7.2 Exploring Galaxy Formation and Evolution

As discussed in Section 1.3.3, various mechanisms have been used to explain the different structures of galaxies. A focus in this thesis has been on galaxy clusters, therefore in this section we explore what effect dense environments have on galaxy morphology. From observations, it is known that early-type galaxies dominate the

centers of clusters, while late-type systems are generally found in the field or low-density environments (e.g. Hubble 1931; Oemler 1977; van der Wel *et al.* 2010). The fraction of spiral galaxies increases with radius from the center of a cluster, while the fraction of ellipticals decreases. This is known as the morphology-density relation (Dressler 1980).

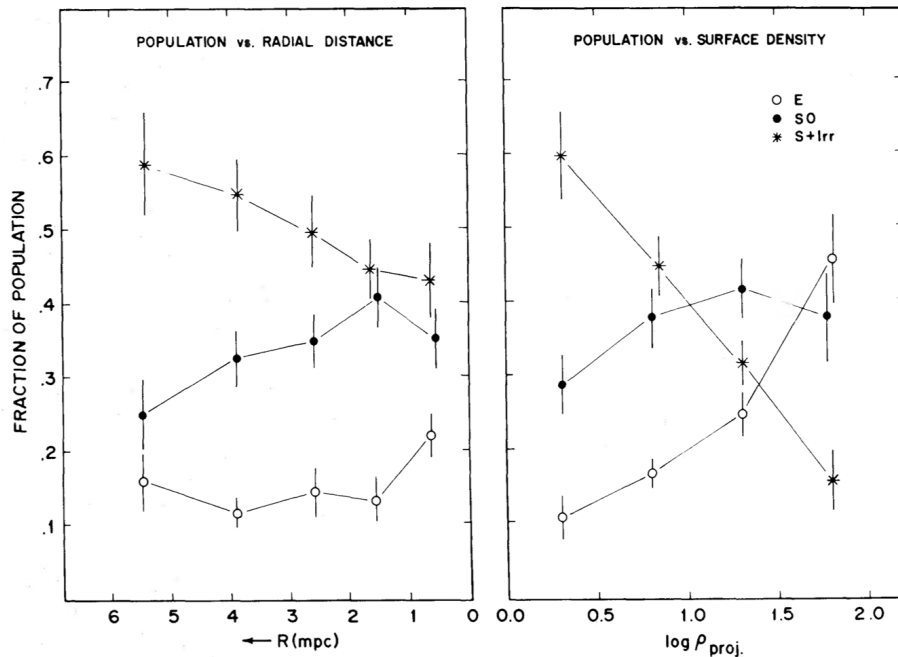


Figure 69: Fraction of galaxy type as a function of clustercentric distance and local density, from Dressler (1980).

Dressler (1980) studied 55 clusters at $z \leq 0.06$. The images were obtained using the Las Campanas 2.5m telescope and KPNO 1.5m telescope. The results of the study for the 55 clusters are depicted in Figure 69. This figure shows the morphology of the galaxy population as a function of radius from the cluster center (left), as well as galaxy type as a function of local surface density, *i.e.* morphology-density relation (right). It can be seen from the two panels that the change in the fraction of galaxy types changes with clustercentric radius and local density.

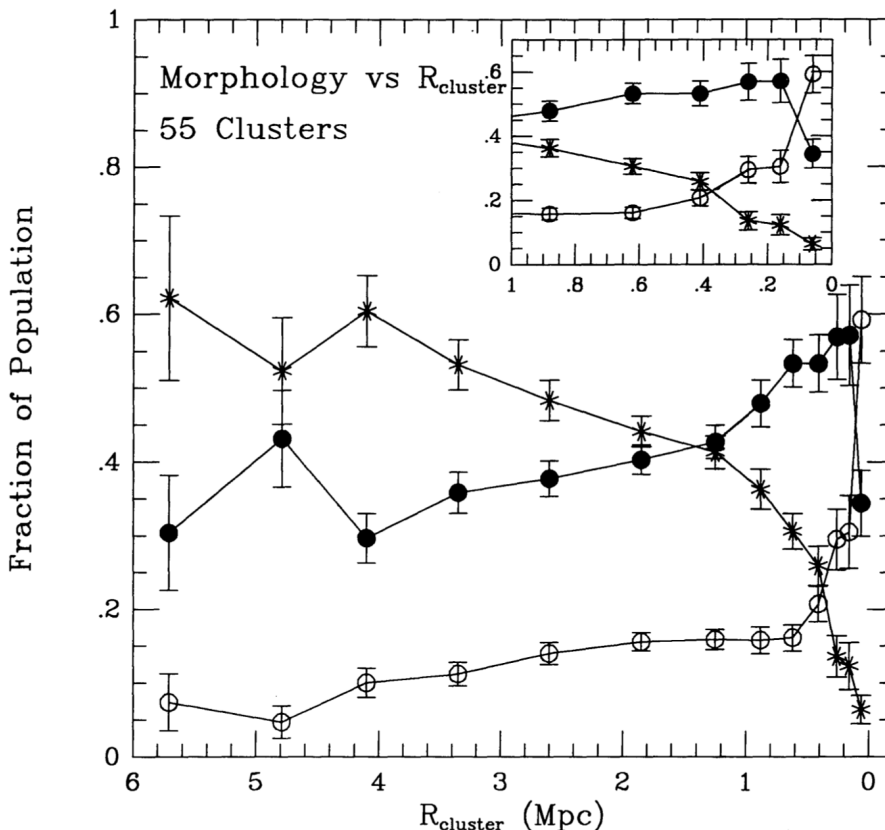


Figure 70: Fraction of galaxy type as a function of clustercentric radius for 55 clusters studied in Dressler (1980). Figure reproduced from Whitmore *et al.* (1993). See text for description.

The motivation for comparing morphology to local density instead of clustercentric radius is that clusters may not be symmetric in shape. Therefore, studying morphology relative to local density may be more fundamental. However, it has been argued that the morphology-clustercentric radius relation may be more significant than the morphology-density relation (Whitmore *et al.* 1993). Figure 70 is the morphology-clustercentric radius relation plot for the 55 clusters studied in Dressler (1980). Open circles represent E-type galaxies, closed circles are S0s, and asterisks are spirals and irregulars.

It has been found that both the morphology-density and morphology-clustercentric

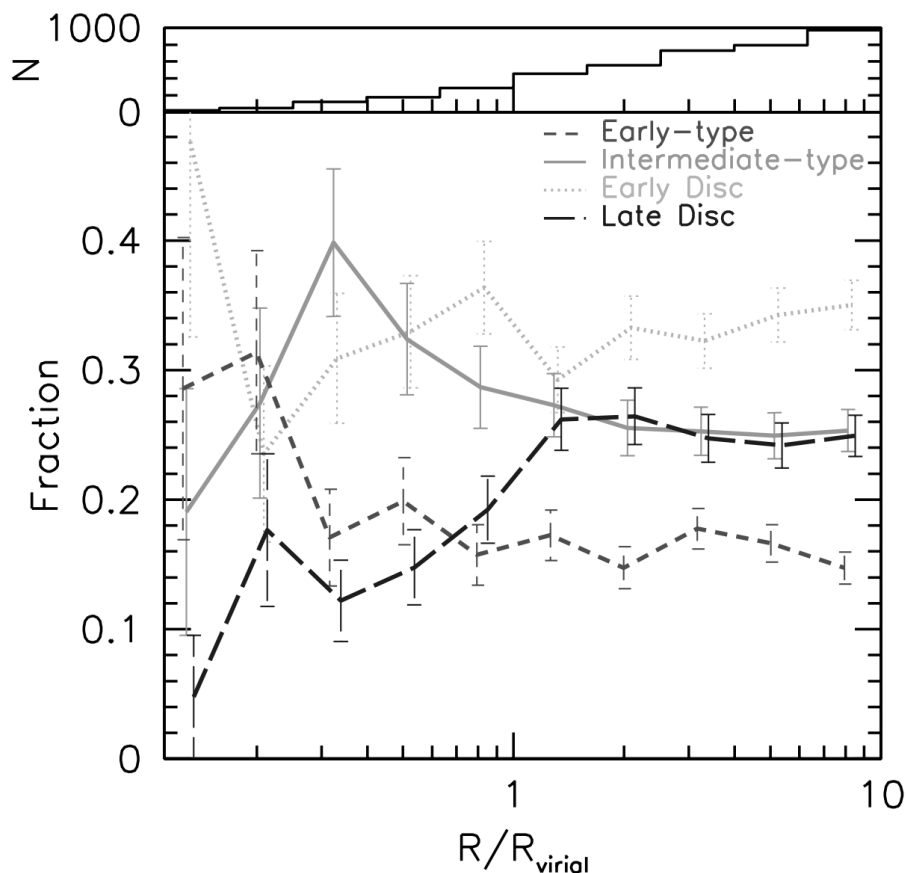


Figure 71: The morphology-radius relation for SDSS galaxies classified using automatic classifiers from Goto *et al.* 2003. See text for description.

radius show that the fractions of early-type galaxies increase towards cluster cores, while the fractions of late-type galaxies decrease toward cluster cores. Figure 71 presents the morphology-clustercentric radius relation for 7,938 galaxies from the SDSS Early Data Release (SDSS EDR) which are classified using automatic classifiers (Goto *et al.* 2003). The galaxies selected have $M_r < -20.5$. These galaxies are classified into four categories: early-type (short-dashed line), intermediate-type (solid line), early-disk (dotted line), and late-types (long-dashed line). Error bars are calculated using Poisson statistics. Figure 71 contains a histogram at the top, which shows the number of galaxies vs. clustercentric

Population vs. radial distance

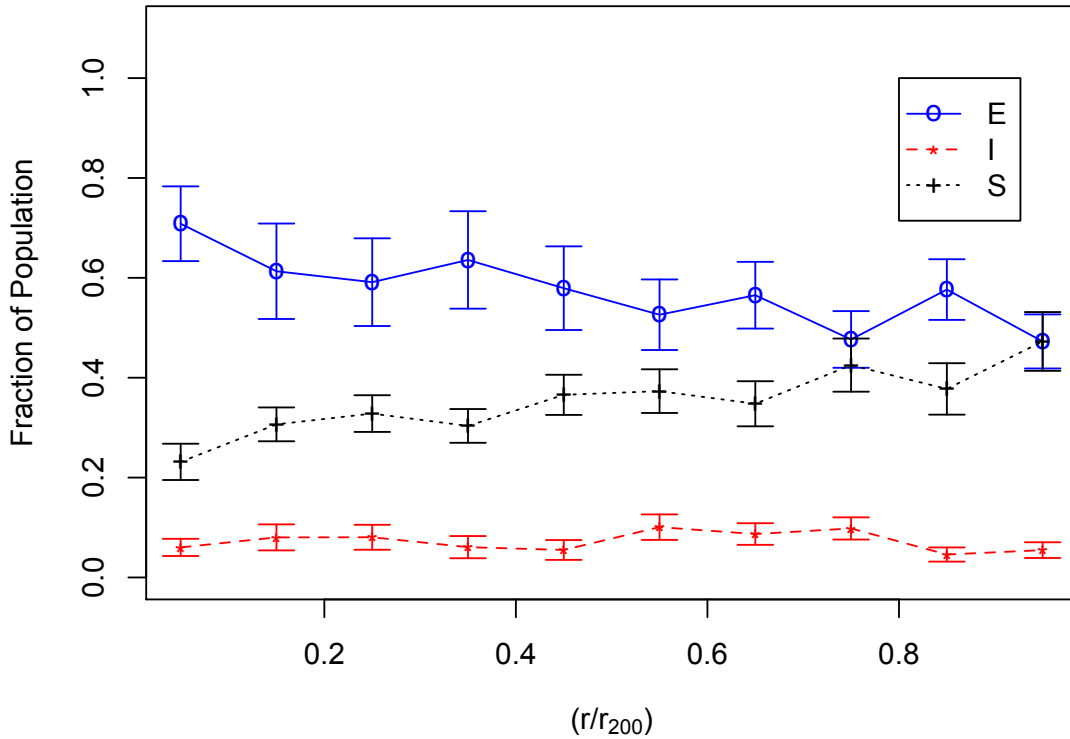


Figure 72: Fraction of galaxies from CFHT data set vs. (r/r_{200}) . See text for description.

radius portion of Figure 71, it can be seen that the fractions of morphological type are approximately constant for $(R/R_{virial}) > 1$. This suggests that the process causing the morphological differences does not have an effect on galaxies in this range (Goto *et al.* 2003). At $(R/R_{virial}) > 1$, late-type systems dominate over the early-type galaxies.

We compare Figure 71 to Figure 72, which is a plot of selected galaxies from the 15 clusters in the CFHT data set studied in this thesis. According to our classification criteria described previously, galaxies are first separated into bright (B) and dim (D) categories. From the B group, we select galaxies with $M_r < -18.1$. In order to classify these bright galaxies, A versus Gini $\rightarrow A$ versus Theil cuts are administered to the

data. In Figure 72, “E” represents early-type galaxies, “S” are late-type galaxies, and “T” are a class of galaxies that appear to be in between the clearly defined early- and late- types. Similar to Figure 71, we find the fraction of late-type galaxies cross-over with the early-type galaxies at approximately $(r/r_{200}) = 1$.

In order to explain the morphology-density or morphology-clustercentric radius relations, various theories exist about the impact the dense cluster center environments have on galaxies: 1) it has been suggested that S0 galaxies in clusters are a result of galaxy mergers or collisions, since in this process gas could be removed from the disk (Dressler 1980; Biviano 2000), 2) ram pressure stripping of gas in the disk due to galaxy motion through the hot inter-cluster medium (ICM) would result in early-type galaxies, 3) tidal stripping of gas due to close confrontations of galaxies would also result in the formation of early-type galaxies (e.g. Dressler 1980; Ferguson & Binggeli 1994).

Galaxy cluster cores are too dense for disks to form, therefore, it is believed that the violent processes arising in this environment are responsible for determining morphological fractions of the population rather than formation processes (Whitmore *et al.* 1993). The environment plays an important role in the development of S0 galaxies. Due to their abundance in cluster environments, it is believed that S0 galaxies develop from spiral galaxies through ram pressure stripping, mergers, etc. (Neistein 1999; Tapia 2017).

As can be observed from Figures 69, 70, 71 and 72, there is strong evidence to suggest that environment has an impact on galaxy formation and evolution, which ultimately produces their observed morphology.

7.3 Exploring Dwarf Galaxies

It is known that the morphology-density relation also holds for dwarf galaxies just as for higher mass galaxies (e.g. Ferguson & Sandage 1990; Ferguson & Binggeli 1994; Lisker 2006). As mentioned previously, cluster cores are hostile, dense environments which have the capacity to disturb and alter galaxy shapes. Unlike nucleated dwarf galaxies, low-mass and “loosely bound” dwarf galaxies (Oh & Lin 2000), such as non-nucleated dE’s, would not be able to retain their shape during tidal interactions with other galaxies in cluster cores. Therefore, it is expected that few non-nucleated dwarf galaxies should be found in cluster cores where tidal interactions are prominent.

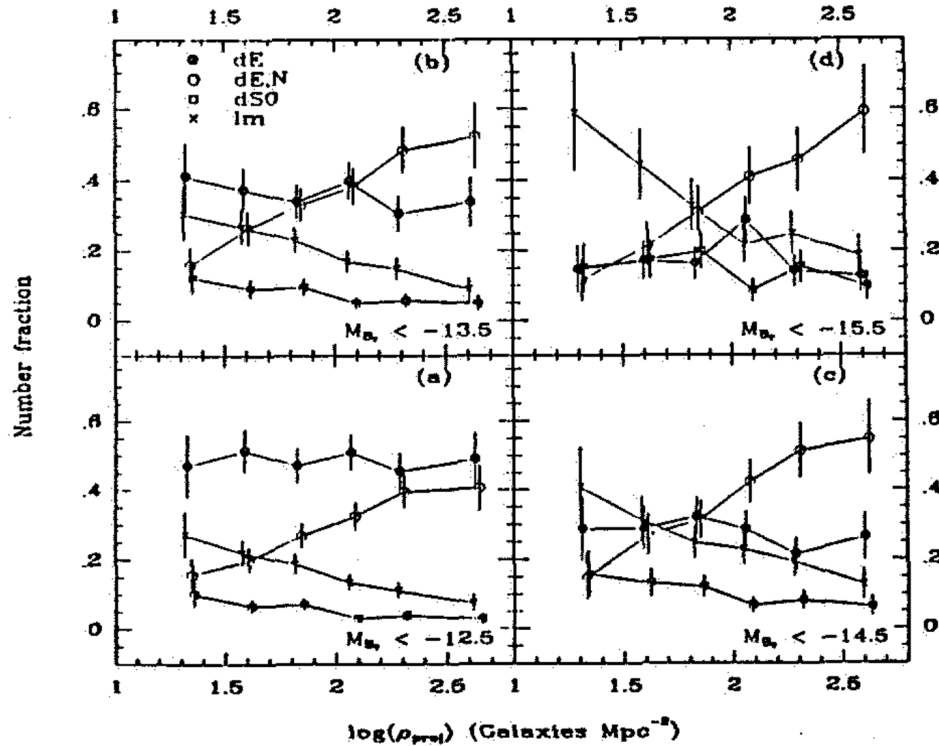


Figure 73: Fraction of galaxy type as a function of density for dwarf galaxies from Ferguson & Sandage (1990). See text for description.

The morphology-density relation for dwarf galaxies from Ferguson & Sandage

(1990) is reproduced in Figure 73. Non-nucleated dE's (labeled as dE) are closed circles, nucleated dE's (labeled as dE, N) are open circles, dwarf lenticular galaxies (dS0) are squares, and late-type galaxies (Im) are crosses. Ferguson & Sandage (1990) find that nucleated dE galaxies are “strongly clustered toward dense environments”.

Ordenes-Briceno *et al.* (2018) studied the distribution of nucleated and non-nucleated dwarf galaxies in the Fornax cluster in terms of the cluster's virial radius. They find that nucleated dwarfs are clustered in the inner regions near the cD galaxy rather than non-nucleated dwarf galaxies. These results are in agreement with Lisker *et al.* (2007), who studied nucleated and non-nucleated dwarf galaxies in the Virgo cluster. Similarly, we study the ratio of nucleated versus non-nucleated dwarf galaxies as a function of clustercentric radius in Figure 44 for dwarf galaxies in our CFHT data set. Unlike the previous studies mentioned, our sample includes dwarf galaxies from 15 clusters rather than one. However, likewise, we find that cluster cores possess a greater number of nucleated than non-nucleated dwarf galaxies (Thomson & Gregory 1993). Dwarf galaxies with high- C values (*i.e.* nucleated) are found in higher numbers in the cluster centers than in the outer regions of the 15 CFHT clusters. We discuss our future study of dwarf galaxies in the next chapter.

CHAPTER VIII

CONCLUSIONS

For this thesis, computer software has been developed in order to classify galaxies by measuring five structural parameters from digital images. I have constructed computer algorithms to calculate: 1) central concentration, the ratio of flux between the inner and outer radius of a galaxy, 2) asymmetry, which is based on how symmetric the light is distributed throughout a galaxy, 3) the Theil index, a statistic that is used to measure economic inequality that I have adopted to measure the distribution of light within individual galaxies, 4) the Gini coefficient, a statistical index similar to the Theil index, and 5) M20, the second-order moment of the brightest regions of a galaxy. The code I have developed is capable of measuring classification parameters for a single postage-stamp image of a galaxy, as well as from images of the full cluster. The CFITSIO library is used to read images in FITS format. Additional compiler flags were introduced to the CFITSIO library in order for the software to read the large cluster images used in this research.

The results of the program have been extensively tested for accuracy by comparing the output of the software with known published galaxy morphologies. Data used to test and train the software includes galaxies from the Galaxy Zoo project as well as the EFIGI catalog (selected from SDSS). The EFIGI catalog includes a large number of visually classified galaxies spanning a wide range of detailed morphologies. Figures 18, 19, and 20 show the distribution of the parameters studied in this thesis with respect to the galaxies' Hubble type. As can be seen from these figures and histograms in

this thesis, the parameters clearly segregate the different classes of galaxies.

Classification parameters were measured for numerous galaxies in a diverse range of redshifts. The results have been compiled into detailed catalogs. Among the data sets studied are high-redshift galaxies observed at CFHT from Rude (2015), low-redshift clusters observed at KPNO from Barkhouse *et al.* (2007), and nearby WINGS clusters from Fasano *et al.* (2003).

Emphasis is placed on the Theil index, since it is a new parameter being tested for classification purposes in this research. The Theil index displays a strong correlation with the Gini coefficient. Much like the Gini coefficient, the Theil index does not rely on the center of an object to be defined, which is an advantage particularly when studying high-redshift disturbed galaxies. We find that using the C versus Gini \rightarrow Gini versus Theil planes to select data segregates the classes of galaxies more efficiently than single parameter selection. This is the method we employ in this research to categorize galaxies that lack visual classifications.

We perform principal component analysis (PCA) in order to quantitatively explore the variance and correlation between classification parameters in our study. In this thesis, we implement PCA through Minitab 18 and the build-in function `prcomp()` in R. From Kaiser’s criteria, we reduce the five primary dimensions of our data sets to two dimensions, *i.e.* we describe our data with two principal components. The principal components display consistency throughout the data sets:

1. For the first principal component (PC1), we find that the Gini coefficient and the Theil index vary together — as Gini increases, so does the Theil statistic. PC1 is strongly correlated with these two variables, meaning, as Gini and Theil values increase, so does PC1. Central concentration and asymmetry do not exhibit a significant influence on PC1, but the variable that affects PC1 the

least is M20. In other words, PC1 can be thought of as a measure of the Theil index and the Gini coefficient.

2. For each data set it can be seen that the second principal component (PC2) is an inverse relation of asymmetry, Theil and M20 versus central concentration and the Gini coefficient. As one set increases, the other parameters decrease. This principal component can be described as predominantly a measure of asymmetry and M20.
3. The biplots are a visual display of the correlation of parameters and the strength of each original parameter on the two new principal components. From the biplots it can be clearly seen that M20 and asymmetry possess a correlation with each other, as well as the Theil index and the Gini coefficient. The central concentration has a greater correlation with the Theil index and Gini coefficient rather than M20 or asymmetry.

The ratio of nucleated versus non-nucleated dwarf galaxies with respect to their distance from the cluster core are studied for the CFHT cluster data from Rude (2015). We find that dwarf galaxies in the cluster core are predominantly nucleated, whereas those at the outer regions of the clusters are generally non-nucleated. As discussed in Section 7.3, this is consistent with observations and the current understanding of dwarf galaxies' development (e.g. Binggeli & Cameron 1991; Oh & Lin 2000; Mistani *et al.* 2015).

8.1 Future Work

In this work, emphasis was placed on studying the morphology of galaxies in dense environments. In the future, I will study the morphology of galaxies in the field. I plan to examine six high-redshift galaxy clusters and their parallel fields from the

HST Frontier Fields project. Using the *HST* Frontier Field data, I will compare morphology between high-redshift galaxies in high-density cluster regions to those in low-density parallel field regions. The *HST* Frontier Field data currently has the “deepest observations of clusters and their lensed galaxies ever obtained, and the second-deepest observations of blank fields (located near the clusters).”¹

Alongside the six *HST* Frontier Field data sets, I will analyze twenty-two massive galaxy clusters from the *HST* Cluster Lensing And Supernova (*CLASH*) survey, which has a median redshift of $z \sim 0.4$ (Postman *et al.* 2012). The morphology results from my software of the *HST* Frontier Field galaxies will be compared with the *HST CLASH* results. The high resolution of the *HST* images will also permit me to more robustly measure galaxy morphology to a given apparent magnitude and surface brightness compared to ground-based telescopes. In total, these data will sample galaxy types from clusters out to a redshift of $z = 0.545$. My software is being used to provide morphological classification of these distant galaxies.

The morphology software developed in this thesis has been applied to r -band images of 10 galaxy clusters ($0.03 < z < 0.15$) from Kalawila *et al.* (2018; in preparation). Images of the 10 clusters were taken at KPNO by the Mayall 4-m telescope with overlapping $H\alpha$ images. Observations show that the presence of $H\alpha$ is an indicator of star formation in galaxies (e.g. Kennicutt 1998). These data are being used to measure star formation of cluster galaxies based on morphology. Thus examining the impact of the high-density cluster environment on star formation and galaxy evolution.

Dense galaxy clusters may contain giant elliptical galaxies (cD) that possess large, diffuse envelopes. It is believed that cD galaxies form by mergers (e.g. Schneider & Gunn 1982). Additionally, it has been observed that there are similarities between cD halos and cluster properties, which suggests that the large halos of the cD galaxies are

¹<http://www.stsci.edu/hst/campaigns/frontier-fields/>

formed by cluster processes (Kormendy 1982). It may be possible that non-nucleated dwarf galaxies in cluster centers are disturbed as the cluster evolves, and become “redistributed throughout the cluster potential” as part of the cD halo (Lopez-Cruz *et al.* 1996). In the future, I will examine the difference of nucleated versus non-nucleated dwarf galaxies in clusters with supergiant cD galaxies as opposed to those without them. Since non-nucleated galaxies may become distributed through the cD halo, we hypothesize that there should be fewer non-nucleated dwarf galaxies in the presence of cD’s.

8.2 Final Thoughts

Through the use of computers to analyze galaxy morphology, this project hopes to make an important contribution to the study of galaxy classification in astronomy. Determining how galaxy type is related to various physical parameters will help us to obtain a more complete understanding of galaxy formation and evolution. This project has the potential to be used as a convenient tool for evaluating galaxy type measured for large area surveys, such as the *DES*, Panoramic Survey Telescope and Rapid Response System (*PanSTARRS*), or *LSST*. Data from large area surveys will be difficult to manage with the aid of citizen projects, thus the use of computer software to classify and analyze the morphology of galaxies will be extremely important in terms of efficiency.

Besides *DES*, *PanSTARRS*, and *LSST*, computer programs such as the one developed in this thesis can be used to analyze data from spacecraft-borne instruments like the *HST* and *James Webb Space Telescope*. Since space telescopes are able to get very high-resolution measurements of distant galaxies, with my software it will be possible to study and classify distant galaxies. The main goal of this project is

to ensure that well-tested automatic galaxy classification software is available when data from large area surveys is released to the community, and to contribute to the understanding of galactic development.

APPENDICES

Below we include samples of computer code used in this thesis.

APPENDIX A: An example of the Theil index FORTRAN code in our software.

```
REAL FUNCTION Thiel(SEGDATA)
  IMPLICIT NONE
  REAL,DIMENSION(:,:),INTENT(in)           :: SEGDATA
  REAL,DIMENSION(SIZE(SEGDATA,1)*SIZE(SEGDATA,2)) :: T1
  REAL,DIMENSION(SIZE(SEGDATA,1)*SIZE(SEGDATA,2)) :: T2
  INTEGER                                       :: i, j, k, kf, n, nx, ny
  INTEGER                                       :: i_counter_0s, kk
  REAL                                         :: Xmean_k, xTt

  CALL WriteFITSImage(SEGDATA,'foobar.fits')
  nx = SIZE(SEGDATA,1);
  ny = SIZE(SEGDATA,2);

! A new way: convert the matrix to a 1-dimensional array T1
  kf = 0
  do i=1, nx
    do j=1, ny
      T1 (kf) = SEGDATA(i,j)
      kf = kf + 1
    enddo
  enddo

  i_counter_0s = 0
  Xmean_k=0 ! an arithm. avarage value of the T1 array
  kk = 1
  do k=1, kf
    Xmean_k = Xmean_k + G1(k)
```

```

        if (T1(k) .ne. 0) then
            T2(kk) = T1(k)
            kk = kk+1
        else if (T1(k) .eq. 0) then
            i_counter_0s = i_counter_0s + 1
        end if
    enddo
kf = kf - i_counter_0s      ! NEW kf
Xmean_k = Xmean_k / kf
T1 = T2

! The Theil index:
xTt = 0
do i = 1, kf
    if (T1(i) .ne. 0) xTt = xTt + G1(i)/Xmean_k * log(G1(i)/Xmean_k)
    !If T1 is not equal to zero then find the Thiel and do the summation
enddo
Thiel = xTt / kf
Thiel = Thiel/ log(Real(kf))

RETURN
END FUNCTION Thiel

```

APPENDIX B: An example of a GALFIT input file.

```

# IMAGE and GALFIT CONTROL PARAMETERS
A) $galaxy           # Input data image (FITS file)
B) $galaxy.out.fits  # Output data image block
C) none              # Sigma image name (made from data if blank or \"none\")
D) none              # Input PSF image and (optional) diffusion kernel
E) 1                 # PSF fine sampling factor relative to data
F) none              # Bad pixel mask (FITS image or ASCII coord list)
G) none              # File with parameter constraints (ASCII file)
H) 100 400 100 400  # Image region to fit (xmin xmax ymin ymax)
I) 100 100           # Size of the convolution box (x y)
J) $mag_ZP           # Magnitude photometric zeropoint
K) $pixscale $pixscale # Plate scale (dx dy) [arcsec per pixel]
O) regular           # Display type (regular, curses, both)
P) 0                 # Choose: 0=optimize, 1=model, 2=imgblock, 3=subcomps
# For object type, the allowed functions are:
#     nuker, sersic, expdisk, devauc, king, psf, gaussian, moffat,
#     ferrer, powersic, sky, and isophote.
#
# Hidden parameters will only appear when they're specified:

```

```

#      CO (diskyness/boxyness),
#      Fn (n=integer, Azimuthal Fourier Modes),
#      R0-R10 (PA rotation, for creating spiral structures).
# -----
#   par)   par value(s)   fit toggle(s)   # parameter description
# -----
# INITIAL OBJECT FITTING PARAMETERS
# Object number: 1
0) expdisk                # object type
1) 250.0 250.0 1 1        # position x, y
3) 20.0   1               # total magnitude
4) 20.5   1               # Rs   [Pixels]
9) 0.5    1               # axis ratio (b/a)
10) 1.0   1               # position angle (PA) [deg: Up=0, Left=90]
Z) 0                    # Skip this model in output image? (yes=1, no=0)

# Object number: 2
0) sersic                 # object type
1) 250.0   250.0 1 1      # position x, y
3) 20.0     1             # Integrated magnitude
4) 7.0      1             # R_e (half-light radius) [pix]
5) 4.0      1             # Sersic index n (de Vaucouleurs n=4)
6) 0.0000   0            # -----
7) 0.0000   0            # -----
8) 0.0000   0            # -----
9) 0.5      1             # axis ratio (b/a)
10) 1.0     1             # position angle (PA) [deg: Up=0, Left=90]
Z) 0                    # output option (0 = resid., 1 = Don't subtract)

# Object number: 3
0) sky                   # object type
1) $sky_value 1          # sky background at center of fitting region [ADUs]
2) 0.0000   0            # dsky/dx (sky gradient in x)
3) 0.0000   0            # dsky/dy (sky gradient in y)
Z) 0                    # output option (0 = resid., 1 = Don't subtract)

```

BIBLIOGRAPHY

- Abraham, R. G., Valdes, F., Yee, H. K., van den Bergh, S. (1994). The morphologies of distant galaxies. 1: an automated classification system. *The Astrophysical Journal*, 432, 75-90.
- Abraham, R. G., Tanvir, N. R., Santiago, B. X., et al. (1996a). Galaxy morphology to I=25 mag in the Hubble Deep Field. *Monthly Notices of the Royal Astronomical Society*, 279, L47-L52.
- Abraham, R. G., van den Bergh, S., Glazebrook, K., et al. (1996b). The morphologies of distant galaxies. II. Classifications from the Hubble Space Telescope Medium Deep Survey. *The Astrophysical Journal Supplement*, 107, 1-17.
- Abraham, R. G., van den Bergh, S., Nair, P. (2003). A new approach to galaxy morphology. I. Analysis of the Sloan Digital Sky Survey Early Data Release. *The Astrophysical Journal*, 588, 218-299
- Adams, W. S. (1941). Some results with the Coude Spectrograph of the Mount Wilson Observatory. *The Astrophysical Journal*, 93, 11A.
- Allison, P. D. (1978). Measures of inequality. *American Sociological Review*, 43, 865-880.
- Alpher, R. A., Herman, R. C. (1948). On the relative abundance of the elements. *Physical Review*, 74(12), 1737.
- Athanassoula, E. (2005). On the nature of bulges in general and of box/peanut bulges in particular: input from N-body simulations. *Monthly Notices of the Royal Astronomical Society*, 358, 1477A.
- Bailey, M. E., Butler, C. J., McFarland, J. (2005). Unwinding the discovery of spiral

- nebulae. *Astronomy & Geophysics*, 46(2), 2.26-2.28.
- Baillard, A., Bertin, E., de Lapparent, V., et al. (2011). The FIGI catalogue of 4458 nearby galaxies with detailed morphology. *Astronomy & Astrophysics*, 532, 74.
- Barkhouse, W. A., Yee, H. K. C., Lopez-Cruz, O. (2007). The Luminosity Function of Low-redshift Abell Galaxy Clusters. *The Astrophysical Journal*, 671, 1471.
- Barkhouse, W. A., Yee, H. K. C., Lopez-Cruz, O. (2009). The Galaxy Population of Low-Redshift Abell Clusters. *The Astrophysical Journal*, 703, 2024-2032.
- Berendzen, R., Hart, R., & Seeley, D. (1976). *Man Discovers the Galaxies*.
- Binggeli, B., Cameron, L. M. (1991). Dwarf galaxies in the Virgo cluster. *Astronomy and Astrophysics*, 252, 27-52.
- Biviano, A. (2000). From Messier to Abell: 200 years of science with galaxy clusters. Invited talk at “Constructing the Universe with clusters of galaxies”, F. Durret D. Gerbal organizers, Paris. *ArXiv Astrophysics e-prints*, arXiv:astro-ph/0010409
- Blanton, M., Dalcanton, J., Eisenstein, D., Loveday, J., et al. (2001). The Luminosity Function of Galaxies in SDSS Commissioning Data. *The Astronomical Journal*, 121, 2358-2380.
- Brinchmann, J., Abraham, R. G., Schade, D., et al. (1998). Hubble space telescope imaging of the cfrs and ldss redshift surveys. I. Morphological properties. *The Astrophysical Journal*, 499, 112-133.
- Brown, J. P. (1997). *A deep imaging survey of brightest cluster galaxies*. (Ph.D Thesis). The Department of Astronomy and Astrophysics, University of Toronto.
- Buta R. J. (2000). Galaxies: Classification. *Encyclopedia of Astronomy and Astrophysics*. (Volume 1).
- Buta R. J. (2011). Galaxy Morphology. *ArXiv Astrophysics e-prints*, arXiv:1102.0550
- Carroll, B. W., Ostlie, D. A. (2006). *An Introduction to Modern Astrophysics and*

Cosmology.

- CFITSIO User's Reference Guide Version 3.2 (2010). Goddard Space Flight Center
- Cheng, J. Y., Faber, S. M., Simard, L., Graves, G. J., et al. (2009). Automated morphological classification of Sloan Digital Sky Survey red sequence galaxies. *Monthly Notices of the Royal Astronomical Society*, 412(2), 727-747.
- Chiosi, C., Merlin, E., Piovan, L., Tantalo, R. (2014). Monolithic view of galaxy formation and evolution. *Galaxies*, 2, 300-381.
- Clements, D. L. (2014). *Infrared Astronomy — Seeing the Heat: from William Herschel to the Herschel Space Observatory.*
- Coghlan, A. (2014). *A little book of R for multivariate analysis.*
- Conselice, C. (1997). The symmetry, color, and morphology of galaxies. *Publications of the Astronomical Society of the Pacific*, 109, 1251C.
- Conselice, C., Bershad, M. A., Jangren, A. (2000). The asymmetry of galaxies: physical morphology for nearby and high-redshift galaxies. *The Astrophysical Journal*, 529, 886-910.
- Conselice, C., Gallagher, J. S., Rosemary, F.G.W. (2001). Galaxy Populations and Evolution in Clusters. I. Dynamics and the Origin of Low-Mass Galaxies in the Virgo Cluster. *The Astrophysical Journal*, 559, 791-811.
- Conselice, C. (2003). The Relationship between Stellar Light Distributions of Galaxies and Their Formation Histories. *The Astrophysical Journal Supplement*, 147(1), 1-28.
- Conselice, C. (2014). The evolution of galaxy structure over cosmic time. *Annual Review of Astronomy & Astrophysics*, 52, 291-337.
- Cozzens, S. & Bobb, K. (2003). Measuring the relationship between high technology development strategies and wage inequality. *Scientometrics*, 58.2, 351-368.
- Delgado R. S., (2010). *The Evolution of the Hubble Sequence: morpho-kinematics of*

- distant galaxies*. (Ph.D Thesis). Astronomy and Astrophysics of Ile De France, Observatoire De Paris.
- de la Calleja, J., Fuentes, O. (2004). Automated classification of galaxy images. *Lecture Notes in Computer Science*, 3215, 411-418.
- de Lapparent, V., Baillard, A., Bertin, E. (2011). The EFIGI catalogue of 4458 nearby galaxies with morphology II. Statistical properties along the Hubble sequence. *Astronomy & Astrophysics*, 532, 75.
- de Vaucouleurs, G. (1957). *The Discovery of the Universe*.
- de Vaucouleurs, G. (1959). *Handbuch der Physik*, 53, 275.
- de Vaucouleurs, G. (1991; 1995). *Third Reference Catalog of Bright Galaxies (RC3)*.
- de Vaucouleurs, G. (1994). *Global Physical Parameters Of Galaxies*. Presentation.
- Deng, X. (2013). A tool for the morphological classification of galaxies: the concentration index. *ArXiv Astrophysics e-prints*, arXiv:1301.4755
- Dressler, A. (1980). Galaxy morphology in rich clusters: Implications for the formation and evolution of galaxies. *The Astrophysical Journal*, 236, 351-365.
- Ellis, R. S., Abraham, R. G., Brinchmann, J., Menanteau, F. (2000). The story of galaxy evolution in full colour. *Astronomy & Geophysics*, 41, 2.10.
- Ebeling, H., Voges, W., Bohringer, H., Edge, A. C., et al. (1996). Properties of the X-ray-brightest Abell-type clusters of galaxies (XBACs) from ROSAT All-Sky Survey data - I. The sample. *Monthly Notices of the Royal Astronomical Society*, 281, 799-829.
- Ebeling, H., Edge, A. C., Bohringer, H., Allen, S. W., et al. (1998). The ROSAT Brightest Cluster Sample - I. The compilation of the sample and the cluster log N-log S distribution. *Monthly Notices of the Royal Astronomical Society*, 301, 881-914.
- Ebeling, H., Edge, A. C., Allen, S. W., Crawford, C. S., et al. (2000). The ROSAT

- Brightest Cluster Sample - IV. The extended sample. *Monthly Notices of the Royal Astronomical Society*, 318, 333-340.
- Engel, A. (1997). *Volume 6: The Berlin Years: Writings, 1914-1917* (English translation supplement). Retrieved from <http://einsteinpapers.press.princeton.edu/vol6-trans/>
- Evans, R. (2015). *The Cosmic Microwave Background: How It Changed Our Understanding of the Universe*.
- Fasano, G., Poggianti, B., Bettoni, D., Pignatelli, E., et al. (2003). The WINGS Survey: a progress report. *Memorie della Societa Astronomica Italiana*, 74, 355.
- Fasano, G., Vanzella, E., Dressler, A., Poggianti, B., et al. (2012). Morphology of galaxies in the WINGS clusters. *Monthly Notices of the Royal Astronomical Society*, 420, 926-948.
- Ferguson, H. C., & Sandage, A. (1990). The morphology-density relation for dwarf galaxies. *NASA, Ames Research Center, The Interstellar Medium in External Galaxies: Summaries of Contributed Papers* 281-283.
- Ferguson, H. C. & Binggeli, B. (1994). Dwarf Elliptical Galaxies. *ArXiv Astrophysics e-prints*, arXiv:astro-ph/9409079
- Ferrari, F., de Carvalho, R. R., Trevisan, M. (2015). Morfometryka — a new way of establishing morphological classification of galaxies. *The Astrophysical Journal*, 814(1).
- Freeman, P. E., Izbicki, R., Lee, A. B., et al. (2013). New image statistics for detecting disturbed galaxy morphologies at high redshift. *Monthly Notices of the Royal Astronomical Society*, 434, 282-295.
- Gadotti, D. A. (2009). Structural properties of pseudo-bulges, classical bulges and elliptical galaxies: a Sloan Digital Sky Survey perspective. *Monthly Notices of the Royal Astronomical Society*, 393, 1531-1552

- Galileo, G. (1610). *Starry Messenger*. translated by A. van Helden.
- Graham, A. W. (2001). An investigation into the prominence of spiral galaxy bulges. *The Astronomical Journal*, 121, 820.
- Grant, N. I., Kuipers, J. A., Phillipps, S. (2005). Nucleated dwarf elliptical galaxies in the Virgo cluster. *Monthly Notices of the Royal Astronomical Society*, 363, 1019-1030.
- Grebel, E. K. (1998). Star Formation Histories of Local Group Dwarf Galaxies. *ArXiv Astrophysics e-prints*, arXiv:astro-ph/9806191
- Greenacre, M., (2010). *Biplots in practice*.
- Greenacre, M., (2012). Biplots: the joy of singular value decomposition. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4 (4), 399-406.
- Hart, R., Berendzen, R. (1971). Hubble's classification of non-galactic nebulae, 1922-1926. *Journal for the History of Astronomy*, 2, 109H.
- Hetherington, N. S. (1993). *Cosmology: historical, literary, philosophical, religious, and scientific perspectives*.
- Holmber, E. (1958). A photographic photometry of extragalactic nebulae. *Lund Meddelande från Lunds Observatorium, Series II*, 135, 1-89.
- Holwerda, B. W., Munoz-Mateos, J.-C., Comeron, S., et al. (2014). Morphological parameters of a spitzer survey of stellar structure in galaxies. *The Astrophysical Journal*, 781, 12.
- Hotelling, H. (1933). *Analysis of a complex of statistical variables into principal components*. *Journal of Educational Psychology*, 24, 417-441.
- Houjun, M., van den Bosch, F., White, S. (2010). *Galaxy Formation and Evolution*.
- Howell, S. B. (2006). *Handbook of CCD Astronomy*.
- Hubble, E. (1922). A general study of diffuse galactic nebulae. *The Astrophysical Journal*, 56, 162.

- Hubble, E. (1926). Extra-galactic nebulae. *The Astrophysical Journal*, 64, 321-369.
- Hubble, E. (1929a). A spiral nebula as a stellar system, Messier 31. *The Astrophysical Journal*, 69, 103.
- Hubble, E. (1929b). A relation between distance and radial velocity among extra-galactic nebulae. *Proceedings of the National Academy of Sciences*, 15, 168-173.
- Hubble, E., Humason, M. L. (1931). The velocity-distance relation among extra-galactic nebulae. *The Astrophysical Journal*, 74, 43.
- Hubble, E. (1936). *Realm of the Nebulae*.
- Janert, P. K. (2010). *Data Analysis with Open Source Tools*.
- Jeans, J. H. (1917). Rotation as a factor in cosmic evolution. *Monthly Notices of the Royal Astronomical Society*, 77, 186-99.
- Jeffreys, H. (1918). On the early history of the Solar System. *Monthly Notices of the Royal Astronomical Society*, 78, 424-442.
- Jones, C., & Forman, W. (1999). Einstein Observatory Images of Clusters of Galaxies. *Astrophysical Journal*, 511, 65.
- Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20, 141-151.
- Kennicutt, R. C. (1998). Star Formation in Galaxies Along the Hubble Sequence. *Annual Review of Astronomy and Astrophysics*, 36, 189.
- Kennicutt, R. C. (1983). Observations of galaxy structure and dynamics. *IN: Morphology and dynamics of galaxies; Proceedings of the Twelfth Advanced Course, Saas-Fee, Switzerland*, 113-288.
- Kormendy, J., & Kennicutt, R. C. (2004). Secular evolution and the formation of pseudobulges in disk galaxies. *Annual Review of Astronomy & Astrophysics*, 42, 603.
- Lay, D. (2012). *Linear Algebra and its applications*, 4th ed.

- Lintott, C., Schawinski, K., Bamford, S., et al. (2010). Galaxy Zoo 1: data release of morphological classifications for nearly 900 000 galaxies. *Monthly Notices of the Royal Astronomical Society*, *410*, 166-178.
- Lisker, T., Glatt, K., Westera, P., Grebel, E. K. (2006). Virgo Cluster Early-Type Dwarf Galaxies with the Sloan Digital Sky Survey. II. Early-Type Dwarfs with Central Star Formation. *The Astronomical Journal*, *132*, 2432-2452.
- Lisker, T., Grebel, E. K., Binggeli, B., Glatt, K. (2007). Virgo cluster early-type dwarf galaxies with the Sloan Digital Sky Survey. III. Subpopulations: distributions, shapes, origins. *The Astronomical Journal*, *660*, 1186-1197.
- Lopez-Cruz, O. (1997). *Photometric properties of low-redshift galaxy clusters*. (Ph.D Thesis). The Department of Astronomy and Astrophysics, University of Toronto.
- Lopez-Cruz, O., Yee, H. K. C., Brown, J. P., Jones, C., & Forman, W. (1997). Are Luminous cD Halos Formed by the Disruption of Dwarf Galaxies?. *The Astrophysical Journal Letters*, *475(2)*, L97-L101.
- Lopez-Cruz, O. (2001). Photometric Properties of Low-Redshift Galaxy Clusters. *Revista Mexicana de Astronomia y Astrofisica Serie de Conferencias*, *11*, 183.
- Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, *9*, 209L.
- Lotz, J. M., Primack, J., Madau, P. (2004). A new nonparametric approach to galaxy morphological classification. *The Astronomical Journal*, *128*, 163.
- Lotz, J. M., Jonsson, P., Cox, T. J., Primack, J. R. (2008). Galaxy Merger Morphologies and Time-Scales from Simulations of Equal-Mass Gas-Rich Disc Mergers. *Monthly Notices of the Royal Astronomical Society*, *391*, 10.
- Luminet, J.P. (2011). Editorial note to: Georges Lemaitre, The beginning of the world from the point of view of quantum theory. *General Relativity and Gravitation*, *43*, 2911-2928.

- Lutz, D. (1991). NGC 3597: formation of an elliptical via merging?. *Astronomy & Astrophysics*, 245, 31L.
- Martinez, P., & Klotz, A. (1997). *A Practical Guide to CCD Astronomy*. Translated by Andre Demers.
- McKellar, A. (1941). The problem of possible molecular identification for interstellar lines. *Publications of the Astronomical Society of the Pacific*, 53, 233.
- McLean, I. S. (2008). *Electronic Imaging in Astronomy: Detectors and Instrumentation*.
- Milne, E. A. (2013). *Sir James Jeans: a Biography*.
- Morgan, W. W., Mayall, N. U. (1957). A spectral classification of galaxies. *Publications of the Astronomical Society of the Pacific*, 69, 291M.
- Morgan, W. W. (1958). A preliminary classification of the forms of galaxies according to their stellar population. *Publications of the Astronomical Society of the Pacific*, 70, 364.
- Morgan, W. W. (1959). A preliminary classification of the forms of galaxies according to their stellar population. II.. *Publications of the Astronomical Society of the Pacific*, 71, 394.
- Morgan, W. W., Kayser, S., White, R. A. (1975). cD galaxies in poor clusters. *The Astrophysical Journal*, 199, 545M.
- Nair, P. (2009). *The Morphology of Local Galaxies and the Basis of the Hubble Sequence*. (Ph.D Thesis). The Department of Astronomy and Astrophysics, University of Toronto.
- Neistein, E., Maoz, D., et al. (1999). A Tully-Fisher relation for S0 galaxies. *The Astronomical Journal*, 117, 2666-2675.
- Oh, K. S., Lin, D. N. C. (2000). Nucleation of dwarf galaxies in the Virgo cluster. *The Astrophysical Journal*, 543, 620-633.

- Oemler, A. (1977). The galaxy content of clusters. *Highlights of Astronomy*, 4(1), 253-260.
- O'Leary, E. (2013). Galaxy merger identification in the GOODS-South Field. *Macalester Journal of Physics and Astronomy*, 1(1), 1-20.
- Odewahn, S. C. (1995). Automated classification of astronomical images. *Publications of the Astronomical Society of the Pacific*, 107, 77.
- Ordenes-Briceno, Y., Eigenthaler, P., Taylor, M.A. et al. (2018). The Next Generation Fornax Survey (NGFS): III. Revealing the Spatial Substructure of the Dwarf Galaxy Population Inside Half of Fornax's Virial Radius. *ArXiv Astrophysics e-prints*, arXiv:1803.10784
- Pasachoff, J. M., Filippenko, A. (2007). *The Cosmos: Astronomy in the New Millennium*.
- Paturel, G., Fouque, P., Bottinelli, L., Gouguenheim, L. (1989). An extragalactic database. I - The Catalogue of Principal Galaxies. *Astronomy & Astrophysics Supplement*, 80, 299-315.
- Paturel, G., Fouque, P., Bottinelli, L., Gouguenheim, L. (1995). VizieR Online Data Catalog: Catalogue of Principal Galaxies (PGC). *Astronomy & Astrophysics Supplement*, 80, 299P.
- Pearson, K. (1901). *On Lines and Planes of Closest Fit to Systems of Points in Space*. *Philosophical Magazine*, 2, 559-572.
- Peng, C.Y., Ho, L. C., Impey, C. D., Rix, H. -W. (2002). Detailed structural decomposition of galaxy images. *The Astronomical Journal*, 124, 266.
- Peng, C.Y. (2003). *Galfit User's Manual*. Retrieved from <https://users.obs.carnegiescience.edu/peng/work/galfit/galfit.html>
- Penzias, A. A., Wilson, R. W. (1965). A measurement of excess antenna temperature at 4080 Mc/s. *The Astrophysical Journal*, 142, 419.

- Peth, M. A. (2016). *Using Machine Learning to Study the Relationship Between Galaxy Morphology and Evolution*. (Ph.D Thesis). Johns Hopkins University. Baltimore, MD.
- Petrosian, V. (1976). Surface brightness and evolution of galaxies. *The Astrophysical Journal*, 209, L1-L5
- Phan, T. (2016). *An Introduction to Principal Component Analysis with Examples in R*. Technical Report. Retrieved from <http://www.octoberraindrops.com>
- Plutarch (1878). *The Morals*. Volume 3. revised by W. W. Goodwin.
- Postman, M., *et al.* (2012). Cluster Lensing And Supernova survey with Hubble (CLASH): An Overview. *The Astrophysical Journal Supplement*, 199, 25.
- Povic, M., Marquez, I., Mesagosa, J., Perea, J., *et al.* (2015). The impact from survey depth and resolution on the morphological classification of galaxies. *Monthly Notices of the Royal Astronomical Society*, 453, 1644-1668.
- Reynolds, J.H. (1927). The classification of the spiral nebulae. *Observatory*, 50, 185.
- Rude, C. (2015). *Tracking Star Formation in Dwarf Cluster Galaxies*. (Ph.D Thesis). The Department of Physics and Astrophysics, University of North Dakota.
- Sandage, A. (1961). *The Hubble Atlas of Galaxies*.
- Sandage, A. & Binggeli, B., (1984). Studies of the Virgo cluster. III - A classification system and an illustrated atlas of Virgo cluster dwarf galaxies. *The Astronomical Journal*, 89, 919-931.
- Scarlata, C., Carollo, M., Lilly, S., Sargent, T., *et al.* (2007). Cosmos morphological classification with the Zurich Estimator of Structural Types (ZEST) and the evolution since $z = 1$ of the luminosity function of early, disk, and irregular galaxies. *The Astrophysical Journal Supplement*, 172, 406-433.
- Schneider, D. P. & Gunn, J. E. (1982). V ZW 311 - The once and future cD. *The*

- Astrophysical Journal*, 263, 14-22.
- Scrucca, L. (2004). qcc: an R package for quality control charting and statistical process control. *R News* 4/1, 11-17.
- Seeds, M., Backman, D. (2011). *Foundations of Astronomy*.
- Shamir, L. (2009). Automatic morphological classification of galaxy images. *Monthly Notices of the Royal Astronomical Society*, 399(3), 1367-1372.
- Shimasaku, K., Fukugita, M., Doi, M., Hamabe, M., et al. (2001). Statistical Properties of Bright Galaxies in the Sloan Digital Sky Survey Photometric System. *The Astronomical Journal*, 122, 1238-1250.
- Simard, L., Willmer, C. N. A., Vogt, N. P., Sarajedini, V. L., et al. (2002). The DEEP Groth Strip Survey. II. Hubble Space Telescope Structural Parameters of Galaxies in the Groth Strip. *The Astrophysical Journal Supplement*, 142(1), 1-33.
- Shapley, H., & Paraskevopoulos, J.S. (1940). Galactic and extragalactic studies, III. Photographs of thirty southern nebulae and clusters. *Proceedings of the National Academy of Sciences*, 26, 31.
- Schiller, J. (2010). *Big Bang & Black Holes*.
- Shu, F. (1982). *Introduction to Astrophysics*.
- Spinrad, H. (2005). *Galaxy Formation and Evolution*.
- Thomson, L. A. & Gregory, S. A. (1993). Dwarf galaxies in the Coma cluster. *The Astronomical Journal*, 106, 2197-2212.
- Theil, H. (1967). *Economics and Information Theory*.
- Tapia, T., Eliche-Moral, C. M. et al. (2017). Formation of S0 galaxies through mergers. *ArXiv Astrophysics e-prints*, arXiv:1706.03803
- Timmons, T. (2012). *Makers of Western Science: The Works and Words of 24 Visionaries from Copernicus to Watson and Crick*.

- Trujillo, I., Asensio Ramos, A., Rubino-Martin, J. A., Graham, A. W., et al. (2002). Triaxial stellar systems following the $r^{1/n}$ luminosity law: an analytical mass-density expression, gravitational torques and the bulge/disc interplay. *Monthly Notices of the Royal Astronomical Society*, 333, 510.
- van den Bergh, S. (1960a). A preliminary luminosity classification of late-type galaxies. *The Astrophysical Journal*, 131, 215.
- van den Bergh, S. (1960b). A preliminary luminosity classification for galaxies of type Sb. *The Astrophysical Journal*, 131, 558.
- van den Bergh, S. (1960c). A reclassification of the northern Shapley-Ames galaxies. *Publications of the David Dunlap Observatory*, 2(6), 159-199.
- van den Bergh, S. (1998). *Galaxy Morphology and Classification*.
- van der Wel, A., Bell, F. E., Holden, B. P., et al. (2010). The Physical Origins of the Morphology?density Relation: Evidence for Gas Stripping From the Sloan Digital Sky Survey. *The Astrophysical Journal*, 714, 1779-1888.
- Viral, P., van der Heyden, K., Ferrari, C., Angus, G., Holwerda, B. (2015). Morphology parameters: substructure identification in X-ray galaxy clusters. *Astronomy & Astrophysics*, 575, A127.
- Weinzirl, T., Jogee, S., Khochfar, S., et al. (2009). Bulge n and B/T in high-mass galaxies: constraints on the origin of bulges in hierarchical models. *The Astrophysical Journal*, 696, 1.
- Wells, D.C., Greisen, E. W., Harten, R. H. (1981). FITS: A Flexible Image Transport System. *Astronomy & Astrophysics Supplement*, 44, 363.
- Whitmore, B. C., Gilmore, D. M., Jones, C. (1993). What determines the morphological fractions in clusters of Galaxies? *The Astrophysical Journal*, 407, 489-509.
- Whitney, C.A. (1971). *The Discovery of Our Galaxy*.
- Willett, K. W., Lintott, C. J., Bamford, S. P., et al. (2013). Galaxy Zoo 2: detailed

- morphological classifications for 304,122 galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, 435(4), 2835-2860.
- Wirth, A. (1984). Visual and automatic classification of galaxy images. *IAU Colloquium*, 78, 129-131.
- Yasuda, N., Fukugita, M., Narayanan, V. K., et al. (2001). Galaxy Number Counts From the Sloan Digital Sky Survey Commissioning Data. *The Astronomical Journal*, 122, 1104-1124.