Universidade Federal do Rio Grande do Sul

Masters Dissertation

Quenching of star formation in galaxies up to large clustercentric distances

Author:
Maitê Mückler

Supervisor:
Drª. Marina Trevisan

A dissertation submitted in fulfilment of the requirements for the degree of Master in Physics

in the

Institute of Physics
Astronomy Department

December 11, 2023

"*Fun I love, but too much fun is, of all things, the most loathsome. Mirth is better than fun, and happiness is better than mirth. I feel that a man may be happy in this world. And I know that this world is a world of imagination and vision. I see everything I paint in this world, but everybody does not see alike. To the eyes of a miser, a guinea is far more beautiful than the Sun, and a bag worn with the use of money has more beautiful proportions than a vine filled with grapes. The tree which moves some to tears of joy is in the eyes of others, only a green thing which stands in the way. Some see nature as all ridicule and deformity, and by these, I shall not regulate my proportions; some scarce see nature at all. But to the eyes of the man of imagination, nature is imagination itself. As a man is, so he sees.*"

William Blake

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Abstract

Institute of Physics
Astronomy Department

Master in Physics

Quenching of star formation in galaxies up to large clustercentric distances

by Maitê Mückler

It is well known that galaxy properties depend on their environment, with high-density regions having a low fraction of star-forming galaxies ($f_{\rm SF}$). However, other galaxy properties correlated with $f_{\rm SF}$ are also dependent on the environment, making it difficult to disentangle what actually affects $f_{\rm SF}$. Fortunately, large-scale surveys have provided an increase in the amount of reliable extragalactic data, which can shed light on these questions. In this context, we use a complete sample of galaxies in the Sloan Digital Sky Survey Data Release 18 along with a logistic regression model to account for nine different galaxy properties across their entire range of values without binning to investigate how the $f_{\rm SF}$ varies with galaxy properties out to $R/r_{\rm vir} = 20$, where R is the clustercentric distance and $r_{\rm vir}$ is the virial radius. We also explore how the variations of $f_{\rm SF}$ with the galaxy environment and properties depend on the adopted estimate of the star formation rates (SFRs), using SFRs estimated by two different methods: one mainly from Hα emission line and other from UV/optical photometry. We find that the increase of $f_{\rm SF}$ with $R/r_{\rm vir}$ depends on the galaxy properties, with velocity dispersion and AGN activity (classified in the WHAN diagram) playing the most important role. Furthermore, when various galaxy properties are taken into account, the dependence of $f_{\rm SF}$ on $R/r_{\rm vir}$ is significantly smaller, being important only for non-AGN galaxies according to the WHAN diagram and galaxies with intermediate $\sigma_{\rm e}$ values ($1.9 \lesssim \log_{10}(\sigma_{\rm e}/[{\rm km\,s^{-1}}]) \lesssim 2.1$). Although the results obtained with different SFRs estimates are qualitatively similar, the absolute variations of $f_{\rm SF}$ with galaxy properties and environment are significantly different. Our work reveals the importance of taking several galaxy properties into account simultaneously to properly estimate the $f_{\rm SF}$ and how distinct methodologies used to measure SFRs can impact the interpretation of the results.

Key-words: Galaxy evolution; Groups of galaxies; Star formation; Data survey; Statistical methods.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Resumo

Instituto de Física
Departamento de Astronomia

Mestre em Física

Supressão da formação estelar em galáxias até grandes distâncias dos aglomerados

by Maitê Mückler

É bem conhecido que as propriedades das galáxias dependem de seu ambiente, de forma que regiões com alta densidade tendem a ter uma baixa fração de galáxias com alta taxa de formação estelar – ou fração de galáxias *star-forming* ($f_{\rm SF}$). No entanto, outras propriedades das galáxias que são correlacionadas com $f_{\rm SF}$ também dependem do ambiente, tornando difícil determinar o que realmente afeta $f_{\rm SF}$. Felizmente, o grande aumento na quantidade e na qualidade de dados extragalácticos obtidos pelos grandes levantamentos de dados recentes podem ajudar a esclarecer essas questões. Neste contexto, usamos uma amostra completa de galáxias do Sloan Digital Sky Survey Data Release 18, juntamente com um modelo de regressão logística para modelar nove diferentes propriedades de galáxias em todos os seus intervalos completes de valores, sem binagem, a fim de investigar como $f_{\rm SF}$ varia com essas propriedades até $R/r_{\rm vir} = 20$, onde R é a distância clustercêntrica e $r_{\rm vir}$ é o raio virial. Exploramos como as variações de $f_{\rm SF}$ com o ambiente juntamente com as propriedades da galáxia dependem do método adotado para estimar as taxas de formação estelar (SFRs), utilizando SFRs estimadas por dois métodos diferentes: um utilizando principalmente linha de emissão Hα e outro utilizando distribuição espectral de energia (SED) no UV/óptico. Descobrimos que a dependência de $f_{\rm SF}$ com $R/r_{\rm vir}$ depende fortemente de propriedades como dispersão de velocidade e atividade AGN (classificada utilizando o diagrama WHAN). Além disso, quando várias propriedades das galáxias são levadas em conta, a dependência de $f_{\rm SF}$ com $R/r_{\rm vir}$ é significativamente menor, sendo importante apenas para galáxias não-AGN (no diagrama WHAN) e galáxias com valores de $\sigma_{\rm e}$ intermediários ($1,9 \lesssim \log_{10}(\sigma_{\rm e}/[\mathrm{km\,s^{-1}}]) \lesssim 2,1$). Embora os resultados obtidos com diferentes estimativas de SFRs sejam qualitativamente semelhantes, as variações absolutas de $f_{\rm SF}$ com as propriedades e ambiente das galáxias são significativamente diferentes. Nosso trabalho revela a importância de utilizar várias propriedades de galáxias conjuntamente para estimar $f_{\rm SF}$ de forma adequada e como diferentes metodologias usadas para estimar SFRs podem afetar a interpretação dos resultados.

Palavras chave: Evolução de galáxias; Grupos de galáxias; Formação estelar; Levantamento de dados; Métodos estatísticos.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Press release: Novo estudo revela correlações complexas entre a atividade de formação estelar nas galáxias, as suas propriedades e o ambiente onde residem

Porto Alegre, 14 de agosto de 2023 - Uma pesquisa científica recente arremessa uma nova luz sobre as intricadas interações entre a formação estelar em galáxias e o ambiente que as cerca, fornecendo uma compreensão mais profunda das complexas relações entre esses fatores. Realizada por uma equipe de cientistas da Universidade Federal do Rio Grande do Sul, o estudo é fruto de uma análise minuciosa envolvendo uma amostra ampla, compreendendo mais de 200 mil galáxias. A investigação se vale de um modelo estatístico avançado, mergulhando nas múltiplas propriedades galácticas.

Há décadas, cientistas têm mergulhado em uma busca incansável para compreender por que certas galáxias interrompem a produção de estrelas. Este enigma cósmico tem intrigado a comunidade científica, conduzindo a uma série de investigações que lançam luz sobre esse fenômeno complexo. Características distintas das galáxias desempenham um papel crucial na identificação de sua atividade de formação estelar. A cor e a forma da galáxia emergem como sinais reveladores, com galáxias em constante formação estelar geralmente apresentando uma tonalidade azulada e uma estrutura em espiral. Por outro lado, aquelas que estão em um estado de quiescência tendem a exibir uma tonalidade avermelhada e uma configuração elíptica mais ampla. No entanto, é fundamental notar que esse processo não é tão simples, uma vez que tais características isoladas não são determinantes por si só.

A compreensão plena do status de formação estelar de uma galáxia exige uma abordagem multifacetada, explorando uma variedade de propriedades intrínsecas. A taxa de formação estelar, a presença de gás e poeira interestelar, a atividade de núcleos ativos, a distribuição das idades das estrelas e a dinâmica de movimento estelar são apenas algumas das peças desse quebra-cabeça cósmico. Portanto, uma análise minuciosa e abrangente é necessária para determinar com precisão se uma galáxia está atualmente gerando novas estrelas ou se alcançou um estado de repouso nesse aspecto fascinante.

O estudo aproveita as vastas quantidades de dados extragalácticos obtidos por meio de levantamentos recentes, e utiliza uma abordagem inovadora ao utilizar diversas características galácticas em uma análise conjunta. A equipe de pesquisa explora como a taxa de galáxias formadoras de estrelas varia com essas propriedades à medida que se afastam do centro de grupos e aglomerados galácticos. Esse olhar atento para as relações entre a posição galáctica e a formação de estrelas promete trazer *insights* profundos sobre os processos cósmicos em jogo.

Além disso, o estudo analisa como as variações na fração de galáxias formadoras de estrelas são afetadas pelos métodos utilizados para estimar as taxas de formação estelar. Duas abordagens principais estão sob escrutínio: o uso das emissões da linha de Hα, bem como a distribuição espectral de energia na faixa do ultravioleta/ótico.

Os resultados revelam que a fração de galáxias formadoras de estrelas observada está intrinsecamente ligado a elementos cruciais, incluindo a dispersão de velocidades e a atividade de núcleos ativos de galáxias (AGN). Além disso, a análise demonstra que, quando várias propriedades galácticas são consideradas em conjunto, a dependência de fração de galáxias formadoras de estrelas em relação a distância ao centro de seus aglomerados é substancialmente reduzida, tornando-se significativa principalmente para galáxias menos massivas, bem como para galáxias com determinados valores de dispersão de velocidade intermediários. Além disso, o estudo mostra que, apesar das difentes estimativas da taxa de formação estelar apontarem para a mesma direção, suas diferenças implicam profundamente na interpretação dos resultados e na ordem de magnitude em que observamos as frações de galáxias formadoras de estrelas.

Com as recentes inovações tecnológicas na astronomia, a pesquisa está posicionada para proporcionar uma visão inédita sobre o panorama cósmico. À medida que os cientistas decifram os códigos das galáxias e suas estrelas, os segredos profundos do universo podem, finalmente, começar a se desvendar.

Palavras chave: Evolução de galáxias; Grupos de galáxias; Formação estelar; Levantamento de dados; Métodos estatísticos.

Acknowledgements

I would certainly be in a very different place today if I were not Luisa's mother, so this is all because of her above all. When she came, she changed the course of my life and for the past 14 years, I have chosen to make the rightest decisions about my life because of her. And at this place of Science that I am now, I only arrived thanks to all the support and partnership of Emílio, who was the primordial ground for me to become a scientist. We have been together for eleven years and I have never seen partners as aligned and dedicated to each other as we are. The two of us are literally one true love. And speaking of love, the purest form that exists is the one that comes from pets. Oh, how can you learn to love when you allow yourself to love and be loved by them! Sofia, Odin and Chico, the gratitude I have for your existence cannot be measured. They are the most special beings to step on this planet. I only got to University thanks to my grandmother, Dolores, who allowed me to study free of charge until I managed to pass the entrance exam. She educated me, supported me, took care of me, took care of my daughter and never doubted that I was doing the right thing. If she did not have the role she has in my life, I would not be defending a master's thesis today. She has been through the ups and downs in her life and I hope that in the not-too-distant future, I will be able to provide her with the peace she has always provided in my life.

Of course, a master's thesis is produced under the surveillance of a supervisor, so evidently, this work exists beyond me. My supervisor, Dr. Marina Trevisan, was the most appropriate choice I could have made for that. I am absolutely grateful, firstly, for this project that she presented me with and for all her teachings and patience she had with me, with my mistakes, with my inexperience, anxieties and insecurities. She embraced my craving to pursue a career in Astronomy after a BA in Statistics with amazing dedication. I also would like to thank Dr. Gary Mamon for the experience he shared with me during this project and for the meaningful contributions he made to make this work better. Thank you also for your dedicated attention and availability. Also, I want to thank all the professors in the Department of Astronomy who taught me during these two years of my master's degree, especially professors Dr. Alejandra Romero, Dr. Allan Schnorr Müller and Dr. Rogério Riffel. My sincere thanks for your patience in teaching me Physics and Astronomy. To colleagues in the Department of Astronomy who also taught me a lot, in addition to providing good conversations and the necessary leisure in a routine that can easily become too stressful, my dearest gratitude.

Finally, thanks to the Universidade Federal do Rio Grande do Sul and the Department of Astronomy for providing its infrastructure and environment where quality science and education can be accomplished and to the Conselho Nacional de Desenvolvimento Científico e Tecnológico for funding this project.

Contents

Abstract **iii**

Resumo **iv**

Press release **v**

Acknowledgements **vii**

1 Introduction **1**
1.1 The bimodality of galaxy properties and relation with their environment . 4
1.2 Quenching of star formation 9
1.2.1 Internal processes (or *"mass quenching"*) 9
1.2.2 External processes (or *environmental quenching*) 11

2 This work **17**
2.1 Motivation 17
2.2 Objectives 17
2.3 Outline of the dissertation 18

3 Data and sample definition **19**
3.1 The initial galaxy sample 19
3.2 Galaxy properties 20
3.2.1 Absolute magnitudes 21
3.2.2 Stellar masses and SFRs from MPA-JHU 21
3.2.3 Stellar masses and SFRs from GSWLC 22
3.2.4 Galaxy effective radii 23
3.2.5 Velocity dispersions 24
3.2.6 Emission line fluxes and equivalent widths 24
3.3 The sample of groups and clusters 24
3.3.1 Conversion from L17 $r_{180,\mathrm{m}}$ to $r_{100,\mathrm{c}}$ 24
Relating $\overline{\rho}_{\mathrm{L17}}$ to $\overline{\rho}_{\mathrm{vir}}$ 25
Relating the concentration parameter c_{L17} to c_{vir} 26
Finding M_{vir} 27
Finding r_{vir} 27
3.4 Final galaxy sample and data 27

4 Assigning galaxies to haloes **29**
4.1 Membership assignment scheme 29
4.1.1 Galaxies at $R > 2.5r_{\rm vir}$ 30
4.1.2 Galaxies at $R \leq 2.5r_{\rm vir}$ 30
4.2 Group mass thresholds for the assignment 32
4.3 Results from assignment 33

5 Final subsamples **35**
5.1 Stellar mass completeness 35
5.2 Star-forming/Quiescent classification 36
5.3 AGN/Non-AGN classification 37
5.4 Final subsamples 38

6 Exploring the relationship between galaxy properties **41**
6.1 Correlation between variables 41
6.2 Logistic regression 51
6.3 Model-building strategy 54
6.3.1 Univariate analysis 54
6.3.2 Multivariable model comparisons 54
6.3.3 Linearity assumption 55

7 Results **61**
7.1 Logistic model best-fits 61
7.2 Analysis of the $f_{\rm SF}$ with GSWLC data 64
7.3 Analysis of the $f_{\rm SF}$ with MPA-JHU data 67
7.4 Variations of $f_{\rm SF}$ trends with halo-mass threshold for assignment ($M_{\rm a}$) . . 70

8 Discussion **75**
8.1 The dependency with the clustercentric distance 75
8.2 The dependency with galaxy velocity dispersion 76
8.3 The dependency with stellar mass 76
8.4 The dependency with *AGN* activity 77
8.5 Effects of pre-processing 78
8.6 Differences between GSWLC and MPA-JHU measurements 78

9 Conclusion and perspectives **81**

A MSG Query **85**

Bibliography **87**

List of Figures

1.1 Hubble morphological diagram . . . 2
1.2 Illustrative representation of a halo merger tree . . . 3
1.3 Colour-magnitude diagram . . . 5
1.4 Colour histogram by morphology . . . 6
1.5 Morphology-density relation . . . 7
1.6 sSFR-stellar mass diagram . . . 8
1.7 sSFR versus velocity dispersion diagram . . . 11
1.8 sSFR versus $\Sigma_{*,1\mathrm{kpc}}$ . . . 12
1.9 Sketch of the evolution of massive galaxies . . . 13
1.10 Quenched fraction of galaxies as a function of local-environment overdensity and galaxy stellar mass . . . 14
1.11 Mean quenched fraction by local-environment overdensity and halo mass 15
1.12 f_{SF} as a function of clustercentric distance . . . 16

3.1 Sky coverage of the SDSS Legacy Survey . . . 20
3.2 SDSS and GALEX filters . . . 22
3.3 Mean random error of SED fitting SFR and stellar mass for GSWLC catalogues . . . 23

4.1 Real-space distance in terms of log projected distance and absolute line-of-sight velocity . . . 32
4.2 Diagram illustrating the two group-mass thresholds in the assignment scheme . . . 33
4.3 Galaxies assigned to a given halo for different values of M_{a} . . . 34

5.1 Mass completeness limit for the two samples, $z_{\mathrm{max}} = 0.03$ and $z_{\mathrm{max}} = 0.1$. 36
5.2 sSFR vs $M_\star$ diagrams . . . 37
5.3 WHAN diagram . . . 38

6.1 Correlation plot for the variables. . . . 48
6.2 Scatter plot for numerical variables . . . 49
6.3 Pearson residuals of χ^2 . . . 51
6.4 Smoother scatter plots of explanatory variables . . . 58

7.1 Importance of explanatory variables in the four fits . . . 63
7.2 f_{SF} predicted by the logistic model compared to observations: `GSWLC01` dataset and $\log_{10}(M_{\mathrm{a}}/\mathrm{M}_\odot) = 12.3$ . . . 66
7.3 f_{SF} predicted by the logistic model compared to observations: `GSWLC03` dataset and $\log_{10}(M_{\mathrm{a}}/\mathrm{M}_\odot) = 12.3$ . . . 67

7.4 f_{SF} predicted by the logistic model compared to observations: MPAJHU01 dataset and $\log_{10}(M_a/M_\odot) = 12.3$ 69
7.5 f_{SF} predicted by the logistic model compared to observations: MPAJHU003 dataset and $\log_{10}(M_a/M_\odot) = 12.3$ 70
7.6 f_{SF} predicted by the logistic model compared to observations: GSWLC01 dataset and $\log_{10}(M_a/M_\odot) = 13.0$ 71
7.7 f_{SF} predicted by the logistic model compared to observations: MPAJHU01 dataset and $\log_{10}(M_a/M_\odot) = 13.0$ 72
7.8 f_{SF} predicted by the logistic model compared to observations: GSWLC01 dataset with no Lim *et al.* (2017) centrals and $\log_{10}(M_a/M_\odot) = 13.0$ . . . 73
7.9 f_{SF} predicted by the logistic model compared to observations: MPAJHU01 dataset with no Lim *et al.* (2017) centrals and $\log_{10}(M_a/M_\odot) = 13.0$ . . . 74

List of Tables

4.1 Final group samples 34

5.1 Summary of the galaxy final subsamples and datasets 39

6.1 Summary of variables used in this work 42
6.2 Summary statistics of numerical variables 44
6.3 Contingency tables for categorical variables for GSWLC sample 45
6.4 Contingency tables for categorical variables for MPA-JHU 46
6.5 χ^2 test of independence information and effect size 50
6.6 Summary of the univariate logistic models 55
6.7 Summary of the first trial multivariate logistic model 56
6.8 Summary of the second trial multivariate logistic model 56
6.9 Percentage change in the coefficients ($\Delta\beta$ %) between the multivariate model 1 and the multivariate model 2 56
6.10 Summary of the third trial multivariate logistic model 57
6.11 Percentage change in the coefficients ($\Delta\beta$ %) between the multivariate model 2 and the multivariate model 3 59
6.12 Summary of the fourth trial multivariate logistic model 59
6.13 Percentage change in the coefficients ($\Delta\beta$ %) between the multivariate model 3 and the multivariate model 4 59

7.1 Summary of the best-fit logistic model for the four samples 62
7.2 Fit MPAJHU003 62
7.3 Optimal cutoff and confusion matrix for each fit 64
7.4 Model accuracy 64

List of Abbreviations

ΛCDM	Λ Cold Dark Matter
ISM	Interstellar Medium
IGM	Intergalactic medium
SFG	Star-forming galaxy
sSFR	Specific star formation rate
MS	Main sequence of SFGs
SMBH	Supermassive black hole
ETG	Early-type galaxy
LTG	Late-type galaxy
CMD	Colour-magnitude diagram
RS	Red sequence
BC	Blue cloud
DP	Density profile (*of dark matter*)
NFW	Navarro-Frenk-White (*density profile*)
SDP	Surface number density profile
SAM	Semi-analytical model
PPS	Projected phase space
IQR	Interquartile range
GLM	Generalised linear model
MLE	Maximum Likelihood Estimation

I dedicate this dissertation to the brave mother scientists who had the courage to pursue this career. Even if society attempts to stop you, do not be afraid to face whatever it takes to become what you want to be.

Chapter 1

Introduction

We currently understand galaxies as systems composed of stars, gas, dust, and dark matter held together by gravity, and that can be considered the building blocks – and tracers – of the large-scale structure of the Universe. But in the early 20th century, our perception of the Universe was limited to the existence of just our own galaxy, the Milky Way, and these objects that now we know as galaxies were thought to be unresolved objects in the Milky Way called *nebulae*. However, in 1923, while observing the *Andromeda Nebula*, an object that for ages was seen as just an extended cloud of light, the American astronomer Edwin Hubble made one of the first observations of resolved stars and the observation of Cepheid variables, which allowed measuring the distance to the now-recognised Andromeda Galaxy. These observations led to the identification of the extragalactic objects that today we call *galaxies*, a discovery that this year marks the completion of 100 years.

Not long after, in 1926, Edwin Hubble, in his work *"Extragalactic nebulae"* (Hubble, 1926), made the first classification of a large number of galaxies based on their observed morphologies, starting the important morphological classification scheme of galaxies, the Hubble sequence, published in the form of a diagram in Hubble (1936). In this fork-shape diagram, galaxies with a predominantly elliptical structure lie in the fork handle, while galaxies with a spiral shape are in the tines, and in the connection between them are the lenticular galaxies (S0), as shown in Figure 1.1. Historically, the galaxies on the left side of Hubble's diagram (ellipticals and lenticulars) were called *early-type galaxies* (ETGs) and those on the right (spirals) as *late-type galaxies* (LTGs), although this nomenclature should not be considered as temporal. Later, the irregular galaxies would be added to the diagram and sub-classifications of ellipticals and spirals (de Vaucouleurs, 1959). These initial findings demonstrated the existence of a diverse population of galaxies, each with unique properties.

Then, in 1929, Edwin Hubble measured the distances to these objects (Hubble, 1929), which led to distance-redshift relations for galaxies, indicating the expansion of the Universe. This brings on the realisation that the Universe is much larger than previously thought, making clear that galaxies played a central role in its structure and evolution. With the advent of large-scale surveys in the late 20th century, astronomers were allowed to study the statistical properties of galaxies and their large-scale distribution in the Universe. This contributed to the current standard cosmological model Lambda Cold Dark Matter (ΛCDM) being proposed in the 1980s (Blumenthal *et al.*, 1984; Turner *et al.*, 1984) and confirmed with the supernovae Ia observations in the late 1990s

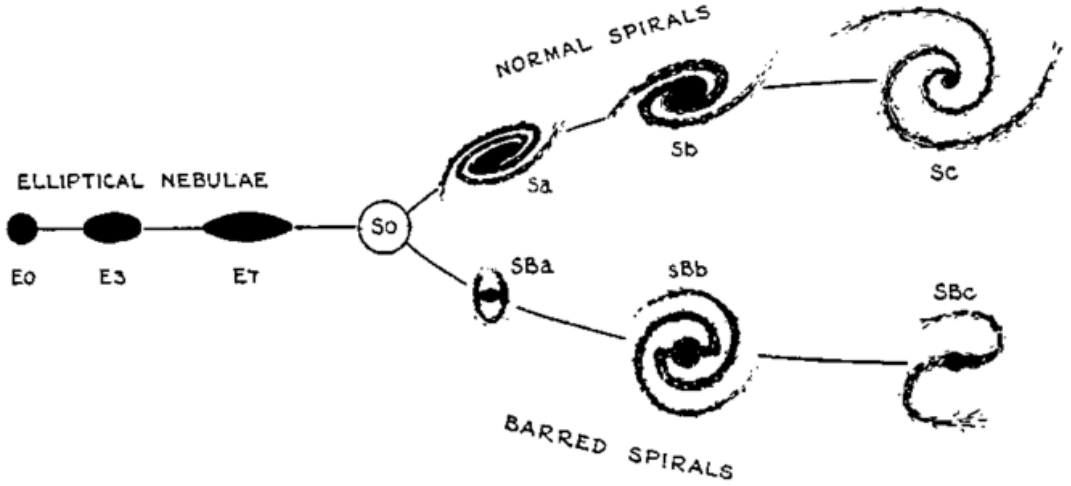

FIGURE 1.1: Hubble morphological diagram as published in Hubble (1936). Here, "nebulae" stands for "galaxies". The index integer number on Elliptical (E) types represents their degree of ellipticity on the sky. Normal spirals (S) are indexed by letters so that Sa has prominent bulges and tightly wound arms, while type Sc has the least prominent bulges and loosely wound arms. The Barred Spirals (SB) follow the same rule. Finally, the Lenticulars (So) display prominent bulges but no spiral arms.

(Perlmutter *et al.*, 1999), establishing the basic content of our Universe which imposed its flat geometry and accelerated expansion. Nowadays, it is estimated that the Universe is composed of approximately 2/3 of dark energy ($\approx$ 69%) and 1/3 of matter ($\approx$ 26% cold dark matter and $\approx$ 5% baryonic, Planck Collaboration *et al.*, 2020).

In the ΛCDM framework and the inflation theory, quantum fluctuations in the early Universe created tiny overdensities that steadily increased and collapsed as the Universe expanded, resulting in halos. Dark matter collapses first as it does not interact with radiation, but after recombination ($z \sim 1100$), the baryons can eventually cool and form structures in these dark matter halos. Fluctuations on smaller scales collapse first and then coalesce over time to form larger halos that would be sites of galaxy formation (Gunn and Gott, 1972; Press and Schechter, 1974; Gott and Rees, 1975; White and Rees, 1978). This process is known as the hierarchical model of galaxy formation and could be represented by a merger tree as the one shown in Figure 1.2. In this schematic representation, a merger occurs at each branch connection and at any given time, the main progenitor halo is the most massive among the progenitors at that time. In these halos, gas will sink, collapse and cool to form new galaxies, and the mechanism of accretion and merging continues until the present day. Galaxies gather together, forming groups of galaxies, and these groups bring more galaxies due to their gravitational attraction, repeatedly colliding and merging, eventually leading to the larger groups and clusters we observe today. In this paradigm of galaxy formation, it is clear that there is a substantial dependence of galaxy properties on the environment.

The primary mechanism driving the cycle of evolution of galaxies is their star formation activity. Through the production of new stars, a galaxy will enrich the interstellar medium (ISM) and the intergalactic medium (IGM) with the chemical elements formed in the core of stars (stellar nucleosynthesis), affecting the properties and

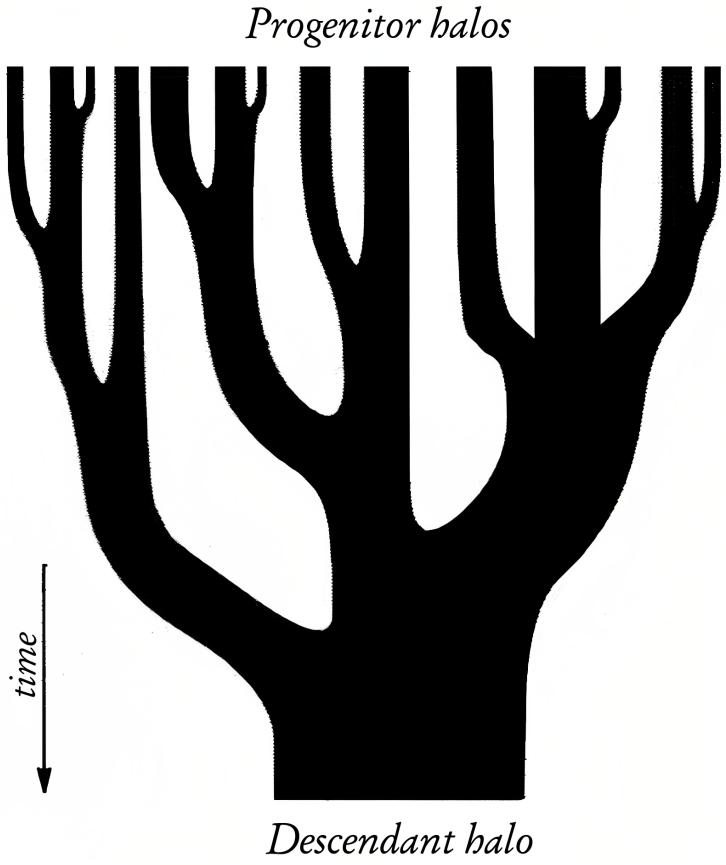

FIGURE 1.2: Illustrative representation of a halo merger tree. The time increases from top to bottom. The descendant halo (bottom of the figure, late times) is the trunk of the tree, and its initial progenitors (top of the figure, early times) are the thinnest branches. Halo mergers correspond to branch connections. A horizontal line represents each halo at any given time, and its length corresponds to the halo mass. Since there is no mass loss in this scenario and it is assumed that the halo only grows through mergers, the sum of the line lengths at any given time remains constant. Adapted from Lacey and Cole (1993).

evolution of the next generations of stars. To fuel the star formation, a galaxy must have enough cold gas ($T \lessapprox 30$ K, Kennicutt and Evans, 2012), allowing a sufficiently high density to form molecular clouds, which will remain in hydrostatic equilibrium as long as the kinetic energy of the gas pressure is in balance with the potential energy of the internal gravitational force. However, if a cloud is massive enough (above Jeans mass), the gas pressure is insufficient to support it. Then, the cloud will undergo gravitational collapse, forming dozens to tens of thousands of stars. Galaxies that are currently forming new stars are called *star-forming galaxies* (SFGs). In contrast, galaxies that are in a phase (permanent or not) of suppression of star formation are called *passive or quiescent galaxies*. Notably, some important characteristics of galaxies seem to be related to whether they are forming stars or not.

1.1 The bimodality of galaxy properties and relation with their environment

The fact that the distributions of some galaxy properties are bimodal has been demonstrated even before the advent of large-scale surveys. In this context, a commonly studied bimodal property is galaxy colour. Galaxies are often classified as "blue" or "red" based on their optical colours (Strateva *et al.*, 2001), with these colours presenting a relationship with the stellar mass ($M_\star$) of galaxies, where red galaxies are relatively more massive than blue galaxies (Baldry *et al.*, 2004). These populations were further denominated as the *red sequence* (RS) and the *blue cloud* (BC), as shown in the colour-magnitude diagram (CMD) in Figure 1.3. The sparse and supposedly transitional population between them was called "green valley" (Bell *et al.*, 2004). Also, it was found that BC galaxies have a higher star-formation rate (SFR), there is, the total mass of stars formed per year, than galaxies in the RS (Brinchmann *et al.*, 2004, Salim *et al.*, 2007). In that sense, the colours of galaxies reflect their dominant stellar populations, whereupon blue galaxies have bluer colours due to the presence of young, hot, and massive stars, while red galaxies exhibit redder colours due to older stellar populations (Humason, 1936; Hubble, 1936). This characteristic is also linked to the environment, as numerous works have already demonstrated (Lewis *et al.*, 2002; Gómez *et al.*, 2003; Kauffmann *et al.*, 2004; Zehavi *et al.*, 2005; Weinmann *et al.*, 2006; Sheth *et al.*, 2006; Tinker *et al.*, 2008; Skibba *et al.*, 2009): RS galaxies tend to be found in dense environments, while BC galaxies tend to be found in less dense regions or isolated.

Another bimodal galaxy property revealed by observations is their morphology (Strateva *et al.*, 2001; Driver *et al.*, 2006) with a distinct mode for ETGs and another for LTGs. Also, it has been known, at least since the late 1930s, that the morphology correlates with the colours of galaxies (Humason, 1936; Hubble, 1936). The greater part of LTGs lies in the BC, and not surprisingly, the majority of ETGs are found in the RS (Strateva *et al.*, 2001) as we can see in Figure 1.4. Also, morphology shows a relation with the density of the environment. This relation was first quantified by Oemler (1974) and Davis and Geller (1976) and establishes that ETGs are more easily found in denser environments. At the same time, LTGs reside mostly in less dense environments and in the field. This connection was reinforced by the nominal work of Dressler (1980), which presented for a sample of about 6000 galaxies in 55 clusters the morphology-density relation in a diagram form, illustrated in Figure 1.5. Although these characteristics are

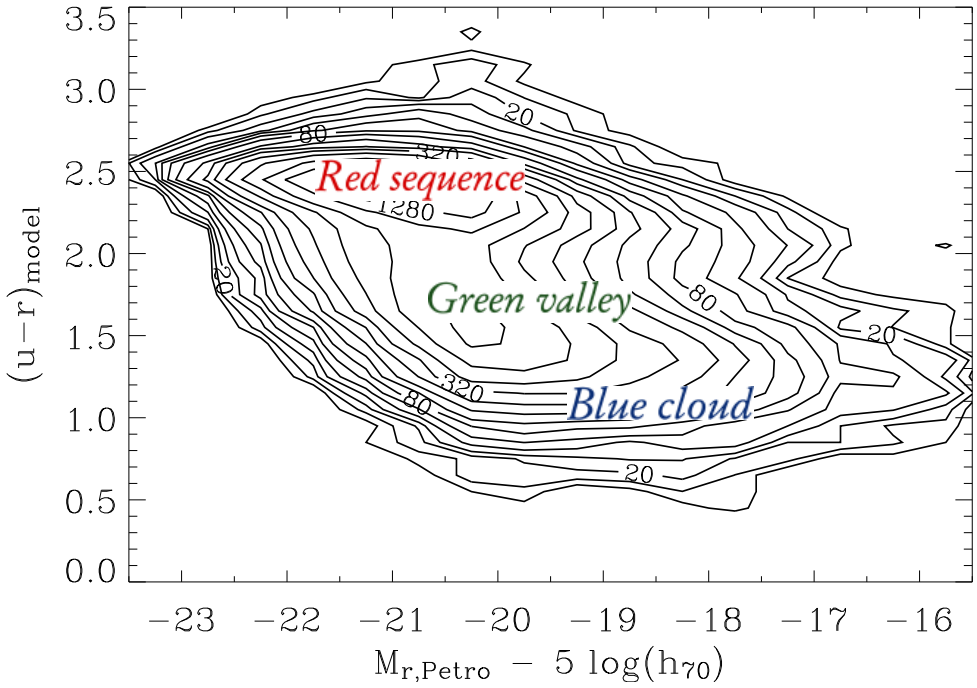

FIGURE 1.3: Colour-magnitude diagram. Observed bivariate distribution of the sample of Baldry *et al.*, 2004. The y-axis is the rest-frame colour (u - r) where *model* stands for a fit to the flux using the best fit of a de-Vaucouleurs and an exponential profile. The x-axis is the Petrosian, absolute magnitude in the r-band (which is the summed flux in an aperture that depends on the surface-brightness profile of the object) with Hubble constant $H_0 = (h_{70})$ 70 km s^{-1} Mpc^{-1}. The contours are determined for galaxy number counts in 0.1 colour × 0.5 mag bins (with a total of 66,846 galaxies). The contour levels are on a logarithmic scale, starting at 10 and doubling every two contours. Adapted from Baldry *et al.* (2004).

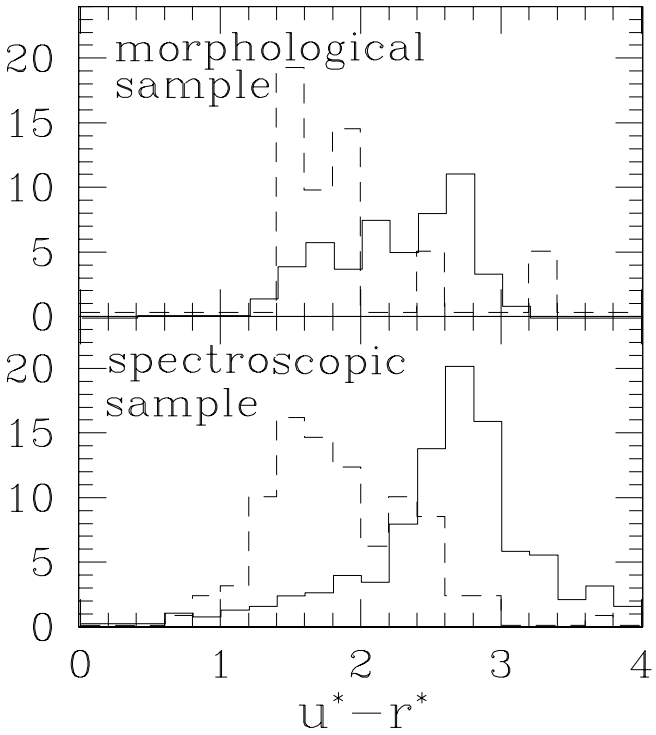

FIGURE 1.4: Colour histograms (u-r) of early-type (solid lines) and late-type (dashed lines) galaxies made from a sample where the morphologies were classified visually (upper panel) and spectroscopically (bottom). Adapted from Strateva *et al.* (2001).

interconnected, they are scattered, i.e., there is a residue from ETGs found in the BC and LTGs populating the RS (Salim *et al.*, 2007; Schawinski *et al.*, 2007; Bamford *et al.*, 2009; Skibba *et al.*, 2009) indicating that colour and morphology are not entirely dependent on each other.

Given all the bimodal relations between the galaxy properties presented so far, one might expect that there will also be a bimodal relationship between these properties and the galaxy's SFR. This is indeed correct, and the SFR also shows dependence on the galaxy's environment. First, observations have been present for a long time in a reasonably tight and linear sequence of SFGs on an SFR versus stellar mass diagram called *main sequence* of SFGs (MS) (Brinchmann *et al.*, 2004; Knobel *et al.*, 2015). Figure 1.6 presents this sequence above the black line that divides the MS and the "quenched sequence". In this figure, the specific star-formation rate (sSFR), which is the star formation per unit mass, was used as an indicator of a star formation history so that galaxies with a larger fraction of recent star formation will have a higher value of the sSFR. Additionally, it has been well-known that the environmental impact on SFR is even stronger than that on morphology (Kauffmann *et al.*, 2004; Blanton *et al.*, 2005; Ball *et al.*, 2008; Bamford *et al.*, 2009), indicating a primary dependence between environment and star formation before morphological transformation.

Finally, the internal kinematics of galaxies, which basically describes how the material in galaxies balances its self-gravity, presents two distinct behaviours: one when the material moves in roughly circular and coplanar orbits, being *rotationally supported*, and another when the material maintains a stable equilibrium by randomly orientated orbits, being *pressure supported*. These two characteristics are closely related to galaxy

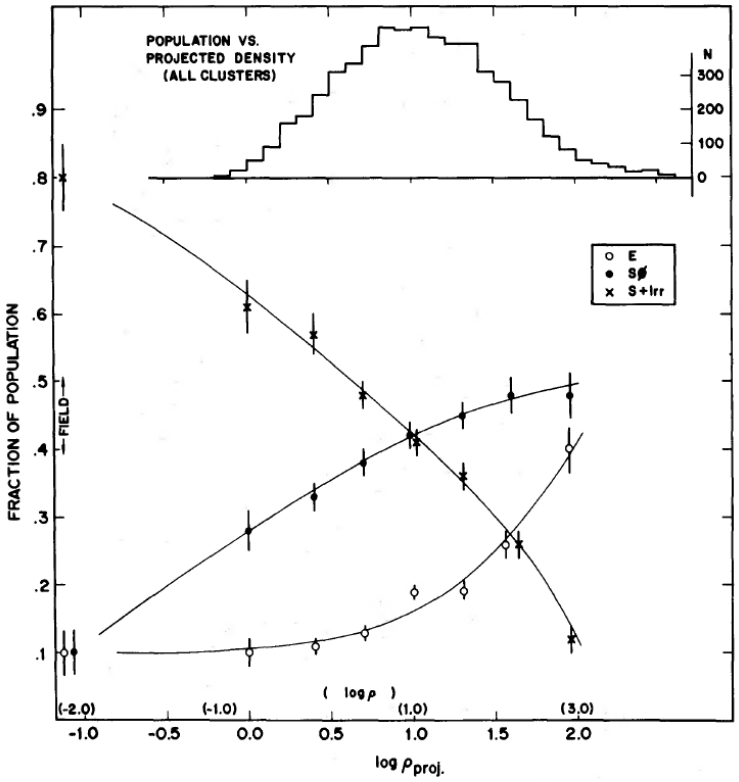

FIGURE 1.5: Morphology-density relation published by Dressler (1980). The fraction of E (elliptical), SO (lenticular), and S (spiral) + I (irregular) galaxies as a function of the log of the projected density in galaxies Mpc^{-2}. The upper histogram shows the number distribution of the galaxies over the bins of projected density.

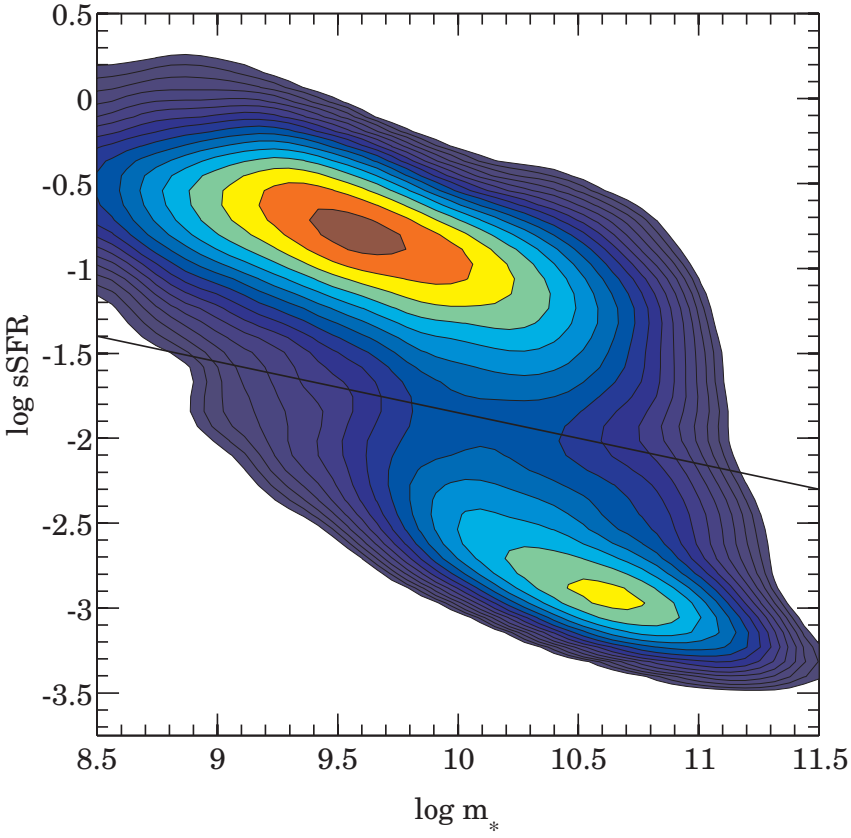

FIGURE 1.6: sSFR-stellar mass diagram from Knobel *et al.* (2015). The colours refer to the (unweighted) number density of galaxies on a logarithmic scale. The black line is visually chosen to separate between MS and the quenched sequence.

morphology, where flattened, disc galaxies are primarily supported by rotation, whilst spheroidal or elliptical galaxies reveal to be mostly, but not exclusively, supported by pressure (van de Sande *et al.*, 2017; van de Sande *et al.*, 2018; Graham *et al.*, 2018). Furthermore, it is well known that both traits have a strong correlation with other galaxy attributes, for example, mass (Dressler, 1980; Baldry *et al.*, 2006; Kelvin *et al.*, 2014), colour (Blanton *et al.*, 2003; Driver *et al.*, 2006), and SFR (Kennicutt, 1983; Kauffmann *et al.*, 1993), demonstrating that they store information about the formation process of galaxies.

The interconnection between all these galaxy properties and their relationship with the environment is fully connected to how these systems evolve. This evolution can happen through different paths, and the mechanisms that make a galaxy eventually stop forming stars are very likely to depend on the path that they took, producing galaxies with different properties and residing in different environments.

1.2 Quenching of star formation

The processes associated with the suppression of star formation in galaxies generally prevent the gas from cooling, remove gas replenishment from the external medium, remove hot gas from the galaxy or, more drastically, remove cold gas from the galaxy. Thus, the galaxy will enter a phase (permanent or not) of suppression of star formation, customarily called *quenching of star formation*. Recent observations and studies have shown that the suppression of star formation can occur in different ways depending on the galaxy's intrinsic characteristics (stellar mass, velocity dispersion, gas fraction, etc.) and extrinsic characteristics (merger history, environment density, halo mass, etc.).

Furthermore, a galaxy being the *central galaxy* of a dark matter halo or the *satellite* of this halo may, in a certain way, restrict the possible mechanisms for its quenching. In this section, we will present the observational pieces of evidence related to the suppression of star formation in galaxies and the mechanisms that may lead to this, first discussing the internal processes, also known as *mass quenching*, and then the external processes related to the environment where the galaxy resides, the *environmental quenching*.

1.2.1 Internal processes (or *"mass quenching"*)

As discussed in the previous section, the relations between galaxy properties, their sSFRs and stellar masses suggest that internal processes are important in quenching star formation. An important work that studied this relationship was the study of Peng *et al.* (2010). The authors constructed a phenomenological quenching model where internal and external processes independently influenced the fraction of quiescent galaxies. In this model, the internal processes were the quenching channel that depends on the galaxy's stellar mass, fully controlling the shape of the Press-Schechter mass function (Press and Schechter, 1974) of the star-forming galaxies, which predicts the number of objects of a certain mass within a given volume of the Universe. These findings with respect to the stellar mass led to the labelling of the internal processes as *"mass quenching"*. However, this nomenclature does not imply that the stellar mass is *causing* the quenching, only that it is somehow related to the internal mechanisms causing quenching. A more contemporary terminology could be "self-quenching" or "internal-quenching", but for

historical reasons, in this text, we will use the expression mass-quenching to name the internal processes.

To explain mass-quenching, several mechanisms are usually invoked, such as supernova explosion, i.e., the violent death of a star, and the feedback of an active galactic nucleus (AGN), by which the activity of the supermassive black hole (SMBH) at the centres of galaxies can release enormous amounts of energy in various forms. In the event of supernova explosions, the gas outflows driven by stellar winds could lead the galaxy to become quenched (Larson, 1974; Dekel and Silk, 1986). Also, it could be responsible for the correlation observed between the mass quenching efficiency and SFR, as proposed by Darvish *et al.* (2016). Moreover, in the AGN feedback, the central SMBH can eject or heat the surrounding gas, preventing its collapse and subsequent star formation (Croton *et al.*, 2006; Fabian, 2012; Fang *et al.*, 2013). Strong relationships can be seen between a galaxy's bulge mass, total stellar mass, the stellar velocity dispersion (σ), and black hole mass (Magorrian *et al.*, 1998; Marconi and Hunt, 2003; Häring and Rix, 2004), which implies that variations in a galaxy's SFR and structure may also be connected to black hole activity.

Nonetheless, recent studies (e.g. Teimoorinia *et al.*, 2016; Bluck *et al.*, 2016; Bluck *et al.*, 2020a; Bluck *et al.*, 2020b; Bluck *et al.*, 2022; Brownson *et al.*, 2022) have shown that, rather than stellar mass, the stellar velocity dispersion is the most important galaxy property that separates star-forming and quenched galaxies. This relation can be clearly seen in Figure 1.7, where most galaxies with low values of σ have higher values of sSFR and, on the opposite, those with high values of σ have low values of sSFR. Since σ is related primarily to a galaxy's bulge/spheroidal component, this relation suggests a deep connection between morphological and kinematic structure and the mechanisms leading to the quenching. This relation with σ could also be interpreted as a probable signature of cumulative AGN feedback. This interpretation arises from the well-known relation between σ and the SMBH mass (Ferrarese and Merritt, 2000; Gebhardt *et al.*, 2000), which, in turn, is a proxy for the amount of AGN activity that the galaxy had integrated over time, particularly in the maintenance (radio) mode (Bluck *et al.*, 2020b).

The velocity dispersion σ is correlated with the concentration of mass at the centre of galaxies, the stellar surface density in the inner kpc ($\Sigma_{\star,1\mathrm{kpc}}$) and, indeed, several works find that $\Sigma_{\star,1\mathrm{kpc}}$ strongly correlated with quenching (Strateva *et al.*, 2001; Blanton *et al.*, 2003; Kauffmann *et al.*, 2003; Wuyts *et al.*, 2011; Woo *et al.*, 2015; Woo *et al.*, 2017). As shown in Figure 1.8 of Woo *et al.* (2017), galaxies with high SFR appear to have less dense central regions than galaxies with low SFR by a factor of 0.2–0.3 dex, whether in the field or as satellites. But satellite galaxies have a higher fraction of quenched galaxies than the field at all masses. These effects are strongest for the lowest masses and decrease in significance for the most massive galaxies. This elucidates the fact that more important than the total value of the stellar mass is how it is distributed, especially in the central region.

To explain the relation between $\Sigma_{\star,1\mathrm{kpc}}$ and the quenching, it has been suggested that galaxies experience a compaction event before quenching (Fang *et al.*, 2013; Tacchella *et al.*, 2015; Tacchella *et al.*, 2016). The "compaction" of a galaxy refers to a process that increases its central density. It involves the redistribution of matter within the galaxy, causing the central region to become denser and more packed. It is believed that compaction events can occur through wet mergers (two gas-rich galaxies collide and merge) because the combination of gravitational interactions can cause the gas and stars to be funnelled towards the centre, resulting in a compact and dense central region in the

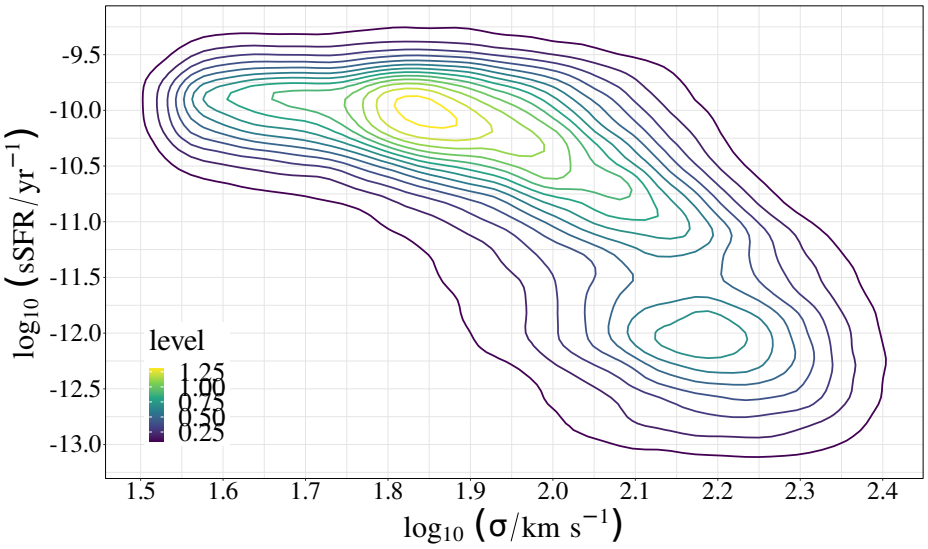

FIGURE 1.7: Specific star formation rate as a function of galaxy velocity dispersion.

newly formed galaxy. Furthermore, violent disc instabilities are believed to trigger the formation of dense clumps or bars in the central region, leading to compaction. During these compaction events, the gas and stars in the outer regions of the galaxy are often driven inward, increasing the density and concentration of mass in the central bulge or nucleus. This process can have implications for the subsequent evolution of the galaxy, influencing its SFR, morphology, and other properties.

According to observations, the mass profiles of quenched galaxies show a denser inner region (1 kpc) than SFGs (Fang *et al.*, 2013). At any time, this central stellar density grows along with stellar mass while the galaxy is in the MS (Hopkins *et al.*, 2009; Feldmann *et al.*, 2010; van Dokkum *et al.*, 2015). The high inner densities seen in quenched galaxies today were already in place by $z \approx 2$ (Patel *et al.*, 2013; Barro *et al.*, 2015; Tacchella *et al.*, 2015), which may suggest that if there is a connection between quenching and high inner density, it likely occurs at high-z. In the sketch shown in Figure 1.9 from Tacchella *et al.* (2015), we can see what would be the evolution of massive galaxies that would undergo compaction in high-z, causing quenching in their centres proceeding inside-out until the present-day.

1.2.2 External processes (or *environmental quenching*)

The term "galaxy environment" is a very general concept that has been used in the literature in diverse ways, with different environment measurements depending on the authors' approach (Muldrew *et al.*, 2012), from local overdensities to the mass of the host halo. Regardless of the type of measurement used, the galaxy environment is intrinsically linked with the environment's *density* where the galaxy resides. Dense environments are conducive to interactions, encounters and exchange of material between galaxies, which can affect their star formation and, consequently, their evolution. It is not by chance that the environment correlates with various properties of galaxies, as already discussed.

As mentioned earlier, stellar mass and velocity dispersion are extremely strong indicators of quenching, but even with these variables fixed, there is a residue left over. In

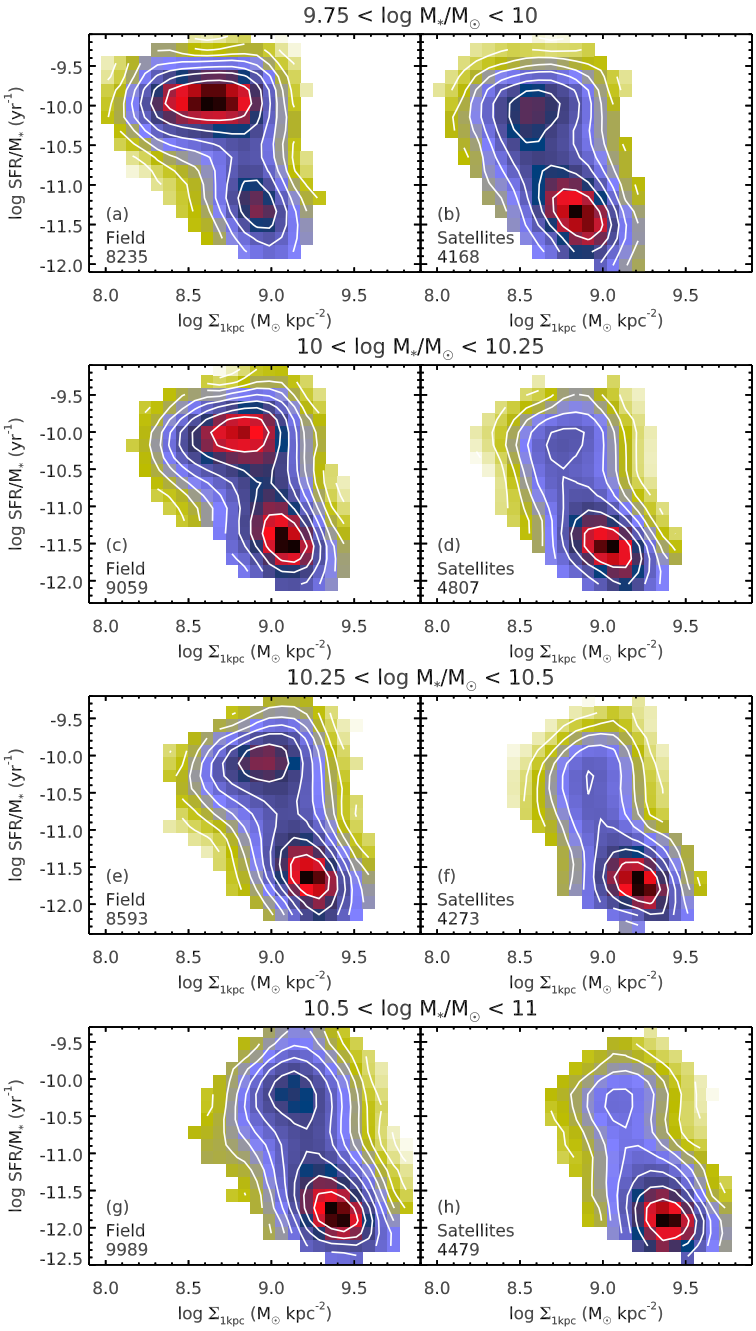

FIGURE 1.8: sSFR versus $\Sigma_{*,1\mathrm{kpc}}$ comparing the field and satellite galaxies in four bins of stellar masses. The contours represent the logarithmic number density in each panel and are separated by 0.2 dex. The colour scale is normalised such that dark red represents the highest number density in the panel. The sample size is indicated in the bottom left. The figure shows that quiescent galaxies tend to have higher values of $\Sigma_{*,1\mathrm{kpc}}$ in all bins of stellar mass and regardless of the environment (i.e. if they are field or satellite galaxies). Adapted from Woo *et al.* (2017).

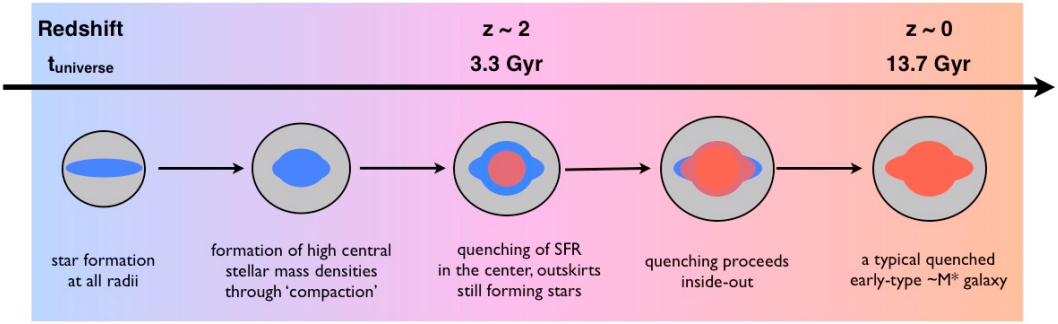

FIGURE 1.9: Sketch of the evolution of massive galaxies Tacchella *et al.* (2015). In this scenario, the total stellar mass and bulge mass grow synchronously in $z \sim 2$ MS galaxies, and quenching is concurrent with their total masses and central densities approaching the highest values observed in massive spheroids in today's Universe.

Figure 1.10 taken from Peng *et al.* (2010), one can see that, even for a fixed stellar mass, the galaxy's fraction of quiescent galaxies changes quadratically as a function of the local environment (the overdensity). Furthermore, Figure 1.11 of Woo *et al.* (2013) shows that even for a fixed overdensity, the fraction of quenched galaxies changes depending on the host halo mass ($M_{\rm h}$), which the total mass (dark and baryonic) of the group. Although there are several environment markers, tracing both local and global environments, where we can observe quiescence, the related physical processes in each regime are not totally clear. Many processes occurring in groups/clusters of galaxies can indeed affect the gas of galaxies. Next, we describe some of the most well-known environmental mechanisms that can lead to the suppression of star formation in galaxies.

According to the *Starvation/Strangulation* scenario, star formation in galaxies is quenched because the inflow of gas from the IGM is suppressed and, as a consequence, star formation can only continue for a limited period of time using the available gas in the galaxy (Larson *et al.*, 1980; Balogh *et al.*, 2000; Balogh and Morris, 2000; Feldmann *et al.*, 2010; Peng *et al.*, 2015). In this case, the metallicity (i.e. the abundance of elements heavier than helium) should increase rapidly due to the lack of dilution from the inflow of external gas. In Peng *et al.* (2015), the analysis of stellar metallicities of 26,000 galaxies in the local universe revealed that, for low/intermediate mass galaxies, passive systems are, on average, much more metal-rich than their star-forming progenitors, which is in agreement with the strangulation scenario. This result supports the idea that "strangulation" is the main quenching mechanism for low/intermediate mass galaxies ($\log_{10} M_\star < 11$).

Furthermore, the circumgalactic and ISM gas can be stripped away when galaxies move through the medium of groups and clusters (intracluster medium, ICM), in what is known as *ram-pressure stripping*. In such a scenario, a gas-rich galaxy moving through the ICM can lose its gas content through interaction with the environment. The rate of star-forming gas loss is related to several factors, but mostly the galaxy's mass and its velocity in the medium (Gunn and Gott, 1972; Abadi *et al.*, 1999; Quilis *et al.*, 2000).

Another way for galaxies to lose gas due to the interaction with the environment is from the cumulative effect of tidal interactions from the numerous high-speed encounters with smaller galaxies. These heat up the galaxy's outer gas content and thus slow down its

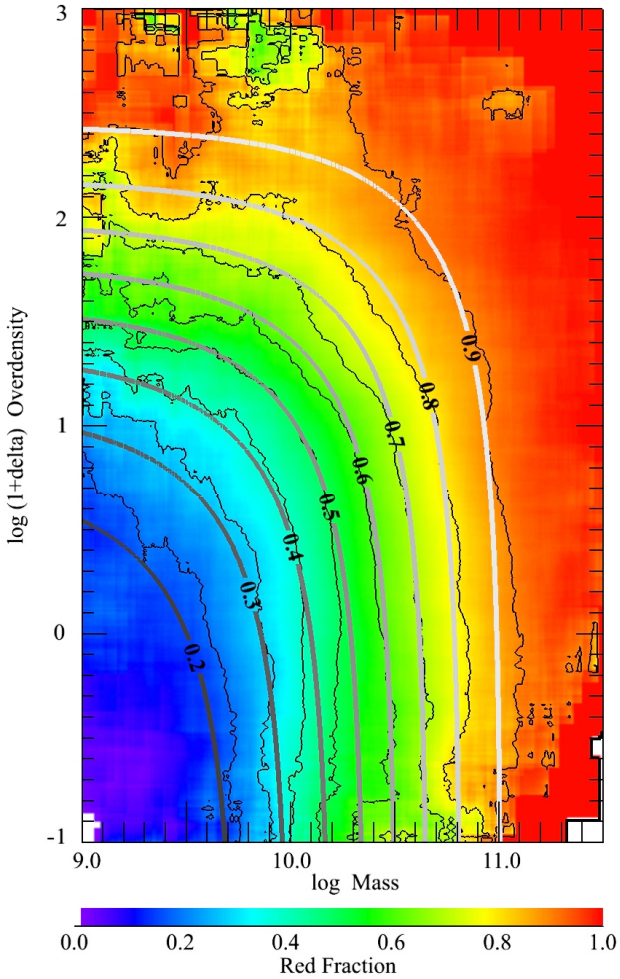

FIGURE 1.10: Quenched fraction of galaxies as a function of local-environment overdensity (delta) and galaxy stellar mass by Peng *et al.* (2010). Here the quenched fraction was called the "red fraction". The overdensity delta in the y-axis was measured through the distance to the 5th nearest neighbour.

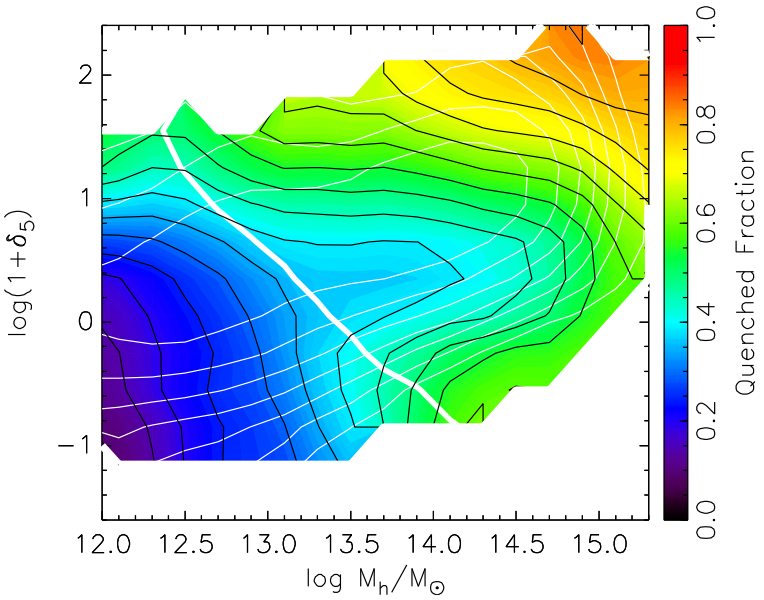

FIGURE 1.11: Mean quenched fraction by local-environment overdensity (δ_5) and halo mass (M_h) for satellite galaxies by Woo *et al.* (2013). The overdensity δ_5 in the y-axis was estimated using the distance to the 5th nearest neighbour.

star formation processes, in events known as *Tidal stripping or Harassment* (Farouki and Shapiro, 1981; Moore *et al.*, 1996; Smith *et al.*, 2010).

Even on the outskirts of a massive structure, quenching mechanisms with environmental origins can occur. *Pre-processing* is a process in which galaxies in groups, infalling on a large group/cluster, may already be suffering some kind of environmental effect, being environmentally impacted at significant distances from the galaxy cluster's centre (Zabludoff and Mulchaey, 1998). In Figure 1.12, we can see that, even at large clustercentric distances, the fraction of star-forming galaxies is below that observed for field galaxies. The pre-processing has been considered as an explanation for reduced star-forming fractions beyond the virial radius by a number of authors (e.g., Berrier *et al.*, 2009; McGee *et al.*, 2009; Mahajan *et al.*, 2011, Lu *et al.*, 2012; Wetzel *et al.*, 2012, Haines *et al.*, 2015).

Although there are several well-known and observed environmental effects, it is not clear which of these processes dominate and when and how exactly they operate. In addition, there is a lack of studies focusing on determining the extent to which galaxies are affected by the group/cluster environment and the role of pre-processing in small groups along the large-scale filamentary structure (i.e., galaxies have their sSFR affected by denser local environments prior to their infall into massive clusters; see, e.g. Ribeiro *et al.*, 2009, Bahé *et al.*, 2013).

In summary, quenching mechanisms can be diverse, interconnected and complex. To unveil different physical processes responsible for the quenching of star formation in galaxies, it is necessary to analyse several galaxy properties simultaneously to account for the correlation between them. In this context, in the next Chapter, we present the objectives of this work and the outline of this dissertation.

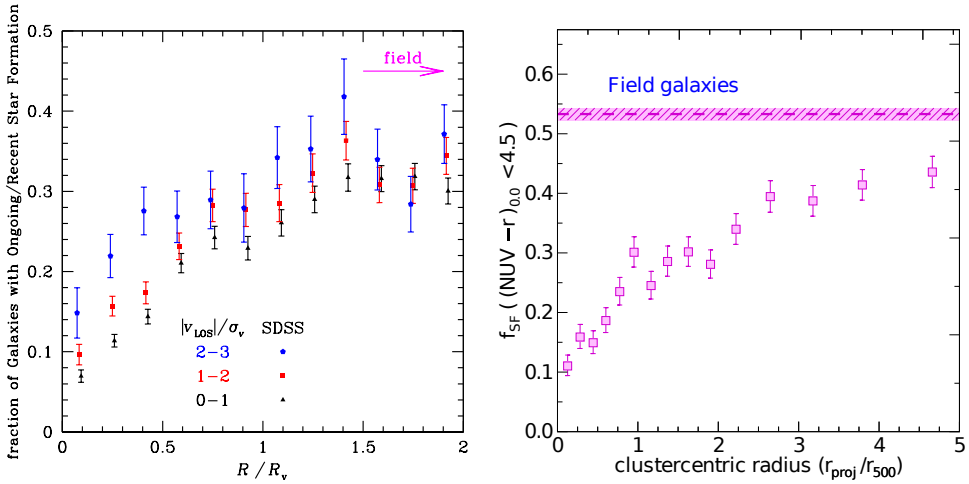

FIGURE 1.12: The fraction of cluster star-forming galaxies as a function of clustercentric distance. *Left panel* by Mahajan *et al.* (2011): fraction of cluster star-forming galaxies vs. projected radius (R) normalised by the virial radius (R_v) in bins of absolute line-of-sight velocity ($v_{\rm LOS}$) normalised by σ_v: 0–1 σ_v (black triangles), 1–2 σ_v (red squares) and $>$ 2 σ_v (blue pentagons), slightly shifted along the x-axis for clarity. The magenta arrow indicates the fraction in the field (R $>$ 3 R_v). *Right panel* adapted from Haines *et al.* (2015): fraction of star-forming cluster galaxies with blue rest-frame UV–optical colours, $(NUV-r)_{0.0} < 4.5$, indicating ongoing unobscured star-formation, vs the clustercentric radius (x-axis), where $r_{\rm proj}$ is the projected radius, and r_{500} is the radius enclosing a mean overdensity of 500 with respect to the critical density of the Universe at the cluster redshift. The magenta dashed line indicates the corresponding fraction of field galaxies with and its 1σ confidence levels (magenta shaded region).

Chapter 2

This work

2.1 Motivation

From what was presented in Chapter 1, it should be clear by now the complexity of determining what causes the quenching of star formation in galaxies. Most of the studies that attempt to investigate the connection between the quenching of star formation and the environment using galaxy properties tackle a few properties at a time. To work with several properties simultaneously, it is essential to employ a sophisticated statistical investigation. Research needs to rely on models that are able to accommodate several variables, and the statistical interpretation of these results needs to be carried out from different angles to be able to disentangle the role of each galaxy property. Besides that, not all observed properties of galaxies have been properly investigated in the literature regarding what concerns the matter. For instance, the velocity dispersion of galaxies cannot be neglected as an important quenching tracer.

Furthermore, to understand the effect that groups in the outskirts of larger halos cause in the fraction of star-forming galaxies, it is important to connect galaxy quenching in cluster environments to the overall environment structure. In this context, Trevisan *et al.* (2017b) presented an assignment scheme that can be used to examine the distribution of galaxies in clusters up to 20 virial radii. It is a natural step to employ such methodology in conjunction with multivariate analysis to answer the aforementioned open questions in the field. Furthermore, in the past decade, large-scale surveys, such as the Sloan Digital Sky Survey, presented an unprecedented amount of multivariate data for galaxies. Therefore, in this work, we will use both these tools with the addition of a robust statistical framework.

2.2 Objectives

The relation of quenching with galaxy properties and environment can unveil different physical processes and provide a more accurate picture of the physics of quenching to finally determine which process dominates in which circumstance. Therefore, this dissertation explores the relationship between different galaxy properties with the probability of star-forming at a given distance from the centre of the galaxy host halo. Specifically, in this study, we tackle the following questions:

- How does the fraction of star-forming galaxies depend on the galaxy stellar mass and velocity dispersion when we account for more galaxy properties?
- How far from the cluster centre do we still see an increase in the fraction of star-forming galaxies?
- What is the effect that small groups in the outskirts of larger halos cause in the fraction of star-forming galaxies?

2.3 Outline of the dissertation

This dissertation is organised as follows: In Chapter 3, we describe the sample and data used in this work. In Chapter 4, we explain the membership assignment scheme that we used to assign galaxies to haloes out to large clustercentric distances. In Chapter 5, we present the galaxies subsamples used in our analysis, defined to be complete in stellar mass within a given volume, and we also describe the methods used to classify the galaxies as star-forming or quiescent and AGN or non-AGN. In Chapter 6, we present the galaxy properties that we want to explore in relation to the fraction of star-forming satellites. We examine the statistical relationship between these properties and their connection with the fraction of star-forming galaxies through descriptive analysis and then present our comprehensive logistic regression to model the relationship between the fraction of star-forming galaxies and galaxy properties. In Chapter 7, we present the results of our logistic regression and compare the best-fit models with observations. In Chapter 8, we discuss the implications of our results with regard to the current literature on the topic. Finally, in Chapter 9, we outline our conclusions and give a brief summary of the perspectives of our study for future work.

Chapter 3

Data and sample definition

This chapter describes the catalogues and data from the literature and the galaxy sample used in this work. In Sections 3.1 and 3.2, we present the main galaxy sample and summarise the additional catalogues from which we extract galaxy properties needed for our analysis. Section 3.3 describes the catalogue of groups and clusters used in this work. Finally, Section 3.4 describes the catalogues matching and cleaning of the galaxy sample before running our halo membership assignment.

3.1 The initial galaxy sample

The Sloan Digital Sky Survey (SDSS) is a massive astronomical survey project that has been ongoing since the late 1990s. It is the first and one of the largest open-source catalogues of galaxies with complete photometric and spectroscopic information on millions of astronomical objects. The SDSS uses primarily a 2.5-meter telescope located at the Apache Point Observatory in New Mexico (Gunn *et al.*, 2006) and a set of specialised instruments, including cameras, spectrographs, and other imaging equipment.

The SDSS has conducted several different phases of observations, each with different scientific goals. Particularly, the SDSS-I/II (2000-2008) focused on the *Legacy Survey* (Abazajian *et al.*, 2003; Abazajian *et al.*, 2004), a program that targeted hundreds of thousands of galaxies and quasars covering more than 7500 square degrees in the sky. In Figure 3.1, we show the footprint of the photometric and spectroscopic observations of the *Legacy Survey*. These objects have images and photometry in the five filters (Figure 3.2), and spectroscopic data covering a wide wavelength range from ~3500 to 10 000 Å– i.e., from near UV to near-infrared – providing us with one of the most complete wavelength coverage for a survey of this scale to this date. The final complete spectroscopic observations of the *Legacy survey* were released in DR7 (Data Release 7, Abazajian *et al.*, 2009).

In this work, the observational sample of galaxies was selected from the *Main Galaxy Sample* (hereafter SDSS-MGS, Strauss *et al.*, 2002) of the Legacy Survey, and the data were retrieved from the SDSS-V DR18 database (Almeida *et al.*, 2023). The MGS is one of the largest and most complete spectroscopic and imaging catalogues of galaxies ever compiled, including galaxies of all types, from early-type ellipticals to late-type spirals and irregulars. The algorithm that selects the main sample of galaxies for spectroscopy from the photometric data obtained by the imaging survey is detailed in Strauss *et al.* (2002).

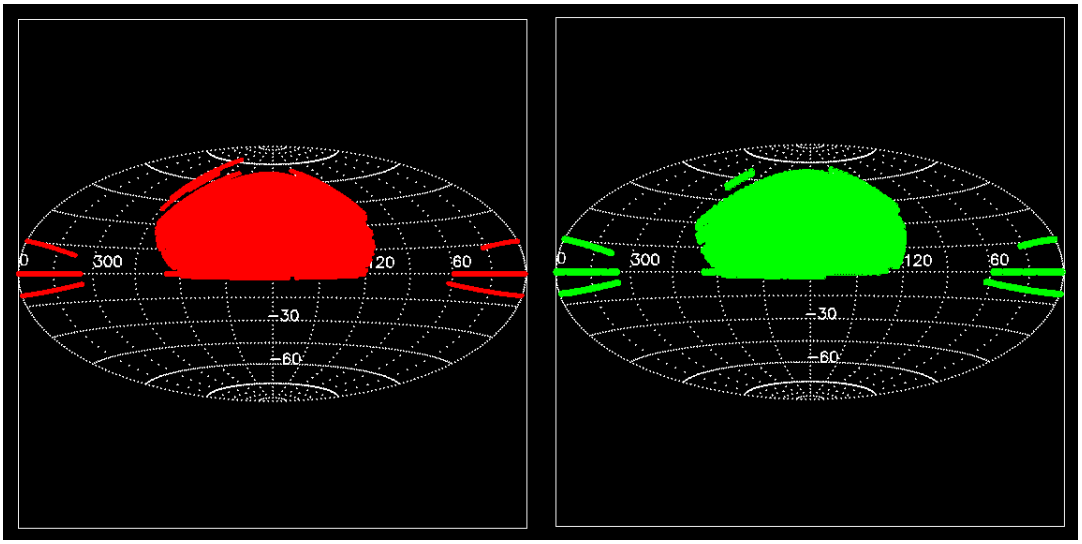

FIGURE 3.1: Regions of the sky covered by imaging (*left panel*) and spectroscopic (*right panel*) observations of the SDSS Legacy Survey.

Essentially, the SDSS-MGS consists of galaxies with r-band Petrosian magnitude $r \leq 17.77$ and r-band Petrosian half-light surface brightness $\mu_{50} \leq 24.5$ mag arcsec^{-2}, resulting in the completeness of the sample exceeding 99%. For almost all of the targeted galaxies (99.9%), the SDSS spectra have a signal-to-noise ratio S/N > 4 per pixel, which is enough to provide reliable redshift measurements with typical statistical errors of less than 30 km s^{-1}.

The SDSS-MGS data used in this work was retrieved through `CasJobs` [1] (*Catalog Archive Server Jobs System*), a web-based query and data management tool to perform SQL queries and analyse large datasets. Galaxies between $0.01 \geq z \geq 0.1$ were selected, where the lower limit is chosen to avoid uncertainties in the definition of the environment due to the galaxy's peculiar velocities. In turn, the upper limit is chosen to have a sufficient sample size since the number of galaxies decreases with increasing redshift within the SDSS survey. The query for acquiring the data is presented in Appendix A and leads to a sample of 325 330 galaxies.

3.2 Galaxy properties

In this section, we describe the galaxy properties used in our analysis. The corrections applied to the galaxy magnitudes are described in Section 3.2.1. The main property that can be used to classify a galaxy as star-forming or quiescent is its sSFR, which can be estimated in different ways with different indicators. To explore how our results depend on the way that the sSFR is determined, we used sSFR derived through different approaches: one inferred mainly from the luminosity of the Hα emission line (estimates by the MPA-JHU group, Brinchmann *et al.*, 2004; Section 3.2.2) and the other based on SDSS photometry combined with UV data from *GALEX–SDSS–WISE* Legacy Catalog (Salim *et al.*, 2016; Section 3.2.3). We also extracted $M_\star$ estimates from these catalogues.

[1] See `https://skyserver.sdss.org/casjobs/`

We also retrieve the galaxy effective radii, r_e, from the catalogue by Simard *et al.* (2011, Section 3.2.4). The r_e was used to standardise the stellar velocity dispersions to a fixed aperture (Section 3.2.5). Finally, to classify the galaxies as AGN or non-AGN, we retrieved the fluxes and equivalent widths of emission lines from the SDSS database (Section 3.2.6).

3.2.1 Absolute magnitudes

In the context of SDSS data, the Petrosian magnitude is often used for galaxies, providing a standardised measure of their brightness. The SDSS has adopted a modified form of the Petrosian, 1976 system, measuring galaxy fluxes within a circular aperture that depends on the surface-brightness profile of the object. We corrected the SDSS apparent Petrosian magnitudes in the r band, m_r^{Petro}, for the Galactic extinction (see the query in Appendix A) and applied the k-corrections, which converts the apparent magnitudes into their respective rest-frames, computed with the `KCORRECT` code (version 4_3) of Blanton and Roweis (2007). Then we calculate the absolute magnitudes, M_r^{Petro}, using the relation between apparent and absolute magnitudes

$$M_r^{\mathrm{Petro}} = m_r^{\mathrm{Petro}} - [5 \log_{10}(\mathrm{D_L}) - 5]\,. \tag{3.1}$$

where $\mathrm{D_L}$ is the luminosity distance in parsec, which depends on the redshift of the galaxy.

3.2.2 Stellar masses and SFRs from MPA-JHU

The MPA-JHU (*Max Planck Institute for Astrophysics - Johns Hopkins University*) catalogue (Kauffmann *et al.*, 2003; Salim *et al.*, 2007; Brinchmann *et al.*, 2004; Tremonti *et al.*, 2004) is a publicly-available database of several quantities inferred from SDSS spectroscopic data. A collaboration between the Max Planck Institute for Astrophysics in Germany and the Johns Hopkins University in the United States created it. The data in the MPA-JHU catalogue were obtained from the SDSS database, and it contains physical properties such as galaxy stellar masses and star formation rates.

The MPA-JHU catalogue's stellar masses were calculated using a Bayesian approach and model grids described in Kauffmann *et al.* (2003). Since the SDSS spectra were obtained through a fibre with an aperture of 3″, they do not probe the entire galaxy. Therefore, the stellar masses are inferred from the *ugriz* galaxy photometry alone. They also corrected the photometry for the small contribution from nebular emission, which they estimated using the spectra. They compute the stellar masses within the SDSS fibre aperture using fibre magnitudes and the total stellar mass using model magnitudes. A Kroupa (2001) initial mass function was assumed. They provide the stellar mass corresponding to the median and 2.5%, 16%, 84%, and 97.5% of the probability distribution function. In this work, we used the median of the total stellar mass (`lgm_tot_p50`) as the measure of stellar masses.

The star formation rates (SFRs) were computed within the galaxy fibre aperture using the nebular emission lines as described in Brinchmann *et al.* (2004). SFRs outside of the fibre were estimated using galaxy photometry following Salim *et al.* (2007). For AGN and galaxies with weak emission lines, SFRs were estimated from the photometry. They report both the fibre SFR and the total SFR at the median and 2.5%, 16%, 84%, and 97.5% of the probability distribution function. In this work, we used the median of total SFR (`sfr_tot_p50`) as a measure of SFR.

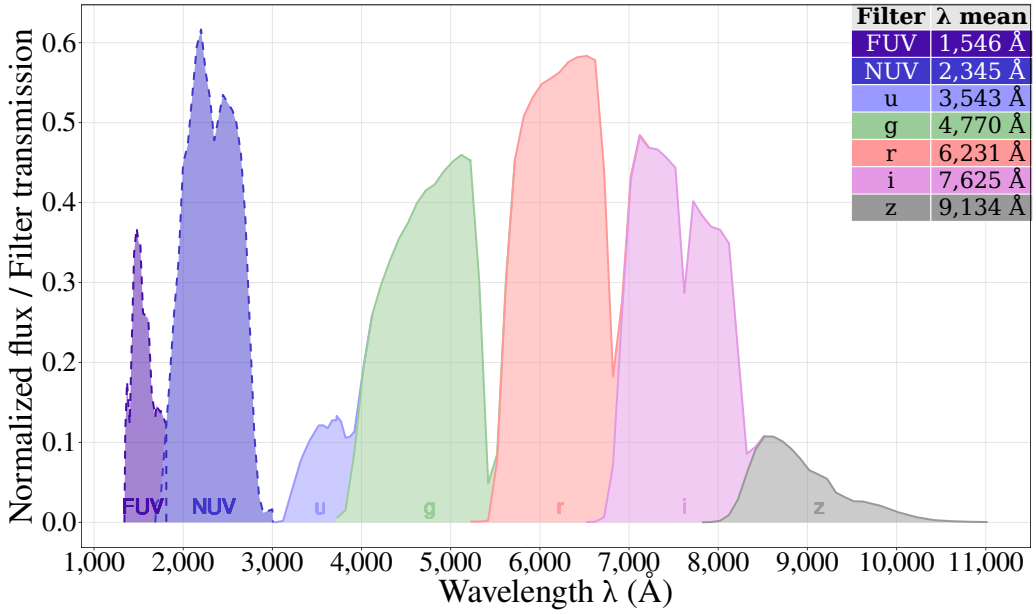

FIGURE 3.2: SDSS *ugriz* (*solid lines*) and GALEX *FUV* and *NUV* (*dashed lines*) filters.

3.2.3 Stellar masses and SFRs from GSWLC

The GSWLC (*GALEX–SDSS–WISE Legacy Catalog*, Salim *et al.*, 2016) is a catalogue of physical properties of galaxies within the GALEX (*Galaxy Evolution Explorer*, Martin *et al.*, 2005) footprint. The GALEX (2003-2013) was a space-based ultraviolet telescope designed to observe ultraviolet light, which can provide important information about the young, hot stars that are responsible for much of the light emitted by galaxies.

The GSWLC catalogue contains physical properties for $\sim$ 700,000 SDSS galaxies between $0.01 \leq z \leq 0.3$ within the GALEX footprint, regardless of UV detection, corresponding to $\sim$ 90% of SDSS galaxy sample. In Figure 3.2, we can see the wavelength range covered by GALEX filters NUV and FUV, their respective transmissions, and the SDSS filters.

The physical properties of GSWLC were obtained from UV/optical spectral energy distribution (SED) fitting following the Bayesian methodology of Salim *et al.* (2007), with improvements such as blending corrections for low-resolution UV photometry, flexible dust attenuation laws, and emission-line corrections. An improved version of GSWLC was made by Salim *et al.* (2018) leading to the GSWLC-2. In this work, we are using the GSWLC-2.

There are three versions of the GSWLC (and, consequently, for the GSWLC-2) catalogue depending on the depth of the UV photometry: GSWLC-A (shallow catalogue), containing 88% of the galaxies in SDSS; GSWLC-M (medium-deep catalogue), with 49% of the SDSS galaxies; and GSWLC-D (deep catalogue), with 7% of SDSS galaxies. There is also a *master catalogue*, the GSWLC-X, that combines the three catalogues, keeping the data from the deepest catalogue for each galaxy, comprising 90% of the SDSS galaxies. The authors of Salim *et al.* (2016) argue that the non-uniform depth in the GSWLC-X may lead to systematic errors in SFRs, especially for galaxies with

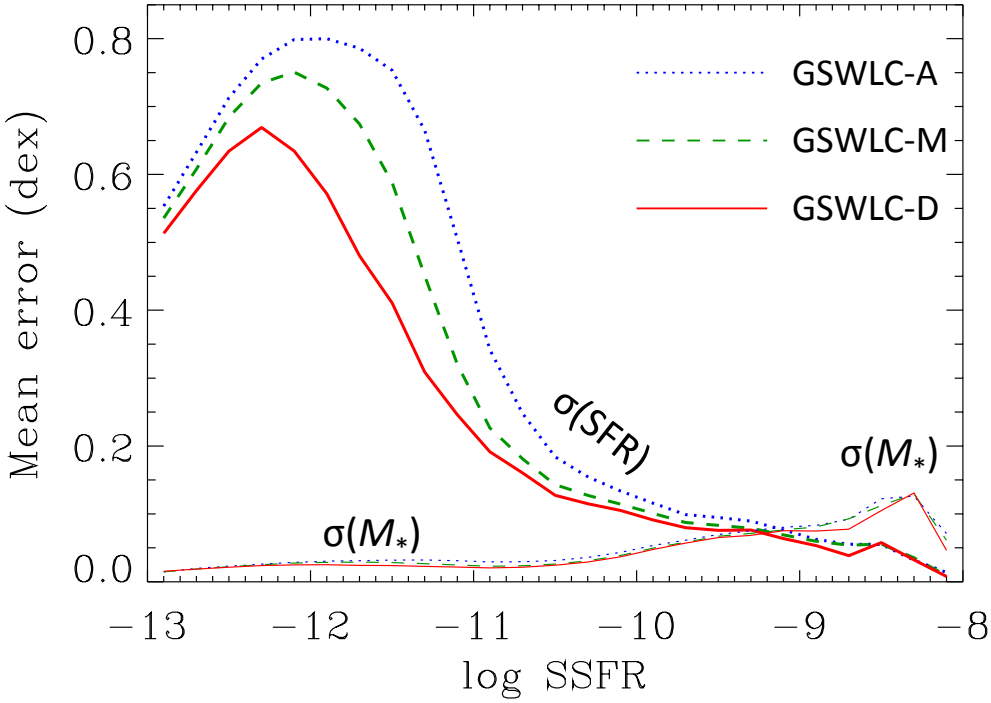

FIGURE 3.3: Mean random error of SED fitting SFR (thick lines) and stellar mass (thin lines) for each GSWLC catalogue from Salim *et al.*, 2016.

$\log_{10}(\mathrm{sSFR/yr^{-1}}) < -11$, as shown in Figure 3.3. For low sSFRs, the UV detection rate and their quality differ from catalogue to catalogue, but the stellar masses are less affected by these biases.

Due to the broad coverage of SDSS and large sample size, which are desirable attributes given the statistical nature of our work, we chose the second version of the master catalogue (GSWLC-X2, hereafter GSWLC) containing 659,229 galaxies. The issue related to the possible systematic errors in $\log_{10}(\mathrm{sSFR/yr^{-1}}) < -11$ does not have much impact on our results since we do not analyse the SFRs *per se* but only classify the galaxies between star-forming or quiescent according to their position relative to the Main Sequence (see Chapter 5). We selected only galaxies of GSWLC that are in the SDSS-MGS (`flag_mgs = 1`) and are at $z \leq 0.1$, which results in a subsample of 301,021 galaxies.

3.2.4 Galaxy effective radii

The galaxy's effective radius (or half-light radius), r_e, is a measure of its size and corresponds to the radius that contains half of the total galaxy luminosity. The r_e values were used to standardise the stellar velocity dispersions to a fixed aperture. We obtained r_e from the catalogue by Simard *et al.* (2011). They fitted Sérsic+Exponential models to the 2D surface brightness profiles of a sample of 1 123 718 galaxies from the Legacy Survey of SDSS-DR7. We adopted r_e = `Rhlr`, where `Rhlr` is the galaxy semi-major axis in the r-band. We converted the galaxy r_e (in units of kpc) to θ_e (in units of arcseconds) using the parameter `Scale` (in arcsec $\mathrm{kpc^{-1}}$) from the S11 catalogue.

3.2.5 Velocity dispersions

To standardise the stellar velocity dispersions, we converted the σ_{ap} retrieved from the SDSS database, which was measured through a fixed aperture with radius $r_{ap} = 1.5$ arcseconds, into the velocity dispersion σ_e that would be measured within r_e (see Subsection 3.2.4). We adopted as r_e the semi-major axis of the half-light ellipse from Sérsic+Exponential fits and then calculate σ_e using the relation of Cappellari *et al.* (2006):

$$\sigma_e = \sigma_{ap}/(r_{ap}/r_e)^{-0.066} . \tag{3.2}$$

3.2.6 Emission line fluxes and equivalent widths

We retrieved fluxes of the emission lines [NII]λ6584, Hα, and Hβ and the equivalent width of Hα, EW(Hα), from the *emissionLinesPort* table available in the SDSS database. The information was obtained with the query given in Appendix A. For these measurements, the galaxy spectra were fitted using an adaptation of the publicly-available codes *Gas AND Absorption Line Fitting* (`GANDALF`, Sarzi *et al.*, 2006) and *penalised PiXel Fitting* (`pPXF`, Cappellari and Emsellem, 2004). The stellar population models used to model the stellar continuum are from Maraston and Strömbäck (2011) and Thomas *et al.* (2011).

3.3 The sample of groups and clusters

The galaxy groups and clusters were selected from the catalogue compiled by Lim *et al.* (2017), who applied a halo-based group finder to redshift surveys, including the SDSS. Their group-finder algorithm is based on that of Yang *et al.* (2007) but with an improved halo mass assignment. The halo mass estimates are based on abundance matching and are assigned to groups according to observed proxies, which can be the total stellar mass or total luminosity of member galaxies. We used the catalogue in which the luminosity was used as a proxy for the halo mass estimates, the SDSS(L) catalogue (hereafter L17). The L17 catalogue comprises 446,495 groups drawn from a sample of 586,025 galaxies from the SDSS-IV DR13 database. We highlight that we retrieved the information about the groups from the catalogue, but we do not use the original membership assignment of galaxies made in L17. Instead, we applied the halo membership assignment using an approach described in Chapter 4.

3.3.1 Conversion from L17 $r_{180,m}$ to $r_{100,c}$

The group halo masses provided by the L17 catalogue (M_{L17}) are defined within a radius (r_{L17}) where the mean density of the halo is 180 times the *mean* density of the Universe at the group redshift. In the context of galaxy groups or clusters, the virial radius (r_{vir}) encompasses the region where the gravitational potential is deep enough to bind the system against the overall cosmic expansion. Since we want to study the quenching of star formation in galaxies in relation to distances to the centre of the group in r_{vir} units, it is more appropriate to convert r_{L17} to r_{vir} and M_{L17} to the virial mass M_{vir}.

Relating $\bar{\rho}_{\rm L17}$ to $\bar{\rho}_{\rm vir}$

The first step is to relate the mean density of the halo defined by L17, the $\bar{\rho}_{\rm L17}$, into its equivalent $\bar{\rho}_{\rm vir}$, but defined in terms of $r_{\rm vir}$. As described above, the mean density of the halo in the L17 catalogue is defined as having 180 times the *mean density of the Universe*, as follows,

$$\bar{\rho}_{\rm L17} = \rho(r_{\rm L17}) = 180\,\bar{\rho}_{\rm U}(z)\,, \tag{3.3}$$

where $\bar{\rho}_{\rm U}(z)$ is the mean density of the Universe at the group redshift. $\bar{\rho}_{\rm U}(z)$ is defined as the total density parameter of the Universe $\boldsymbol{\Omega}$ times the critical density of the Universe ρ_c so

$$\bar{\rho}_{\rm U}(z) = \boldsymbol{\Omega}\rho_c\,. \tag{3.4}$$

The critical density of the Universe $\rho_{\rm c}$ represents a balance between the gravitational attraction (which tends to make the Universe collapse) and the expansion driven by dark energy, being related to the density needed for the system being gravitationally bound and in virial equilibrium. It is mathematically defined as

$$\rho_c(z) = E^2(z)\rho_{\rm c,0}\,, \tag{3.5}$$

where $\rho_{\rm c,0}$ is the critical density at $z = 0$ and $E^2(z)$ is a function that depends on the content and shape of the Universe, defined as follows

$$E^2(z) \equiv \boldsymbol{\Omega}_r(1+z)^4 + \boldsymbol{\Omega}_m(1+z)^3 + \boldsymbol{\Omega}_k(1+z)^2 + \boldsymbol{\Omega}_\Lambda\,, \tag{3.6}$$

where $\boldsymbol{\Omega}_r$, $\boldsymbol{\Omega}_m$, $\boldsymbol{\Omega}_k$ and $\boldsymbol{\Omega}_\Lambda$ are the radiation, matter, curvature and dark energy density parameters, respectively. In a Universe with Euclidean geometry (flat, $\boldsymbol{\Omega}_k = 0$), estimated by the Wilkinson Microwave Anisotropy Probe (WMAP) measurements (Hinshaw *et al.*, 2013; Bennett *et al.*, 2013), the Equation 3.6 above can be reduced to

$$E^2(z) = \boldsymbol{\Omega}_{m,0}(1+z)^3 + 1 - \boldsymbol{\Omega}_{m,0}\,. \tag{3.7}$$

Still from Equation 3.4, the total density parameter $\boldsymbol{\Omega}$ is a dimensionless quantity that describes the ratio of the actual density of the Universe to the critical density. For a flat Universe $\boldsymbol{\Omega} = 1$ and, therefore, $\bar{\rho}_{\rm U}(z) = \rho_c$. Hence the Equation 3.3 can be rewrite using $\boldsymbol{\Omega} = 1$ and Equation 3.7 as

$$\bar{\rho}_{\rm L17} = 180\,\boldsymbol{\Omega}_{m,0}(1+z)^3 + 1 - \boldsymbol{\Omega}_{m,0}\,. \tag{3.8}$$

Now, let's define the mean halo density equivalent to Equation 3.8, but for the $r_{\rm vir}$. Formally, $r_{\rm vir}$ is defined as the radius at which the mean density of the halo is equal to the *critical density of the Universe* $\rho_{\rm c}$ at the redshift of the system, multiplied by an overdensity constant $\Delta_{\rm c}(z)$, that is,

$$\bar{\rho}_{\rm vir} = \rho(r_{\rm vir}) = \Delta_{\rm c}(z)\,\rho_{\rm c}(z)\,. \tag{3.9}$$

From Equation 3.5, we can rewrite 3.9 as

$$\bar{\rho}_{\rm vir} = \Delta_{\rm c}(z)\, E^2(z)\, \rho_{\rm c,0}\,. \tag{3.10}$$

The value of $\Delta_{\rm c}(z)$ in Equation 3.10 is obtained from the solution to the collapse of a spherical top-hat perturbation, presuming that the cluster has recently achieved virialisation (Peebles, 1980), given by Bryan and Norman (1998), i.e.,

$$\Delta_{\rm c}(z) = 18\pi^2 + 82x - 39x^2\,, \tag{3.11}$$

where $x \equiv \Omega_m(z) - 1$, and $\Omega_m(z)$ is the mass density in the Universe, with value depending on the cosmological model adopted. For the ΛCDM cosmology, $\Delta_{\rm c}(z)$ varies with redshift, decreasing from $\Delta_{\rm c}(z) \approx 180$ at high redshifts (where $\Omega_m(z = 10^3) \approx 1$, e.g. Voit, 2005) to $\Delta_{\rm c}9z) \approx 100$ in the present Universe (where $\Omega_m(z = 0) \approx 0.3$, e.g. Hinshaw *et al.*, 2013). Therefore, in this work, we adopt $\Delta_{\rm c}(z) = 100$, so that our virial radius and virial mass are defined as $r_{\rm vir} = r_{100}$ and $M_{\rm vir} = M_{100}$, respectively.

From Equations 3.8 and 3.10, the ratio of mean densities is

$$\frac{\bar{\rho}_{\Delta_{\rm c}}}{\bar{\rho}_{\rm L17}} = \frac{\Delta_{\rm c}}{180\,\Omega_{\rm m,0}}\,\frac{E^2(z)}{(1+z)^3} \simeq 1.97\,\frac{E^2(z)}{(1+z)^3} \tag{3.12}$$

for $\Delta_{\rm c} = 100$ and for $\Omega_{\rm m,0} = 0.282$ adopted in L17.

Relating the concentration parameter $c_{\rm L17}$ to $c_{\rm vir}$

Now we have to relate the concentration parameter from L17 $c_{\rm L17}$ with that calculated for $r_{\rm vir}$, the $c_{\rm vir}$. The concentration parameter, denoted as c, is a measure of how centrally concentrated the mass distribution is within a dark matter halo. Assuming an NFW model (Navarro *et al.*, 1997) for the mass distribution in the groups, the mean density profile can be expressed as

$$\bar{\rho}(r) = \bar{\rho}(a)\widetilde{\rho}\left(\frac{r}{a}\right) = \bar{\rho}(a)\,\frac{\widetilde{M}(r/a)}{(r/a)^3}\,, \tag{3.13}$$

where r is the halo radius and a is the scale radius. Also, from Equation 3.13,

$$\begin{aligned} \widetilde{\rho}(x) &= \frac{\widetilde{M}(x)}{x^3} \\ \widetilde{M}(x) &= \frac{M(a\,x)}{M(a)} = \frac{\ln(x+1) - x/(x+1)}{\ln 2 - 1/2} \end{aligned}\,, \tag{3.14}$$

with $\widetilde{\rho}(1) = \widetilde{M}(1) = 1$.

Since the scale radius a is fixed, Equations 3.12 and 3.13 lead to

$$\frac{\widetilde{\rho}(c_{\rm vir})}{\widetilde{\rho}(c_{\rm L17})} = \frac{\Delta_{\rm c}(z)}{180\,\Omega_{\rm m,0}}\,\frac{E^2(z)}{(1+z)^3}\,, \tag{3.15}$$

where $c_{\rm vir} = r_{\rm vir}/a$ and $c_{\rm L17} = r_{\rm L17}/a$.

We can use the concentration-mass relation for ΛCDM halos (that of Macciò *et al.*, 2008), $c_{\rm vir} = c_{\Lambda{\rm CDM}}(M_{\rm vir})$, and we write the L17 concentration parameter as

$$
\begin{aligned}
c_{\rm L17} = c_{\rm vir}\,\frac{r_{\rm L17}}{r_{\rm vir}} = c_{\rm vir}\left(\frac{M_{\rm L17}}{M_{\rm vir}}\right)^{1/3}\left(\frac{\bar{\rho}_{\rm vir}}{\bar{\rho}_{\rm L17}}\right)^{1/3} = \\
= \left(\frac{\Delta_{\rm c}(z)}{180\,\Omega_{\rm m,0}}\right)^{1/3}\frac{E^2(z)}{1+z}\,c_{\Lambda{\rm CDM}}(M_{\rm vir})\left(\frac{M_{\rm L17}}{M_{\rm vir}}\right)^{1/3}.
\end{aligned}
\tag{3.16}
$$

Finding $M_{\rm vir}$

Given the virial radius and the overdensity convention, the virial mass $M_{\rm vir}$ can be found through the relation

$$
M_{\rm vir} = \frac{4}{3}\pi r_{\rm vir}^3 \Delta_{\rm c}(z)\rho_c\,. \tag{3.17}
$$

Combining the Equations 3.12 and 3.16 leads to finding $M_{\rm vir}$ by solving

$$
\begin{aligned}
&\frac{\widetilde{\rho}\left[\left(\Delta_{\rm c}(z)/(180\,\Omega_{m,0})\right)^{1/3}\,\left(M_{\rm L17}/M_{\rm vir}\right)^{1/3}\,c_{\Lambda{\rm CDM}}(M_{\rm vir})E^{2/3}(z)/(1+z)\right]}{\widetilde{\rho}\left[c_{\Lambda{\rm CDM}}(M_{\rm vir})\right]} = \\
&= \frac{180\,\Omega_{\rm m,0}}{\Delta_{\rm c}(z)}\,\frac{(1+z)^3}{E^2(z)}\,,
\end{aligned}
\tag{3.18}
$$

using Equations 3.14, where the $(\ln 2 - 1/2)$ term in the latter scales out.

Finding $r_{\rm vir}$

The virial radius $r_{\rm vir}$ is then obtained by inverting Equation (3.17) to give

$$
r_{\rm vir} = \left(\frac{3}{4\pi}\frac{M_{\rm vir}}{\Delta_{\rm c}(z)\rho_c}\right)^{1/3} \approx 0.13\left(\frac{M_{\rm vir}}{\rho_c}\right)^{1/3}\,{\rm kpc}\,, \tag{3.19}
$$

for $\Delta_{\rm c}(z) = 100$ as adopted by Lim *et al.*, 2017.

3.4 Final galaxy sample and data

To obtain the final sample of galaxies that we assign to the L17 groups (see Chapter 4), we matched the galaxy samples from different catalogues as follows. First, we match the SDSS-MGS, GSWLC and S11 catalogues using the triple match function of `Topcat` software[2] by the `Sky` algorithm with a maximum error of 5.0 arcsec, keeping only galaxies that are in all three catalogues. This step results in a sample of 288,408 galaxies. Then, we remove all galaxies with null values of $\log_{10}$ SFR and $\log_{10} M_\star$ in both MPA-JHU and GSWLC-X2 catalogues, null values of [NII]λ6584 and Hα fluxes, EW(Hα), $M_{\rm r}^{\rm Petro}$ and $\log\sigma_{\rm e}$. After matching and removing objects with null galaxy property values from the initial sample ($\approx$ 1.6% of the galaxies), we get our final sample with 283 846 galaxies.

[2] Topcat Version 4.8-7, see `https://www.star.bris.ac.uk/~mbt/topcat`

Chapter 4

Assigning galaxies to haloes

In this chapter, we explain the methodology used to assign galaxies to the halos of groups and clusters, presenting the membership assignment scheme proposed by Trevisan *et al.* (2017b) that will allow us to explore the structure of the halo's outer regions, which is important to understand the effect that groups in the outskirts of larger halos cause on the $f_{\rm SF}$. In this work, we adopt the ΛCDM cosmology with $\Omega_{\rm m} = 0.315$, $\Omega_\Lambda = 0.685$ and $H_0 = 67.4 {\rm kms}^{-1}{\rm Mpc}^{-1}$ (Planck Collaboration *et al.*, 2020).

4.1 Membership assignment scheme

Ideally, to assign a given galaxy to a halo, one could measure the three-dimensional halo-centric distances between that galaxy and all the groups in the sample and assign the galaxy to the closest group in units of the group virial radius. However, observational data is acquired in redshift space, and their physical real-space halo-centric distances are difficult to evaluate due to redshift distortions. If a galaxy is very far from any group halo and is not detached from the Hubble flow, its distance to the group halo can be measured using the redshift-space distances. However, if a galaxy is close to the centre of a halo, the galaxy peculiar velocities make it difficult to accurately measure this distance.

Taking this into account, Trevisan *et al.* (2017b) proposed a scheme to assign galaxies to their nearest group in units of the group $r_{\rm vir}$, which combines the non-linear virialized regions of the clusters with their outer linear parts. They define a threshold projected distance, R_n, so that if the galaxy lies at projected distances $R > R_n$ from the group centre, its halo-centric distance is simply given by the distance in redshift space ($d_{\rm outer}$, Subsection 4.1.1). If $R \leq R_n$, they define a proxy to the halo-centric distance, which corresponds to the density contrast in the projected phase space ($d_{\rm inner}$, Subsection 4.1.2). Using simulations, Trevisan *et al.* (2017b) demonstrated that the distances obtained in redshift space with this method are in very good agreement with the three-dimensional distances.

Here we will briefly describe the assignment approach made by Trevisan *et al.*; detailed information on the method can be found directly in their paper. In this work, we adopt $R_n = 2.5\, r_{\rm vir}$ based tests made by Trevisan *et al.* (2017b) using simulations. We adopt the notations of r for the real-space (3D) halo-centric distance, while capital R refers to the projected distances to the halo centres.

4.1.1 Galaxies at $R > 2.5r_{\rm vir}$

If the galaxy is at $R > 2.5r_{\rm vir}$ from a halo, its halo-centric distance is given by the redshift-space distances

$$d_{\rm outer}(R, \Delta z) = \sqrt{\frac{\Delta_{\rm c}}{2}\left[\frac{c\Delta z}{v_{\rm vir}(1+z_{\rm group})}\right]^2 + \left(\frac{R}{r_{\rm vir}}\right)^2} \tag{4.1}$$

where R is the projected distance to the group centre, $\Delta z = z_{\rm group} - z_{\rm galaxy}$, $\Delta_{\rm c}$ is the critical overdensity for virialisation (see Equation 3.11), $r_{\rm vir}$ is the virial radius, $v_{\rm vir}$ is virial velocity given by $v_{\rm vir} = \sqrt{(\Delta_{\rm c}/2)}H(z)r_{\rm vir}$, and c is the the speed of light. As mentioned in Section 3.3.1, we adopt $\Delta_{\rm c} = 100$ to define the virial quantities.

4.1.2 Galaxies at $R \leq 2.5r_{\rm vir}$

For galaxies lying in the regions at $R \leq 2.5r_{\rm vir}$, their halo-centric distance is related to the phase-space density contrast, similar to the approach described in Yang *et al.* (2007) and Duarte and Mamon (2016). Assuming that the distribution of galaxies in phase space follows that of the dark matter particles, the density contrast of galaxies in redshift space around the group centre can be defined as

$$P_M(R, \Delta z) = \frac{H(z)}{c}\frac{\Sigma(R)}{\bar{\rho}(z)}p(\Delta z)\,, \tag{4.2}$$

where $\bar{\rho}(z)$ is the average density of the Universe given by

$$\bar{\rho}(z) = \frac{3\Omega_{\rm m}H(z)^2}{8\pi G}\,. \tag{4.3}$$

The projected surface mass density, $\Sigma(R)$, of a halo with a Navarro *et al.* (1996, hereafter NFW) mass profile is given by Lokas and Mamon (2001)

$$\Sigma(R) = \frac{c_{\rm v}^2 g(c_{\rm v})}{2\pi}\frac{M_{\rm vir}}{r_{\rm vir}^2}f(\widetilde{R})\,, \tag{4.4}$$

where $M_{\rm vir}$ is the virial mass, $\tilde{R} = R/r_{\rm vir}$, and

$$g(c_{\rm v}) = \frac{1}{\ln(1+c_{\rm v}) - c_{\rm v}/(1+c_{\rm v})}\,, \tag{4.5}$$

$$f(\widetilde{R}) = \frac{1 - |c_{\rm v}^2\widetilde{R}^2 - 1|^{-1/2}\, C^{-1}[1/(c_{\rm v}\widetilde{R})]}{c_{\rm v}^2\widetilde{R}^2 - 1}\,, \tag{4.6}$$

with

$$C^{-1}(x) = \begin{cases} \mathrm{acos}(x) & \text{if } R > r_s \\ \mathrm{acosh}(x) & \text{if } R < r_s\,. \end{cases} \tag{4.7}$$

In the equations above, r_s is the scale radius and $c_{\rm v}$ is the concentration parameter defined as $c_{\rm v} = r_{\rm vir}/r_s$.

The function $p(\Delta z)$ in Equation (4.2) corresponds to the distribution of galaxy redshifts along the line-of-sight relative to the group centre. It is assumed to be Gaussian and constant with R so that

$$p(\Delta z) = \frac{1}{\sqrt{2\pi}} \frac{c}{\sigma(1+z_{\rm group})} \exp\left[-\frac{(c\Delta z)^2}{2\sigma^2(1+z_{\rm group})^2} \right], \tag{4.8}$$

where σ is the rest-frame velocity dispersion of the group, given by $\sigma = \eta v_{\rm vir}$, with $\eta = 0.65$ (Mamon *et al.*, 2010).

To combine the inner and outer distances estimates, we normalise $d_{\rm inner}$ so that $d_{\rm inner} = d_{\rm outer}$ at $R = R_n = 2.5\, r_{\rm vir}$. Therefore, the P_M values given by Equation (4.2) were converted according to

$$P_M \to P'_M = d_{\rm inner}(R, \Delta z) = \left(\frac{\ln P_M - a}{b} \right)^{1/2}, \tag{4.9}$$

where a and b are given by

$$a = \frac{\widetilde{R}_n^2}{\Delta_{\rm c}\eta^2} + \ln\left(\frac{2}{3} \frac{H(z)}{H_0} \frac{c_{\rm v}^2 g(c_{\rm v})}{\Omega_{\rm m}\eta(1+z)} \sqrt{\frac{\Delta_{\rm c}}{\pi}} f(\widetilde{R}_n) \right), \quad b = -\frac{1}{\Delta_{\rm c}\eta^2}. \tag{4.10}$$

In summary, the halo-centric distance is given by Equation (4.1) for galaxies lying at $R > 2.5 r_{\rm vir}$ and by Equation (4.9) for $R \leq 2.5 r_{\rm vir}$, i.e.

$$d(R) = \begin{cases} d_{\rm outer}\ (\text{Equation 4.1}) & \text{if } R > 2.5 r_{\rm vir}\,, \\ d_{\rm inner}\ (\text{Equation 4.9}) & \text{if } R \leq 2.5 r_{\rm vir}\,. \end{cases} \tag{4.11}$$

and the galaxy is assigned to the halo that leads to the smallest value of $d(R)$.

To test our assignment scheme, we built a mock galaxy sample in projected phase space out to 20 virial radii by assigning each galaxy in real space to the nearest halo using the semi-analytical models of galaxy formation (SAM) by Henriques *et al.* (2015). In Figure 4.1, we show the physical and projected halo-centric distances and compare them with the distance estimates given by Equation (4.11). Comparing the 3D distances with the curves shown in Figure 4.1, it can be seen that the distances given by the assignment scheme described in this section are in very good agreement with the distances in real space.

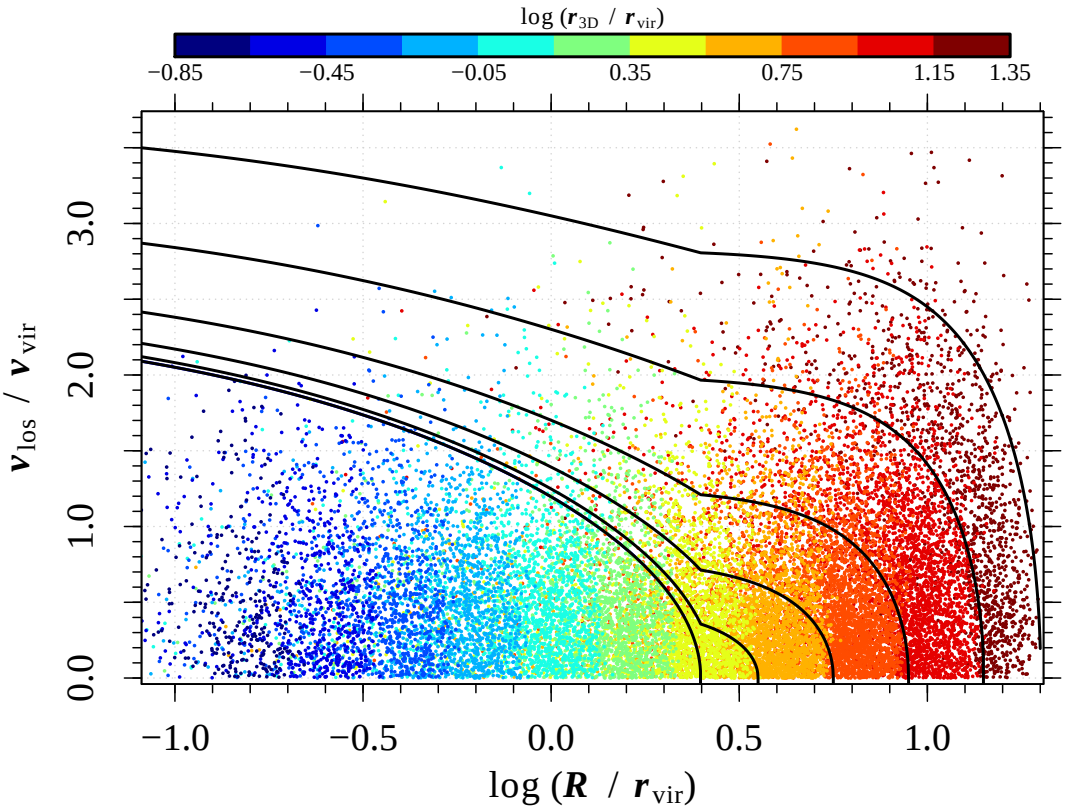

FIGURE 4.1: Real-space distance in terms of log projected distance and absolute line-of-sight velocity, obtained from the mock built from Henriques *et al.* (2015) semi-analytical model. The galaxy is assigned in real space to the nearest halo with mass above $\log_{10}(M_a/M_\odot) = 12.3$, and the colours indicate the distance values. The curves show the distances in the projected space obtained with the assignment scheme adopted in this work. All quantities are in virial units.

4.2 Group mass thresholds for the assignment

To explore the 1-halo term by measuring the density profiles of groups traced by their galaxies out to 20 r_{vir}, Trevisan *et al.* (2017b) considered two group mass thresholds. The first one, M_s, corresponds to the minimum halo mass of groups in the sample. The second, M_a, is the lower mass limit adopted in the assignment procedure, i.e., galaxies can be assigned only to haloes with $M_{vir} \geq M_a$, where $M_a \leq M_s$. These two mass thresholds are illustrated in figure 4.2. In the scenario **(A)**, no galaxy can be assigned to *halo 2* since its mass is below the threshold ${}^{A}M_a$. On the other hand, in scenario **(B)**, a lower mass limit (${}^{B}M_a < M_{vir,2}$) is adopted, so *halo 2* is now included in the assignment procedure. Therefore, the galaxies indicated in the figure are now assigned to *halo 2*, no longer contributing to the density profile of *halo 1*. In practice, this approach is designed to remove the two-halo term contribution to the *halo 1* density profile. Figure 4.3 exemplifies the change of M_a for a given group. In the left panel, $\log_{10}(M_a/M_\odot) = 14.0$, so all galaxies are assigned to the halo in the centre. In the middle and right panels,

$\log_{10}(M_a/M_\odot)$ = 13.0 and 12.3, respectively, so that many galaxies are now assigned to other haloes and will not contribute to the f_{SF} profile of the halo in the centre.

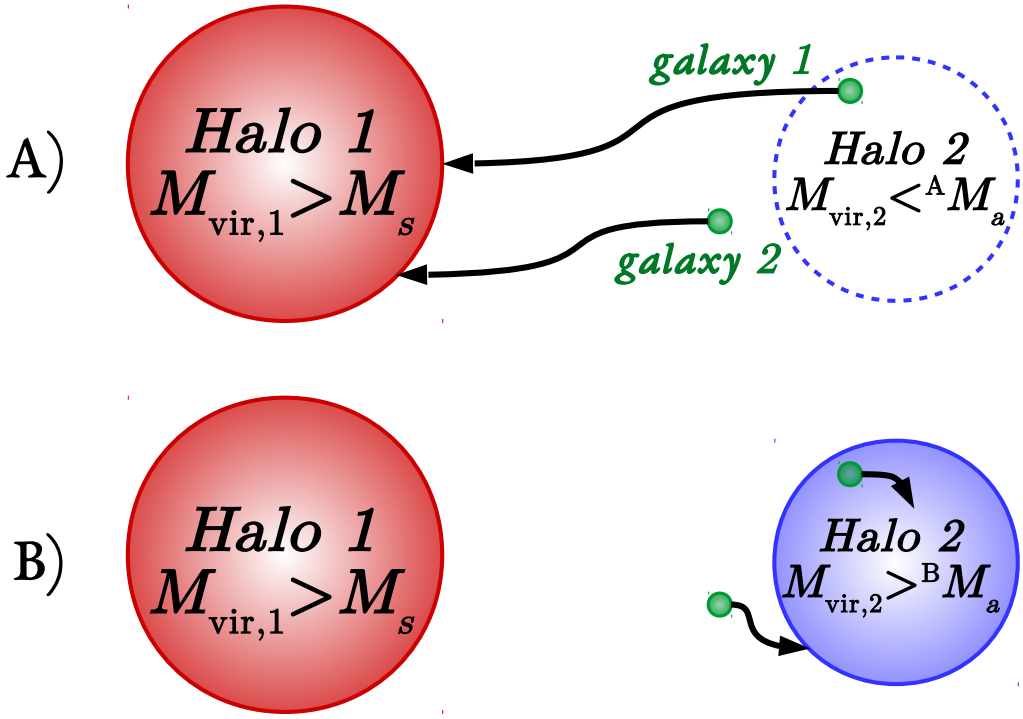

FIGURE 4.2: Diagram illustrating the two group-mass thresholds in the assignment scheme. Two scenarios are shown in the figure: in scenario **(A)**, the mass of *halo 2* is below the threshold ${}^{A}M_a$. Hence, no galaxy can be assigned to it, and even galaxies lying within the virial radius of *halo 2* (as galaxy 1) are assigned to *halo 1*. In scenario **(B)**, we adopt a lower mass limit, ${}^{B}M_a < M_{vir,2}$, so *halo 2* is now included in the assignment procedure. The galaxies indicated in the figure are then assigned to *halo 2*, no longer contributing to the density profile of *halo 1*.

Trevisan *et al.* (2017b) concluded that if M_a is small, the outer density profile of groups falls rapidly, while if M_a is large, it falls slowly. So they find that there is an optimal M_a for which the stacked density profiles follow the NFW model within 0.1 dex accuracy out to $\approx 13 r_{vir}$ in real space, corresponding to $\approx 10 r_{vir}$ in projection, for both the simulations and the observations. In this work, we adopt two values of M_a: the optimal value found by Trevisan *et al.* (2017b), $\log_{10}(M_a/M_\odot)$ = 12.3, and $\log_{10}(M_a/M_\odot)$ = 13.0. The second M_a value was chosen to remove small haloes from the vicinity of larger ones and explore how the pre-processing affects the variation of f_{SF} with halo-centric distance.

4.3 Results from assignment

To avoid incomplete profiles of SDSS groups, we ensure that at least 95% of the region within 20 r_{vir} from the group centre lies within the SDSS coverage area. In order to do this, we adopted the SDSS-DR7 spectroscopic angular selection function mask provided

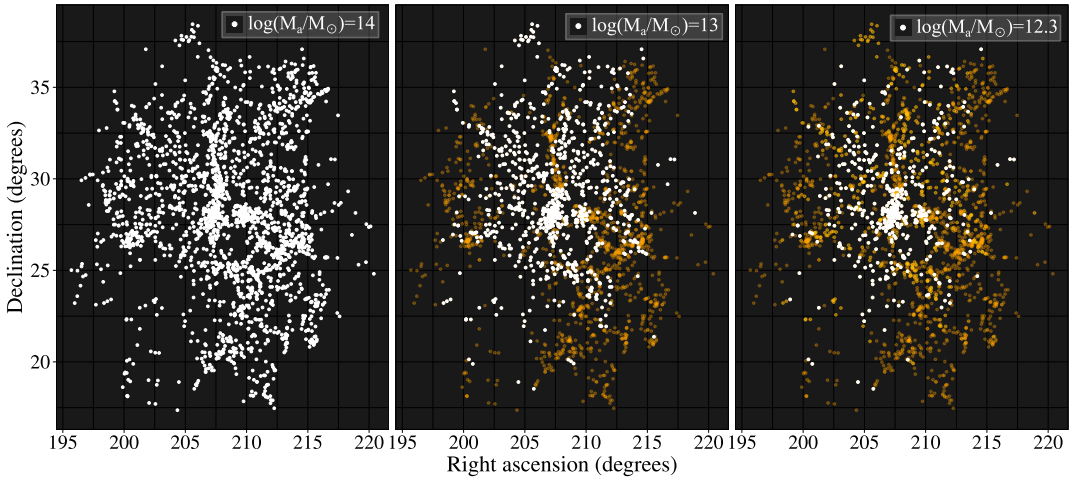

FIGURE 4.3: Illustration of how the number of galaxies assigned to a given halo changes for different values of the mass threshold M_a. In all panels, we show the same set of galaxies. White dots indicate a halo with the same central galaxy and a different M_a, while orange dots represent galaxies assigned to other halos. In the left panel, $\log_{10}(M_a/M_\odot) = 14.0$ and all galaxies are assigned to the same halo. In the middle and right panels, haloes with masses greater than $\log_{10}(M_a/M_\odot) = 13.0$ and $\log_{10}(M_a/M_\odot) = 12.3$, respectively, are allowed to host galaxies.

by the NYU Value-Added Galaxy Catalog team (Blanton *et al.*, 2005) and assembled with the package `Mangle 2.1` (Hamilton and Tegmark, 2004; Swanson *et al.*, 2008).

Furthermore, all groups for which 20 $R_{\rm vir}$ is not entirely contained within the redshift limits of our samples are discarded. In Table 4.1, we show the final results of this cleaning step. We assign the `flag_good` indicator to a value of 1 for groups which are within the SDSS area and redshift limits up to 20 $R_{\rm vir}$.

TABLE 4.1: Final group samples. Columns are, from left to right, M_a, the maximum redshift, the maximum halo mass, the number of groups, the number of satellites and the final number of galaxies in haloes with `flag_good=1`.

$\log_{10}(M_a/{\rm M}_\odot)$	$z_{\rm max}$	$\log_{10}(M_{\rm h,max}/{\rm M}_\odot)$	$N_{\rm groups}$	$N_{\rm gals}$	`flag_good=1`
12.3	0.03	14.68	1070	16614	8957
13.0	0.03	14.68	219	15708	7079
12.3	0.1	15.32	48061	194865	178742
13.0	0.1	15.32	10728	227296	204858

Chapter 5

Final subsamples

In this chapter, we describe the galaxies subsamples used in our analysis, selected from the sample from Section 3.4. The subsamples are defined to be complete in stellar mass within a given volume, as explained in Section 5.1. We also describe how we classify the galaxies as star-forming or quiescent (Section 5.2) and as AGN or non-AGN (Section 5.3). A summary of the subsamples is given in Section 5.4.

5.1 Stellar mass completeness

When estimating fractions of star-forming galaxies, we must take into account that, at a given stellar mass, galaxies with active star formation will be more luminous. We must thus ensure that the passive galaxies are not missed; hence $f_{\rm SF}$ is not overestimated because of this selection effect against high mass-to-light ratio galaxies. Therefore, a sample complete in stellar mass is desirable.

We estimated the minimum $M_\star$ for which the sample of galaxies within a given volume is complete following an approach similar to those described in La Barbera *et al.* (2010), Trevisan *et al.* (2017a) and Mamon *et al.* (2020). To cover a wide range of galaxy stellar masses, we considered two different volumes within $z_{\rm max} = 0.03$ and $z_{\rm max} = 0.1$, and estimated the minimum stellar mass, $M_{\star,\rm min}$, for which the subsamples are 95% complete. We first determine the 95 percentile of apparent Petrosian magnitude, $m_r^{\rm Petro}$, in bins of stellar mass and then perform a linear fit to the 95-percentile points. Then, the value of $\log M_\star$ where the best-fit line intersects $m_r^{\rm Petro} = 17.77$ defines the minimum stellar mass for 95% completeness of the subsample, $M_{\star,\rm min}$. This procedure is illustrated in Figure 5.1. We obtain $\log_{10}(M_{\star,\rm min}/{\rm M}_\odot) = 9.17$ and 10.5 for $z_{\rm max} = 0.03$ and $z_{\rm max} = 0.1$, respectively.

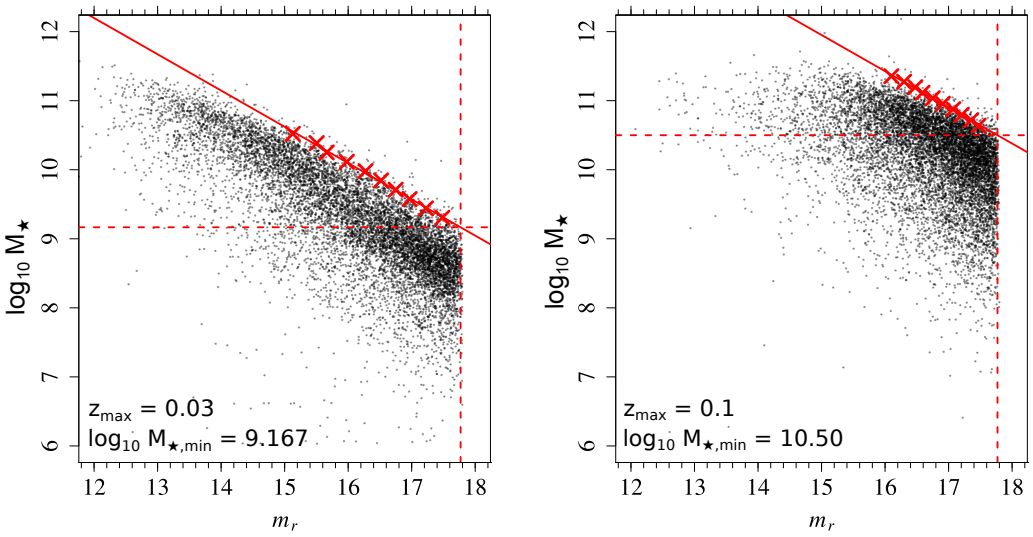

FIGURE 5.1: Mass completeness limit for the two samples, $z_{\rm max} = 0.03$ (*left panel*) and $z_{\rm max} = 0.1$ (*right panel*). On the x-axis is the Petrosian magnitude in the *r* band, and on the y-axis is the $\log_{10} M_\star$. In both panels, the red symbols are the 95 percentile of $m_r^{\rm Petro}$ in bins of $\log_{10} M_\star$, and the solid red lines are the linear best-fit these points. The vertical dashed red lines indicate the completeness limit of the SDSS spectroscopic catalogue ($m_r^{\rm Petro} = 17.77$), and the horizontal dashed red lines show the minimum stellar mass for 95% completeness.

5.2 Star-forming/Quiescent classification

To classify the galaxies between *star-forming* or *quiescent*, we identify their positions in the sSFR *vs.* $M_\star$ diagram. Galaxies that are located in the upper part of the diagram have high sSFRs, i.e., are actively forming stars at a high rate ("star-forming" galaxies). On the other hand, galaxies that are located in lower regions in the diagram have low sSFRs and are no longer forming stars at a significant rate ("quiescent" galaxies). In Figure 5.2, we show the 2D density plot of the sSFR *vs.* $M_\star$ diagram for the MPA-JHU (left) and GSWLC (right) measures. The level parameter (colours) denotes the quantiles of density, i.e., the areas where the density within the contour lines is 0%, 25%, 50%, 75%, and 100% of the maximum density. The line that divides the two main regions in the MPA-JHU sSFR *vs.* $M_\star$ diagram was determined by Knobel *et al.* (2015). The line in the GSWLC diagram was determined by eye in order to optimise the separation of the two regions. It is worth mentioning that we applied some clustering techniques to define the dividing line and found that the result is not significantly different from that shown in Figure 5.2. We opted to keep the line defined by eye to be consistent with Knobel *et al.* (2015), who adopted the same approach. The coefficients of the lines are

$$\log(\mathrm{sSFR}/\mathrm{yr}^{-1}) = \begin{cases} -7.85 - 0.30\log(M_\star/\mathrm{M}_\odot) & \text{for MPA} - \text{JHU} \\ -6.55 - 0.45\log(M_\star/\mathrm{M}_\odot) & \text{for GSWLC}\,. \end{cases} \tag{5.1}$$

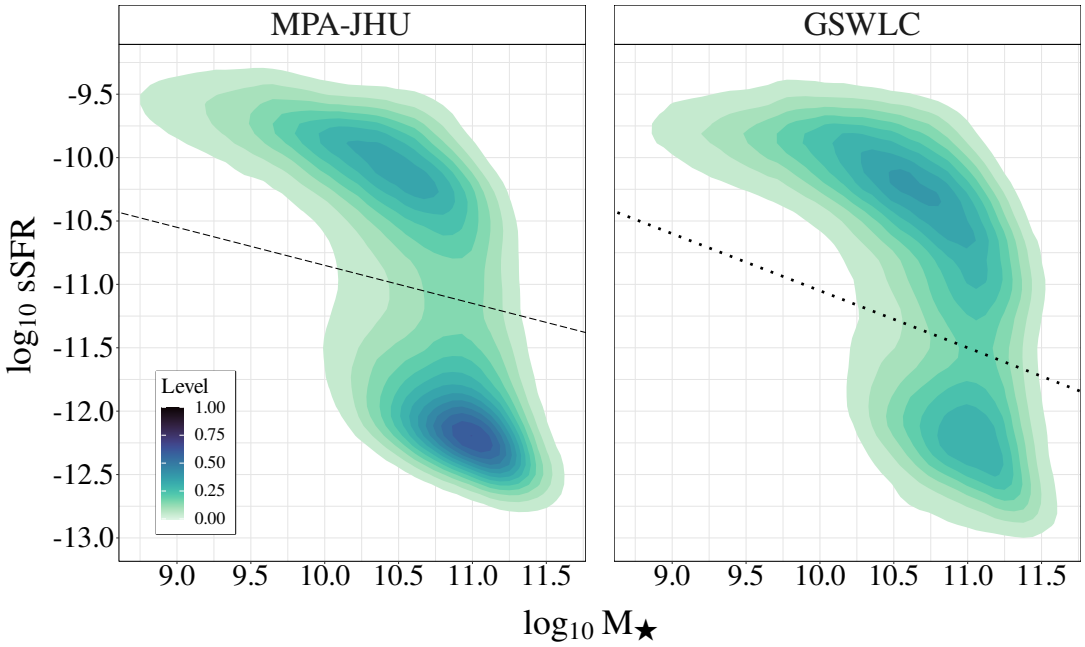

FIGURE 5.2: Specific star formation rates as a function of stellar masses. The *left* and *right panels* show sSFR and $M_\star$ estimates from the MPA-JHU and GSWLC catalogues, respectively. The separation between star-forming and quiescent galaxies is shown as dashed and dotted lines.

5.3 AGN/Non-AGN classification

A large population of weak line galaxies is often left out of statistical studies on emission line galaxies due to the absence of an adequate classification scheme since classical diagnostic diagrams, like [O III]/Hβ vs [N II]/Hα (the BPT diagram, Baldwin *et al.*, 1981), require the measurement of at least 4 emission lines. To avoid this, we opted to use more economical diagrams regarding the number of lines required (but see discussion in chapter 8). Therefore, to classify galaxies between AGN and non-AGN, we use the WHAN diagram proposed by Fernandes *et al.* (2011). The WHAN diagram, shown in Figure 5.3, requires only [NII]λ6584 and Hα emission lines. Using this diagram, we classify a galaxy as AGN if the flux ratio between the [NII]λ6584 and Hα emission lines is $\log_{10}$([NII]λ6584/Hα) ≥ -0.4 and the equivalent width of Hα is EW(Hα) ≥ 3.

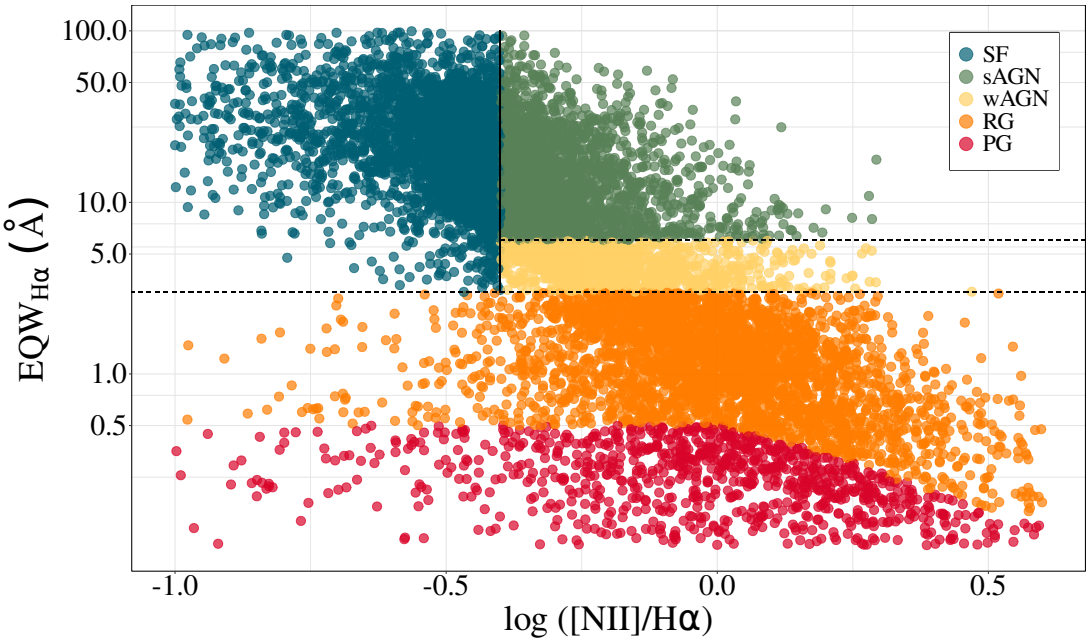

FIGURE 5.3: WHAN diagram. Different classes of galaxies are indicated by different colours: star-forming (*blue*); strong AGN (*green*); weak AGN (*yellow*); retired galaxies (*orange*); and quiescent galaxies (*red*). Only 2.5% of the objects are plotted to avoid overcrowding.

5.4 Final subsamples

In order to account for a variety of galaxy stellar masses, we defined two subsamples of galaxies within different volumes: one within $z_{\rm max} = 0.03$ and complete in stellar mass down to $\log(M_{\star,\rm min}/{\rm M}_\odot) = 9.17$; and other within $z_{\rm max} = 0.1$ and completeness limit $\log(M_{\star,\rm min}/{\rm M}_\odot) = 10.5$. For each of these two subsamples, we classify the galaxies as star-forming or quiescent using two sets of sSFR and $M_\star$ estimates, MPA-JHU and GSWLC. Therefore, we have four sets of data: `GSWLC003` and `GSWLC01`, containing galaxies within $z_{\rm max} = 0.03$ and 0.1 and SFR/$M_\star$ estimates from the GSWLC catalogue; and `MPAJHU003` and `MPAJHU01` containing galaxies within $z_{\rm max} = 0.03$ and 0.1 and SFR/$M_\star$ estimates from the MPA-JHU catalogue. A summary of the four datasets is given in Table 5.1. The table contains, from left to right, the sample ID, the catalogue for the measures of SFR and $M_\star$, the maximum redshift, the minimum value of $M_\star$, the number of clusters/groups in the sample and the number of galaxies.

TABLE 5.1: Summary of the galaxy final subsamples and datasets. The columns are, from left to right, sample identification, the catalogue from which the measures of SFR and $M_\star$ were taken, the maximum redshift, the minimum value of $M_\star$, and the number of clusters/groups and the number of galaxies. Galaxies were assigned using $\log_{10}(M_a/M_\odot) = 12.3$.

Sample	Catalogue	z_{max}	$\log_{10}(M_{\star,min}/M_\odot)$	N_{groups}	N_{gals}
`GSWLC01`	GSWLC	0.10	≥ 10.5	14612	38786
`GSWLC003`	GSWLC	0.03	≥ 9.17	532	4174
`MPAJHU01`	MPA-JHU	0.10	≥ 10.5	12873	31697
`MPAJHU003`	MPA-JHU	0.03	≥ 9.17	527	3728

Chapter 6

Exploring the relationship between galaxy properties

The general objective of this work is to investigate how the fraction of star-forming satellite galaxies (f_{SF}) varies across large distances away from the centre of their host halos. To accomplish this, we want to model if a satellite galaxy is star-forming or quiescent ($SF_{\mathrm{satellite}}$), the *outcome variable*, as a function of several galaxy properties, the *explanatory variables*. We are looking for properties that are strongly related to the outcome and, at the same time, minimally related to each other since correlated variables, in addition to not adding new information to the model, break an important assumption of regression models: that the explanatory variables are independent of each other. In this way, we want to select only the necessary explanatory variables while achieving a desired level of goodness of fit (*principle of parsimony*).

To accomplish this, in this Chapter, we explore the relationship between explanatory variables and their connection with the $SF_{\mathrm{satellite}}$. In Section 6.1, we present the initial galaxy properties and their basic statistics while examining the association between them. In Section 6.2, we quickly go over basic terms and concepts related to probabilistic models before presenting the logistic model functional form, an appropriate regression for modelling the dichotomous outcomes as the $SF_{\mathrm{satellite}}$. Finally, in Section 6.3, we present the statistical strategy chosen to build a parsimonious and robust logistic model to explain the $SF_{\mathrm{satellite}}$ through galaxy properties.

The analysis and methods applied in this part of the work are based on the sample of galaxies `GSWLC01` that uses estimates of SFR and $\log_{10} M_\star$ from GSWLC catalogue with maximum redshift $z_{\mathrm{max}} = 0.1$. Also, variables that correspond to group/cluster information come from groups with $\log_{10} M_a = 12.3$ (the use of this value has already been explained in Section 4.2).

6.1 Correlation between variables

First of all, we present in Table 6.1 all variables used in this work. The table contains the variable name, their function in modelling the $SF_{\mathrm{satellite}}$, their type and a brief description of their measurements and how they were obtained. The $SF_{\mathrm{satellite}}$ variable is our dichotomous categorical outcome. It indicates whether a satellite galaxy is star-forming or not (taking the value one (1) when the galaxy is star-forming or zero (0) if the galaxy is

Table 6.1: Summary of variables used in this work. The first column is the variable name, the second column is their function in the model, the third column is their type, and the fourth column is a brief description of each variable.

Variable	Function	Type	Description
$SF_{\text{satellite}}$	Outcome	Dichotomous categorical	Indicator if the satellite galaxy is star-forming (=1) or quiescent (=0) based on the classification made in Section 5.2.
$\log_{10} M_{\text{h}}$	Explanatory	Numerical continuous	Common logarithm of host group/cluster halo mass, in units of $M_{\odot}$. The measurement was obtained from Lim *et al.* (2017).
$\log_{10} M_{\star}$	Explanatory	Numerical continuous	Common logarithm of the galaxy's stellar mass, in units of $M_{\odot}$. The measurement was obtained from Kauffmann *et al.* (2003) and Salim *et al.* (2007).
$\log_{10} R_{\text{proj}}/R_{\text{vir}}$	Explanatory	Numerical continuous	Common logarithm of the ratio of the projected galaxy radius to the virial radius.
$\log_{10} \sigma_{\text{satellite}}$	Explanatory	Numerical continuous	Common logarithm of the velocity dispersion of the satellite galaxy within a half-light aperture radius (the σ_{e} from Equation 3.2), in units of km/s. The measurement was obtained through the Appendix A query.
$\log_{10} \sigma_{\text{central}}$	Explanatory	Numerical continuous	Common logarithm of the velocity dispersion of the central galaxy within a half-light aperture radius, in units of km/s. The measurement was obtained through the Appendix A query.
$v_{\text{los}}/v_{\text{vir}}$	Explanatory	Numerical continuous	Ratio of line-of-sight velocity to virial velocity.
SF_{central}	Explanatory	Dichotomous categorical	Indicator if the central galaxy is a central star-forming (=1) or a central quiescent (=0) based on the classification made in Section 5.2.
$AGN_{\text{satellite}}$	Explanatory	Dichotomous categorical	Indicator if the satellite galaxy is AGN (=1) or non-AGN (=0) based on the classification made in Section 5.3.
AGN_{central}	Explanatory	Dichotomous categorical	Indicator if the central galaxy is a central AGN (=1) or a central non-AGN (=0) based on the classification made in Section 5.3.

quiescent). The other nine features are explanatory variables (six numerical and three categorical). When there is the same variable referring to satellite galaxies and another referring to central galaxies, this is informed by the subscript. If not informed, the measurement will always refer to satellites. This table is important to define the range of values of the variables we are working with and clarify that the results are valid in this regime.

Explanatory variables that are correlated with each other would explain a reasonable part of the outcome in the same way. But, besides that, correlated explanatory variables cause a regression problem called *multicollinearity*. If the multicollinearity is high enough, it could affect the fitted model and the interpretation of their coefficients. A regression coefficient represents the mean change in the outcome variable for each 1-unit change in an explanatory variable when you hold all of the other explanatory variables constant. So if the explanatory variables are correlated, it indicates that changes in one

variable are associated with shifts in another variable. The stronger the correlation, the more difficult it is to change one variable without changing another. Multicollinearity affects the coefficient estimates, causing them to oscillate wildly and making them very susceptible to small changes in the model. Also, the precision of the estimated coefficients is reduced, which weakens the statistical power of the regression model, making the p-values unreliable. In this way, before trying to do any inferential model, we need to understand the variables' basic aspects and analyse how and how much they relate to each other.

Table 6.2 presents the summary statistics of the six numerical variables used in this work for the two samples of $z_{\max}$. For each feature is presented the minimum value, the 25th percentile, the median value, the 75th percentile, the maximum value, the mean value and the standard deviation. Table 6.3 describes two samples of $z_{\max}$ the six contingency tables for the two-by-two combination of the four categorical variables. Table A presents the total amount of galaxies and the proportion for each combination of the levels of the outcome variable, $SF_{\text{satellite}}$, with the three explanatory categorical variables. Table B presents the total amount of galaxies and the proportion for the combination of the levels of the three explanatory categorical variables with each other. The amounts and proportions presented here are related to the measures of SFR taken from the GSWLC catalogue, but the same contingency tables for the MPA-JHU measures can be found in Table 6.4.

To investigate the association between the numerical variables, we compute Pearson's correlation coefficient (r, Bravais 1846) given by

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{Y})^2}}, \tag{6.1}$$

where n is the sample size, x_i, y_i are the individual sample points of the numerical variables x and y, respectively, and $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean of variable X (and analogously for $\overline{Y}$).

The association between numerical and categorical dichotomous variables can be measured using the point-biserial correlation (r_{pb}, Gupta 1960), which is mathematically equivalent to Pearson's correlation coefficient. If X is a continuous variable and Y is a dichotomous variable, this implies $r_{XY} = r_{pb}$. The point-biserial correlation is defined by

$$r_{pb} = \frac{M_1 - M_0}{s_n}\sqrt{\frac{n_1 n_0}{n^2}} \tag{6.2}$$

where M_1 is the mean of X for all data points where $Y = 1$, M_0 is the mean of X for all data points where $Y = 0$, n_1 is the number of data points which $Y = 1$, n_0 is the number of data points which $Y = 0$, n is the total sample size and s_n is the sample standard deviation of X given by

$$s_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}. \tag{6.3}$$

The calculated correlations between pairs of numerical variables and dichotomous ones with numerical ones are presented in Figure 6.1. The colour grid indicates the correlation intervals, where the reddest colours indicate a negative correlation and the bluest colours indicate a positive correlation (grey squares are relative to pairs of

TABLE 6.2: Summary statistics of numerical variables. Table (A) refers to the sample with $z_{max} = 0.1$, and table (B) refers to the sample with $z_{max} = 0.03$. Each column displays, from left to right, the variable name, the minimum value (Min), the 25th percentile (25th perc), the median (Median), the 75th percentile (75th perc), the maximum value (Max), the mean (Mean) and the standard deviation (SD).

(A) Sample with $z_{max} = 0.1$

Variable	Min	25th perc	Median	75th perc	Max	Mean	SD
$\log_{10} M_{h}$	12.3	12.829	13.334	13.925	15.323	13.419	0.716
$\log_{10} M_{\star\,(\mathrm{GSWLC})}$	10.5	10.575	10.661	10.78	11.936	10.701	0.166
$\log_{10} M_{\star\,(\mathrm{MPA-JHU})}$	10.5	10.572	10.655	10.774	12.133	10.698	0.168
$\log_{10} R_{proj}/R_{vir}$	-2.207	-0.258	0.145	0.622	1.293	0.149	0.562
$\log_{10} \sigma_{satellite}$	1.102	2.009	2.105	2.19	2.933	2.096	0.14
$\log_{10} \sigma_{central}$	1.078	2.138	2.255	2.338	3.014	2.223	0.164
v_{los}/v_{vir}	10^{-5}	0.171	0.376	0.68	2.889	0.476	0.398

(B) Sample with $z_{max} = 0.03$

Variable	Min	25th perc	Median	75th perc	Max	Mean	SD
$\log_{10} M_{h}$	12.3	12.608	12.99	13.342	14.157	13.043	0.512
$\log_{10} M_{\star\,(\mathrm{GSWLC})}$	9.167	9.446	9.788	10.15	11.073	9.818	0.418
$\log_{10} M_{\star\,(\mathrm{MPA-JHU})}$	9.167	9.444	9.763	10.108	11.235	9.795	0.404
$\log_{10} R_{proj}/R_{vir}$	-1.43	0.047	0.498	0.794	1.293	0.399	0.503
$\log_{10} \sigma_{satellite}$	1.042	1.583	1.729	1.891	2.919	1.721	0.265
$\log_{10} \sigma_{central}$	1.033	1.697	1.94	2.082	2.884	1.879	0.266
v_{los}/v_{vir}	10^{-5}	0.222	0.467	0.824	3.226	0.587	0.485

TABLE 6.3: Contingency tables for categorical variables of the GSWLC sample. The values are the number of galaxies in each combination of categorical variable pair levels followed by the proportion inside the parentheses.

(A) Contingency tables between the outcome variables $SF_{\text{satellite}}$ and the explanatory categorical variables

		$SF_{\text{satellite}}$ Star-forming		$SF_{\text{satellite}}$ Quiescent		Total	
		$z_{\text{max}} = 0.1$	$z_{\text{max}} = 0.03$	$z_{\text{max}} = 0.1$	$z_{\text{max}} = 0.03$	$z_{\text{max}} = 0.1$	$z_{\text{max}} = 0.03$
SF_{central}	Star-forming	10034 (26%)	2075 (50%)	7129 (18%)	645 (15%)	17163 (44%)	2720 (65%)
	Quiescent	11712 (30%)	1014 (24%)	9911 (26%)	440 (11%)	21623 (56%)	1454 (35%)
$AGN_{\text{satellite}}$	AGN	12431 (32%)	639 (15%)	744 (2%)	72 (2%)	13175 (34%)	711 (17%)
	Non-AGN	9315 (24%)	2450 (59%)	16296 (42%)	1013 (24%)	25611 (66%)	3463 (83%)
AGN_{central}	AGN	4862 (13%)	842 (20%)	3272 (8%)	341 (8%)	8134 (21%)	1183 (28%)
	Non-AGN	16884 (44%)	2247 (54%)	13768 (35%)	744 (18%)	30652 (79%)	2991 (72%)
	Total	21746 (56%)	3089 (74%)	17040 (44%)	1085 (26%)	38786 (100%)	4174 (100%)

(B) Contingency tables between explanatory categorical variables

		AGN_{central} Central AGN		AGN_{central} Central Non-AGN		Total	
		$z_{\text{max}} = 0.1$	$z_{\text{max}} = 0.03$	$z_{\text{max}} = 0.1$	$z_{\text{max}} = 0.03$	$z_{\text{max}} = 0.1$	$z_{\text{max}} = 0.03$
SF_{central}	Central Star-forming	7412 (19%)	1133 (27%)	9751 (25%)	1587 (38%)	17163 (44%)	2720 (65%)
	Central Quiescent	722 (2%)	50 (1%)	20901 (54%)	1404 (34%)	21623 (56%)	1454 (35%)
	Total	8134 (21%)	1183 (28%)	30652 (79%)	2991 (72%)	38786 (100%)	4174 (100%)

		$AGN_{\text{satellite}}$ AGN		$AGN_{\text{satellite}}$ Non-AGN		Total	
		$z_{\text{max}} = 0.1$	$z_{\text{max}} = 0.03$	$z_{\text{max}} = 0.1$	$z_{\text{max}} = 0.03$	$z_{\text{max}} = 0.1$	$z_{\text{max}} = 0.03$
AGN_{central}	Central AGN	3064 (8%)	196 (5%)	5070 (13%)	987 (24%)	8134 (21%)	1183 (28%)
	Central Non-AGN	10111 (26%)	515 (12%)	20541 (53%)	2476 (59%)	30652 (79%)	2991 (72%)
	Total	13175 (34%)	711 (17%)	25611 (66%)	3463 (83%)	38786 (100%)	4174 (100%)

		$AGN_{\text{satellite}}$ AGN		$AGN_{\text{satellite}}$ Non-AGN		Total	
		$z_{\text{max}} = 0.1$	$z_{\text{max}} = 0.03$	$z_{\text{max}} = 0.1$	$z_{\text{max}} = 0.03$	$z_{\text{max}} = 0.1$	$z_{\text{max}} = 0.03$
SF_{central}	Central Star-forming	6218 (16%)	469 (11%)	10945 (28%)	2251 (54%)	17163 (44%)	2720 (65%)
	Central Quiescent	6957 (18%)	242 (6%)	14666 (38%)	1212 (29%)	21623 (56%)	1454 (35%)
	Total	13175 (34%)	711 (17%)	25611 (66%)	3463 (83%)	38786 (100%)	4174 (100%)

TABLE 6.4: Contingency tables for categorical variables of the MPA-JHU sample. The values are the number of galaxies in each combination of categorical variable pair levels followed by the proportion inside the parentheses.

(A) Contingency tables between the outcome variables $SF_{\text{satellite}}$ and the explanatory categorical variables

		$SF_{\text{satellite}}$ Star-forming		$SF_{\text{satellite}}$ Quiescent		Total	
		$z_{\text{max}} = 0.1$	$z_{\text{max}} = 0.03$	$z_{\text{max}} = 0.1$	$z_{\text{max}} = 0.03$	$z_{\text{max}} = 0.1$	$z_{\text{max}} = 0.03$
SF_{central}	Star-forming	2328 (7%)	1153 (31%)	4749 (15%)	562 (15%)	7077 (22%)	1715 (46%)
	Quiescent	6848 (22%)	1110 (30%)	17772 (56%)	903 (24%)	24620 (78%)	2013 (54%)
$AGN_{\text{satellite}}$	AGN	6894 (22%)	540 (14%)	3045 (10%)	166 (4%)	9939 (31%)	706 (19%)
	Non-AGN	2282 (7%)	1723 (46%)	19476 (61%)	1299 (35%)	21758 (69%)	3022 (81%)
AGN_{central}	AGN	2098 (7%)	627 (17%)	4383 (14%)	428 (11%)	6481 (20%)	1055 (28%)
	Non-AGN	7078 (22%)	1636 (44%)	18138 (57%)	1037 (28%)	25216 (80%)	2673 (72%)
	Total	9176 (29%)	2263 (61%)	22521 (71%)	1465 (39%)	31697 (100%)	3728 (100%)

(B) Contingency tables between explanatory categorical variables

		AGN_{central} Central AGN		AGN_{central} Central Non-AGN		Total	
		$z_{\text{max}} = 0.1$	$z_{\text{max}} = 0.03$	$z_{\text{max}} = 0.1$	$z_{\text{max}} = 0.03$	$z_{\text{max}} = 0.1$	$z_{\text{max}} = 0.03$
SF_{central}	Central Star-forming	4442 (14%)	723 (19%)	2635 (8%)	992 (27%)	7077 (22%)	1715 (46%)
	Central Quiescent	2039 (6%)	332 (9%)	22581 (71%)	1681 (45%)	24620 (78%)	2013 (54%)
	Total	481 (20%)	1055 (28%)	25216 (80%)	2673 (72%)	31697 (100%)	3728 (100%)

		$AGN_{\text{satellite}}$ AGN		$AGN_{\text{satellite}}$ Non-AGN		Total	
		$z_{\text{max}} = 0.1$	$z_{\text{max}} = 0.03$	$z_{\text{max}} = 0.1$	$z_{\text{max}} = 0.03$	$z_{\text{max}} = 0.1$	$z_{\text{max}} = 0.03$
AGN_{central}	Central AGN	2262 (7%)	196 (5%)	4219 (13%)	859 (23%)	6481 (20%)	1055 (28%)
	Central Non-AGN	7677 (24%)	510 (14%)	17539 (55%)	2163 (58%)	25216 (80%)	2673 (72%)
	Total	9939 (31%)	706 (19%)	21758 (69%)	3022 (81%)	31697 (100%)	3728 (100%)

		$AGN_{\text{satellite}}$ AGN		$AGN_{\text{satellite}}$ Non-AGN		Total	
		$z_{\text{max}} = 0.1$	$z_{\text{max}} = 0.03$	$z_{\text{max}} = 0.1$	$z_{\text{max}} = 0.03$	$z_{\text{max}} = 0.1$	$z_{\text{max}} = 0.03$
SF_{central}	Central Star-forming	2508 (8%)	351 (9%)	4569 (14%)	1364 (37%)	7077 (22%)	1715 (46%)
	Central Quiescent	7431 (23%)	355 (10%)	17189 (54%)	1658 (44%)	24620 (78%)	2013 (54%)
	Total	9939 (31%)	706 (19%)	21758 (69%)	3022 (81%)	31697 (100%)	3728 (100%)

categorical variables, and their correlation will be measured using another indicator later in this section). The intensity of the colour indicates the strength of the correlation. In the following list, we summarise the correlations that draw the most attention. Also, the mentioned correlations are highlighted in the diagram of Figure 6.1 with a black stroke.

Negatively correlated variables:

- $\log_{10}\sigma_{\text{central}}$ and SF_{central} ($r_{pb} = -0.584$);
- $SF_{\text{satellite}}$ and $\log_{10}\sigma_{\text{satellite}}$ ($r_{pb} = -0.522$);
- $AGN_{\text{satellite}}$ and $\log_{10}\sigma_{\text{satellite}}$ ($r_{pb} = -0.368$);
- $\log_{10}\sigma_{\text{central}}$ and AGN_{central} ($r_{pb} = -0.361$);
- $\log_{10} M_\star$ and $\log_{10} R_{\text{proj}}/R_{\text{vir}}$ ($r_{XY} = -0.289$);
- $\log_{10} M_{\text{h}}$ and SF_{central} ($r_{pb} = -0.276$).

Positively correlated variables:

- $\log_{10} M_{\text{h}}$ and $\log_{10}\sigma_{\text{central}}$ ($r_{XY} = 0.499$);
- $\log_{10} M_\star$ and $\log_{10}\sigma_{\text{satellite}}$ ($r_{XY} = 0.385$).

The scatter plots in Figure 6.2 can help visualise the correlations between the numerical variables. It shows the points for a 10% random sample of our data (due to file size constraints) with a simple regression line. The correlations between the numerical variables already mentioned are highlighted in the diagram with a black stroke.

However, we cannot apply Pearson's correlation coefficient to measure the association between categorical variables. Instead, we can use the χ^2 test of independence (McHugh, 2013) to determine whether two categorical variables are likely to be related or not. The χ^2 test of independence is a statistical hypothesis test for contingency tables like the ones in Table 6.3. Its null hypothesis (H_0) states that there is no association between the two variables, and the alternative hypothesis (H_a) states that there is an association between the two variables. The χ^2 statistic is given by the equation

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \tag{6.4}$$

where O_{ij} is the observed value of the cell in row i and column j of the contingency table and E_{ij} is the expected value for that cell calculated as

$$E_{ij} = \frac{\sum_{k=1}^{c} O_{ij}\sum_{k=1}^{r} O_{kj}}{N}, \tag{6.5}$$

where N is the sum of all cells in the contingency table.

For the six pairs of categorical variables, the H_0 of the χ^2 test was rejected with very small p-values, as shown in Table 6.5, indicating that all pairs of categorical variables are associated. Nonetheless, the p-value depends on the sample size of the data being tested, which produces small p-values as the sample size increases (Lin *et al.* 2013,

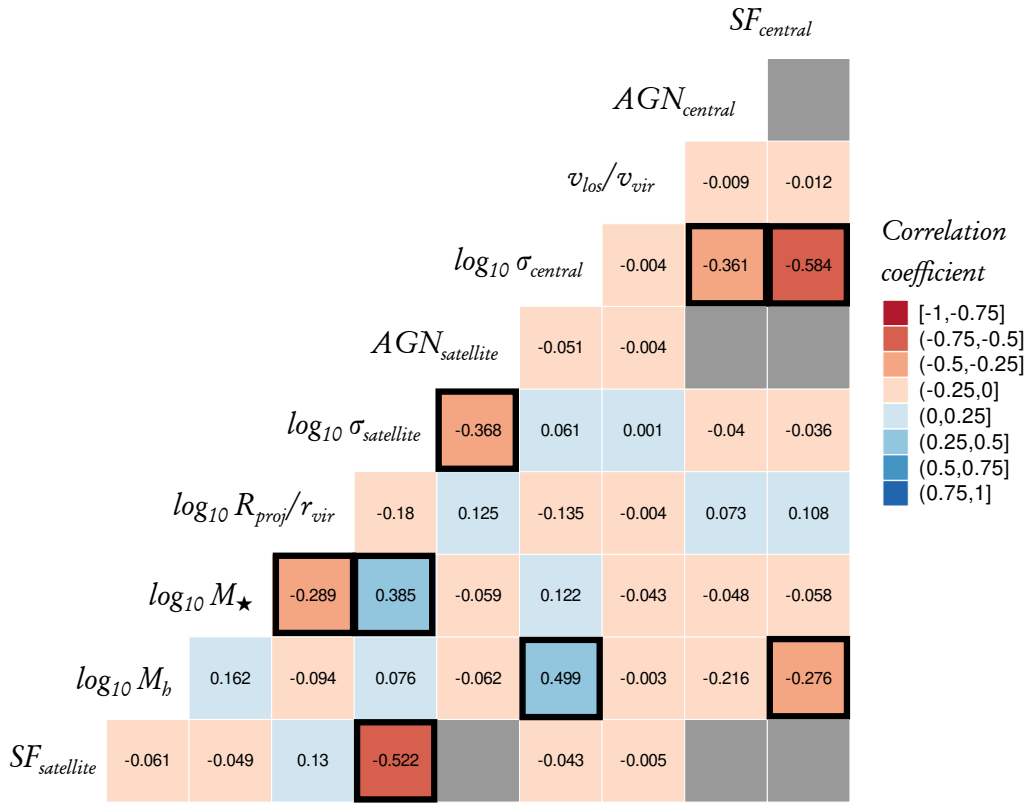

FIGURE 6.1: Correlation plot for the variables. The correlation between two numerical variables was computed using Pearson's correlation coefficient, and between a numerical and a categorical variable was computed using the point-biserial correlation. The colour grid indicates the correlation coefficient intervals, where the reddest colours indicate a negative correlation and the bluest colours indicate a positive correlation (grey squares are relative to pairs of categorical variables, and their correlation will be measured using another indicator later in this section). The intensity of the colour indicates the strength of the correlation. The squares with black strokes refer to the correlations mentioned in the text.

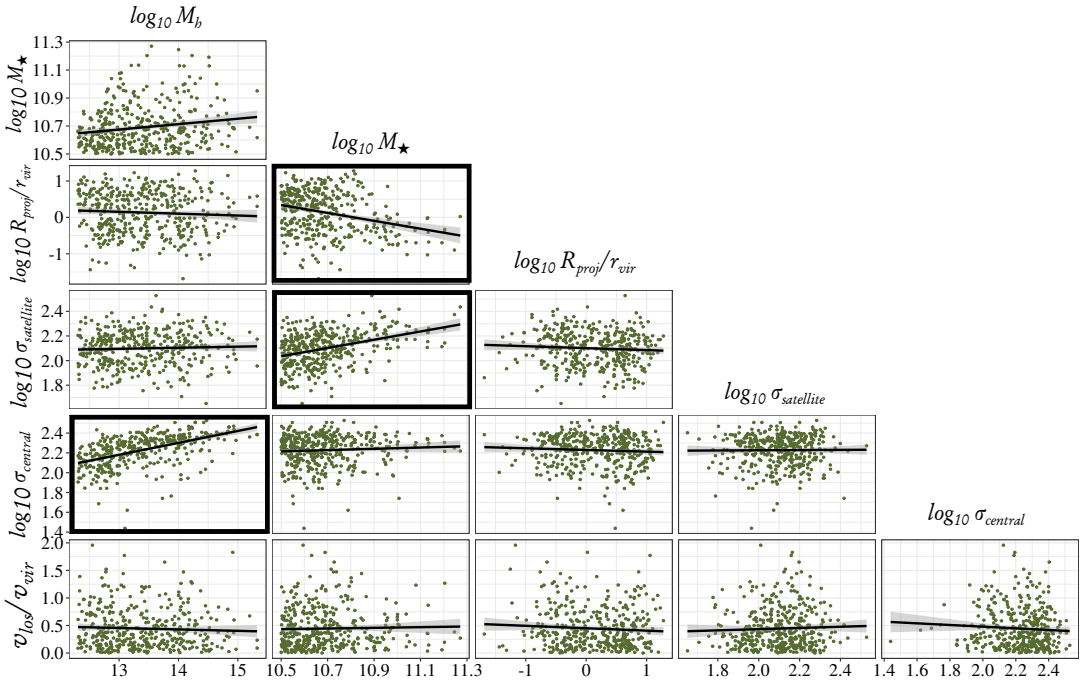

FIGURE 6.2: Scatter plot for numerical variables. The figure shows the points for a 10% random sample of our data (due to file size constraints) with a simple regression line. The rectangles with black strokes refer to the correlations mentioned in the text.

Gómez-de Mariscal *et al.* 2021). To overcome this problem, one can measure the effect size of the relationship between two variables (Kirk 1996, Sullivan and Feinn 2012). The effect size measures the strength of an association between two groups or variables. In this case, for 2x2 contingency tables, one of the most commonly used effect size measures is the ϕ coefficient (Yule, 1912), given by the equation

$$\phi = \sqrt{\frac{\chi^2}{N}}\,. \tag{6.6}$$

According to Cohen 1992, a ϕ of 0.1 indicates a small effect, 0.3 is a medium effect, and 0.5 implies a large effect. The ϕ coefficient for each pair of categorical variables is shown in Table 6.5. We can see that only $SF_{\text{satellite}}$ with $AGN_{\text{satellite}}$ and SF_{central} with AGN_{central} present a strong association with each other.

Finally, one might then be interested in finding what combination in the variable levels (the cell in the contingency table) contributes the most to the χ^2 statistic. To answer this, we can analyse the differences between observed and expected values (O_{ij} - E_{ij}), i.e., the residuals. The cells with the largest residuals might contribute the most to the χ^2 statistic. However, cells with larger cell counts will also have larger residuals, so to compare the residuals more fairly, we must standardise it by dividing by $\sqrt{E_{ij}}$. The standardised residuals are commonly called *Pearson residuals* and are defined as

$$r_{P,ij} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}} \tag{6.7}$$

In Figure 6.3, we show the Pearson residuals for the two pairs of categorical variables

TABLE 6.5: χ^2 test of independence information and effect size. The columns are, from left to right, the pair of variables, the calculated χ^2 statistic of the test, the p-value from the test and the effect size measured by the ϕ coefficient.

Pair of variables	χ^2 statistic	χ^2 p-value	ϕ coefficient
$SF_{\text{satellite}}$ vs. $AGN_{\text{satellite}}$	11871.380	2.2e-16	0.55
SF_{central} vs. AGN_{central}	9164.231	2.2e-16	0.49
AGN_{central} vs. $AGN_{\text{satellite}}$	62.632	2.5e-15	0.04
SF_{central} vs. $AGN_{\text{satellite}}$	69.964	2.2e-16	0.04
$SF_{\text{satellite}}$ vs. AGN_{central}	57.235	3.9e-14	0.04
$SF_{\text{satellite}}$ vs. SF_{central}	71.597	2.2e-16	0.04

where we find a strong association due to the great effect size, the $SF_{\text{satellite}}$ with $AGN_{\text{satellite}}$ and SF_{central} with AGN_{central}. Each circle is the Person residuals for the level combination of each categorical variable. The size of the circle and the intensity of its colour represent the absolute magnitude of the Pearson residuals, with blue colours representing positive residual values (observed values greater than expected) and red colours representing negative residual values (observed values smaller than expected). In the first panel (left to right), the number of central galaxies that are AGN and star-forming is *greater* than expected ($r_P = 63.55$, highlighted blue circle), indicating that the found association between SF_{central} and AGN_{central} is highly derived by the great number of central galaxies that are AGN and star-forming. In the second panel, the number of satellite galaxies that are AGN and quiescent is *lesser* than expected ($r_P = -66.3$, highlighted red circle), indicating that this relationship is the one that contributes the most to the association between $SF_{\text{satellite}}$ and $AGN_{\text{satellite}}$. The magnitude of Pearson residuals indicates that these two combinations of levels of these categorical variables are driving the significance of the χ^2 test, i.e., the differences found by the test are totally related to these two phenomena.

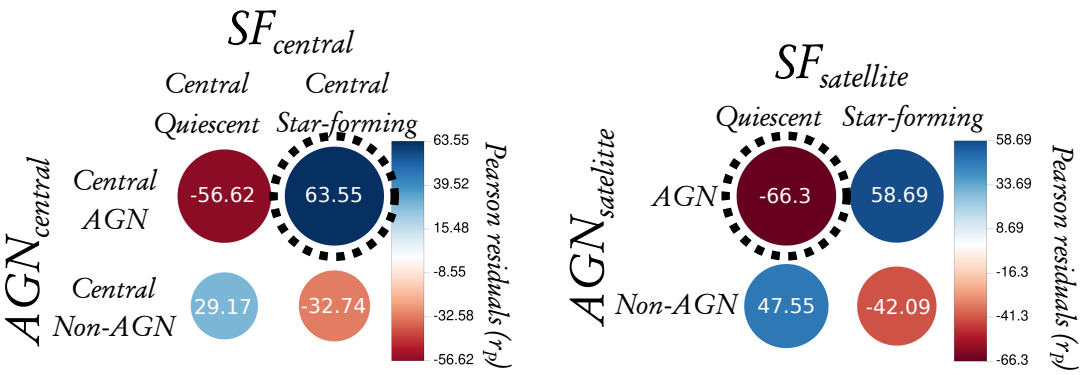

FIGURE 6.3: Pearson residuals of χ^2. Each circle is the level combination of each categorical variable. The size of the circle and the intensity of its colour represent the magnitude of the Pearson residuals, with blue colours representing positive residual values ($O_{ij} > E_{ij}$) and red colours representing negative residual values ($O_{ij} < E_{ij}$). The dashed-contoured circles indicate the residuals commented on the text.

With the analysis of the variable's descriptive statistics and the correlations between pairs of variables, we conclude that the following pairs of variables are the most associated with each other

- $\log_{10}\sigma_{\rm satellite}$ and $\log_{10}M_\star$;
- $\log_{10}M_\star$ and $\log_{10}R_{\rm proj}/R_{\rm vir}$;
- $\log_{10}M_{\rm h}$ and $\log_{10}\sigma_{\rm central}$;
- $\log_{10}\sigma_{\rm central}$ and $SF_{\rm central}$;
- $SF_{\rm satellite}$ and $\log_{10}\sigma_{\rm satellite}$;
- $AGN_{\rm satellite}$ and $\log_{10}\sigma_{\rm satellite}$;
- $\log_{10}M_{\rm h}$ and $SF_{\rm central}$;
- $\log_{10}\sigma_{\rm central}$ and $AGN_{\rm central}$;
- $SF_{\rm central}$ and $AGN_{\rm central}$;
- $SF_{\rm satellite}$ and $AGN_{\rm satellite}$.

However, to measure the impact of their correlation and finally choose which variables will collaborate more for a robust and parsimonious model, we must also evaluate their individual and combined effects within the model. But before that, we will briefly review some ideas about modelling and terminology and present the model's functional form chosen to describe how the $f_{\rm SF}$ galaxies vary along the clustercentric distance, the logistic regression.

6.2 Logistic regression

A model is a simplified abstraction of reality in that it provides an approximation of some relatively more complex phenomenon. Models can be broadly classified as *deterministic* or *probabilistic*. In a deterministic model, the outcome is *precisely* defined, often by a set of equations (e.g., the ideal gas law, $PV = nRT$, and the first law of thermodynamics, $\oint dW = j\oint dQ$). However, in a probabilistic model, the outcomes will exhibit *variability* because the model either contains random elements or is impacted in some way by random influences. In this work, we are interested in modelling the $SF_{\rm satellite}$, our

categorical outcome, through a probabilistic model. More precisely, $SF_{\text{satellite}}$ is categorical and *dichotomous*, as it can only have two results: "quiescent" or "star-forming".

Before looking at the methods for modelling the probability of occurrence of one of the two outcomes of a categorical dichotomous variable, let us remember, through modelling a continuous outcome, the essence of what we call linear regression. We can model a generic outcome variable (Y) by decomposing it into a function $f(\mathbf{X})$ dependent on a set of p generic explanatory variables ($\mathbf{X} = X_1, \ldots, X_p$) plus a random error term ε, in such a way that

$$Y_i = f(X_1, \ldots, X_p) + \varepsilon_i = f(\mathbf{X}) + \varepsilon_i \,. \tag{6.8}$$

where the index i is an observation of the outcome variable Y, with $i = 1, \ldots, n$.

Nevertheless, it is usually too ambitious to model all aspects of the outcome variable. Thus, it is common practice to analyse just how the *mean of* Y changes in relation to changes in $\mathbf{X}$. Thus, a typical statistical model can be expressed as an equation that associates the mean of the outcome variable to some function of a linear combination of the explanatory variables as follows

$$Y_i = \mathbb{E}[Y_i|\mathbf{X}] + \varepsilon_i \,, \tag{6.9}$$

which results in $f(\mathbf{X}) = \mathbb{E}[Y_i|\mathbf{X}]$. In the Equation 6.9, the quantity $\mathbb{E}[Y_i|\mathbf{X}]$ is called the *conditional mean*, i.e., the mean of Y given the presence of the set of explanatory variables X. In linear regression, the function $f(\mathbf{X})$ is called *linear predictor*, because it has the form

$$f(\mathbf{X}) = \eta(\mathbf{X};\boldsymbol{\beta}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \,, \tag{6.10}$$

where $f(\mathbf{X}) = \eta(\mathbf{X};\boldsymbol{\beta})$ is a *linear combination* of the explanatory variables. Here, it should be noted that $\eta(\mathbf{X};\boldsymbol{\beta})$ is linear in terms of the *parameters* (the βs) but not necessarily in terms of explanatory variables (in their original form): some of the explanatory variables may themselves be functions of the original explanatory variables. Higher powers (such as squares, cubes, etc.) of original variables are often referred to as *higher-order terms*. Additionally, the product of some explanatory variables is referred to *interaction terms*.

In this work, the outcome variable Y is a categorical variable having only two results (*star-forming* or *quiescent*), so its domain is not in $\mathbb{R}$ nor a categorical variable could In fact, our outcome variable is Bernoulli-distributed, given the explanatory variables, i.e.,

$$Y_i|X_1, \ldots, X_p \sim \text{Ber}(\pi_i) \,, \tag{6.11}$$

where the parameter π is the expected value of a Bernoulli random variable taking values in $0 \leq \pi \leq 1$. The probability mass function of a Bernoulli distribution is given by

$$P_{Y|X}(y) = \begin{cases} \pi & \text{for } y = 1 \\ 1 - \pi & \text{for } y = 0 \\ 0 & \text{otherwise} \,, \end{cases}$$

where, in our context, $y = 1$ means that the galaxy is star-forming, and $y = 0$ means that it is quiescent.

The mean π of the Bernoulli distribution can be considered as a *probability*. Therefore, to describe probabilities as functions of explanatory variables, we can use the

generalised linear model (GLM) where the dichotomous outcome will have the conditional mean belonging to $0 \leq \mathbb{E}(Y|x) \leq 1$. To model this conditional mean using a linear model, we have to link the average outcome π to the linear predictor $\eta(\mathbf{X};\boldsymbol{\beta})$ via a suitable transformation that maps the linear combination of explanatory variables onto the Bernoulli probability distribution with a domain from 0 to 1. In this case, the *link function* $h(.)$ that associate $\eta(\mathbf{X};\boldsymbol{\beta})$ to π is the *logit function*,

$$h(\pi) = \ln\left(\frac{\pi}{1-\pi}\right) = \eta(\mathbf{X};\boldsymbol{\beta})\,. \tag{6.12}$$

To obtain the value of the conditional mean, we need to isolate π. For this, we need to take the inverse of the logit function so that

$$h^{-1}(\pi) = \frac{\pi}{1-\pi} = \exp[\eta(\mathbf{X};\boldsymbol{\beta})]\,, \tag{6.13}$$

where $h^{-1}(\pi)$ is called the *sigmoid function*. Then we can isolate π so that

$$\pi = \frac{\exp[\eta(\mathbf{X};\boldsymbol{\beta})]}{1+\exp[\eta(\mathbf{X};\boldsymbol{\beta})]} = \frac{\exp(\beta_0+\beta_1X_1+\cdots+\beta_pX_p)}{1+\exp(\beta_0+\beta_1X_1+\cdots+\beta_pX_p)}\,, \tag{6.14}$$

where the equation 6.14 is the well-know *logistic function* (Berkson 1944; Cox 1958). Therefore, we use the logistic function to model the dichotomous outcome through the *logistic regression*. In a regression problem, the model's parameters $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ are unknown and must be estimated. The process of estimating these parameters (and calculating their errors, as well as various goodness of fit statistics) from the data will be referred to as *fitting* the model. To fit the model, we need to find the coefficients that minimise the difference between predicted probabilities and actual outcomes through Maximum Likelihood Estimation (MLE, Fisher 1922a; Fisher 1922b). The likelihood function represents the probability of observed outcomes given the logistic model, with the log-likelihood function simplifying calculations. Optimisation algorithms adjust coefficients to maximise the log-likelihood, converging to optimal or near-optimal values.

Finally, before applying a strategy to use logistic regression to model $SF_{\text{satellite}}$, we need to define the nomenclature used from now on. For the sake of simplicity, we will call the mathematical functional form of how the outcome variable relates to the explanatory variables as a *model*. Fitting this model to a sample will be called a *fit*. In this context, for example, the expression $y = a + bx + cz$ is a *model*, but $y = a_1 + b_1x + c_1z$ and $y = a_2 + b_2x + c_2z$ are two different *fits* of this model. Realise that the values of the adjustment coefficients can change depending on the sample used (the coefficient b, for example, has value b_1 in the first fit and value b_2 in the second fit), but the outcome variable (y), the explanatory variables (x and z) and the functional form (a simple linear model) with which they are related is fixed. Here, we continue using the `GSWLC01` sample as a foundation for choosing the best explanatory variables and functional form to model the $SF_{\text{satellite}}$. After this, we will use this same functional form and set of explanatory variables to independently model our three other samples, `GSWLC03`, `MPAJHU01` and `MPAJHU03`.

6.3 Model-building strategy

This work aims to study how the f_{SF} galaxies vary with large clustercentric distances through a logistic model, which will allow us to evaluate the f_{SF} in the entire available range of the explanatory variables, without the need to binning the data. To choose which explanatory variable we will include in it, we need to follow a model-building strategy that allows us to choose only relevant variables to explain the $SF_{\text{satellite}}$, developing a robust and parsimonious model. We have already evaluated in Section 6.1 the basic properties of variables and their associations with each other, and now, we will evaluate how these variables behave within the model.

We follow an adapted version of the model-building strategy for logistic models proposed by Zhang (2016). This procedure originally contained five steps: univariate analysis, multivariable model comparisons, linearity assumption, interactions among explanatory variables and assessing the fit of the model. However, the interactions among explanatory variables were not applied during this work due to time constraints but are planned to be analysed in the future (see Section 9). Also, the final step of the Zhang model-building strategy, which assesses the fit of the model, will be carried out in the *Results* chapter (Chapter 7), since we do not have the intention of modifying the functional form or the set of explanatory variables of the model if, by chance, some of the four samples are not perfectly adjusted to the chosen model. Thus, with regard to this work, we see the assessing model's fit as an evaluative approach and not a procedure that assists in the construction of the model.

6.3.1 Univariate analysis

The first step uses univariable analysis to explore the unadjusted association between variables and the outcome. Each of the ten explanatory variables will be included in a logistic regression model, one at a time. Table 6.6 shows the results of the univariable regression for each explanatory variable. The p-values are relative to the *z-test*, (Casella and Berger, 2001) that examine whether the coefficient estimate $\widehat{\beta}$ is statistically significant, that is, whether $\widehat{\beta} \neq 0$. The z value statistics is the ratio of the estimated coefficient $\widehat{\beta}$ to its standard error $\widehat{\sigma}(\beta)$. It measures the number of standard deviations that the estimated coefficient is away from zero. In this step, a p-value smaller than 0.25 can be included for further multivariable analysis (Bendel and Afifi 1977; Mickey and Greenlan 1989). Just the variable $v_{\text{los}}/v_{\text{vir}}$ would not be included for further analysis (p-value $\approx$ 0.33). Despite this, variables of known relevance or that are of interest in the study should be included in the multivariate analysis, even if they are statistically insignificant.

6.3.2 Multivariable model comparisons

This step fits the multivariable model comprising all variables identified in step one. Variables that do not contribute to the model should be eliminated and a new, smaller model should be fitted. Here we will adopt a confidence level of 95% due to the ease of obtaining low p-values in tests with large sample sizes, as previously mentioned. In this case, only coefficients whose p-value is less than 0.05 will be understood as statistically significant coefficients. Table 6.7 shows the first trial multivariate model results. Here, the

TABLE 6.6: Summary of the univariate logistic models. The columns are, from left to right, the variable name, the coefficient estimate, the standard error, the z value, the confidence interval for the coefficient and the p-value associated with the z-test. The row highlighted in yellow indicates the variable that will be removed from the model.

Variable	$\widehat{\beta}$	$\widehat{\sigma}(\beta)$	z value	Confidence interval	P-value
$\log_{10} M_{\rm h}$	-0.172	0.014	-12.04	[-0.199 — -0.145]	<2e-16
$\log_{10} M_{\star}$	-0.587	0.061	-9.56	[-0.707 — -0.467]	<2e-16
$\log_{10} R_{\rm proj}/R_{\rm vir}$	0.470	0.019	25.38	[0.433 — 0.507]	<2e-16
$\log_{10} \sigma_{\rm satellite}$	-11.624	0.130	-89.61	[-11.879 — -11.369]	<2e-16
$\log_{10} \sigma_{\rm central}$	-0.534	0.063	-8.44	[-0.657 — -0.411]	<2e-16
$v_{\rm los}/v_{\rm vir}$	-0.025	0.026	-0.98	[-0.076 — 0.026]	0.32949
$AGN_{\rm satellite}$	3.375	0.040	84.56	[3.297 — 3.453]	<2e-16
$AGN_{\rm central}$	0.192	0.025	7.57	[0.143 — 0.241]	<2e-16
$SF_{\rm central}$	0.175	0.021	8.47	[0.134 — 0.216]	<2e-16

p-values are also associated with the z-test. We can see that the coefficient for the variable $AGN_{\rm central}$ is statistically insignificant (p-value $\approx$ 0.96). This way, the $AGN_{\rm central}$ variable will be eliminated from the model and a new one will be fitted.

Table 6.8 shows the second trial multivariate model results, without $AGN_{\rm central}$. We can see that all the coefficients are statistically significant at the 95% level.

Also, in the parsimonious model, the coefficients of variables should be compared to the coefficients in the original one. If a change of coefficients ($\Delta\beta$) is more than 20%, the removed variables have provided an important adjustment of the effect of the remaining variables. Such variables should be added back to the model. This process of deleting, adding variables to the model, fitting and refitting continues until all variables excluded are statistically unimportant, while variables that remain in the model are important. Table 6.9 shows the percentage of the change in the magnitude of the coefficients from multivariate model 1 (Table 6.7) to multivariate model 2 (Table 6.8). Since no percentage change is greater than 20%, we do not need to cycle the process again.

Until now the variables $\log_{10} M_{\rm h}$, $\log_{10} M_{\star}$, $\log_{10} R_{\rm proj}/R_{\rm vir}$, $\log_{10} \sigma_{\rm satellite}$, $\log_{10} \sigma_{\rm central}$, $AGN_{\rm satellite}$ and $SF_{\rm central}$ seem to be somehow relevant to the outcome $SF_{\rm satellite}$ and could be considered the preliminary main effects model. Next, we need to check the linearity assumption for the numerical variables of this set.

6.3.3 Linearity assumption

In this step, numerical variables are checked for their linearity in relation to the logit of the outcome. One way to find out is by examining the smoothed scatter plot by the logit of the outcome and a numeric explanatory variable, as in Figure 6.4. In addition to the points, the plot shows a *loess regression* (Cleveland, 1979), which is a nonparametric technique that uses locally weighted regression to fit a smooth curve through points in a scatter plot. When there is a linear relationship between the logit of the outcome and the explanatory variable, we expect to see a reasonably linear loess regression. The smoothed scatter plots show that variables $\log_{10} M_{\rm h}$, $\log_{10} M_{\star}$, $\log_{10} R_{\rm proj}/R_{\rm vir}$ and $\log_{10} \sigma_{\rm satellite}$ are all linearly

TABLE 6.7: Summary of the first trial multivariate logistic model. The columns are, from left to right, the variable name, the coefficient estimate, the standard error, the z-value, the confidence interval for the coefficient and the p-value associated with the z-test. The row highlighted in yellow indicates the variable that will be removed from the model.

Variable	$\widehat{\beta}$	$\widehat{\sigma}(\beta)$	z value	Confidence interval	P-value
Intercept	-13.516	0.961	-14.07	[-15.4 — -11.632]	<2e-16
$\log_{10} M_{\rm h}$	-0.140	0.023	-6.04	[-0.185 — -0.095]	<2e-16
$\log_{10} M_{\star}$	3.647	0.100	36.56	[3.451 — 3.843]	<2e-16
$\log_{10} R_{\rm proj}/R_{\rm vir}$	0.275	0.027	10.16	[0.222 — 0.328]	<2e-16
$\log_{10} \sigma_{\rm satellite}$	-11.724	0.159	-73.87	[-12.036 — -11.412]	<2e-16
$\log_{10} \sigma_{\rm central}$	0.264	0.121	2.18	[0.027 — 0.501]	0.02945
$AGN_{\rm satellite}$	2.879	0.043	67.31	[2.795 — 2.963]	<2e-16
$AGN_{\rm central}$	0.002	0.041	0.05	[-0.078 — 0.082]	0.95676
$SF_{\rm central}$	0.100	0.038	2.64	[0.026 — 0.174]	0.00822

TABLE 6.8: Summary of the second trial multivariate logistic model. The columns are, from left to right, the variable name, the coefficient estimate, the standard error, the z value, the confidence interval for the coefficient and the p-value associated with the z-test.

Variable	$\widehat{\beta}$	$\widehat{\sigma}(\beta)$	z value	Confidence interval	P-value
Intercept	-13.514	0.960	-14.08	[-15.396 — -11.632]	<2e-16
$\log_{10} M_{\rm h}$	-0.141	0.023	-6.05	[-0.186 — -0.096]	<2e-16
$\log_{10} M_{\star}$	3.647	0.100	36.56	[3.451 — 3.843]	<2e-16
$\log_{10} R_{\rm proj}/R_{\rm vir}$	0.275	0.027	10.17	[0.222 — 0.328]	<2e-16
$\log_{10} \sigma_{\rm satellite}$	-11.724	0.159	-73.88	[-12.036 — -11.412]	<2e-16
$\log_{10} \sigma_{\rm central}$	0.264	0.121	2.18	[0.027 — 0.501]	0.02923
$AGN_{\rm satellite}$	2.879	0.043	67.31	[2.795 — 2.963]	<2e-16
$SF_{\rm central}$	0.101	0.035	2.85	[0.032 — 0.17]	0.00436

TABLE 6.9: Percentage change in the coefficients ($\Delta\beta$ %) between the multivariate model 1 and the multivariate model 2.

Variable	$\Delta\beta$ %
Intercept	0.015 %
$\log_{10} M_{\rm h}$	-0.045 %
$\log_{10} M_{\star}$	0.001 %
$\log_{10} R_{\rm proj}/R_{\rm vir}$	0.008 %
$\log_{10} \sigma_{\rm satellite}$	-0.001 %
$\log_{10} \sigma_{\rm central}$	-0.205 %
$AGN_{\rm satellite}$	0.001 %
$SF_{\rm central}$	0.733 %

TABLE 6.10: Summary of the third trial multivariate logistic model. The columns are, from left to right, the variable name, the coefficient estimate, the standard error, the z value, the confidence interval for the coefficient and the p-value associated with the z-test. The row highlighted in yellow indicates the variable that will be removed from the model.

Variable	$\widehat{\beta}$	$\widehat{\sigma}(\beta)$	z value	Confidence interval	P-value
Intercept	-13.271	0.953	-13.92	[-15.139 — -11.403]	<2e-16
$\log_{10} M_{\rm h}$	-0.119	0.021	-5.66	[-0.16 — -0.078]	<2e-16
$\log_{10} M_{\star}$	3.654	0.100	36.66	[3.458 — 3.85]	<2e-16
$\log_{10} R_{\rm proj}/R_{\rm vir}$	0.272	0.027	10.06	[0.219 — 0.325]	<2e-16
$\log_{10} \sigma_{\rm satellite}$	-11.724	0.159	-73.88	[-12.036 — -11.412]	<2e-16
$AGN_{\rm satellite}$	2.878	0.043	67.31	[2.794 — 2.962]	<2e-16
$SF_{\rm central}$	0.060	0.030	1.99	[0.001 — 0.119]	0.04662

associated with the $SF_{\rm satellite}$ outcome in the logit scale. However, the variable $\log_{10} \sigma_{\rm central}$ is not related to the $SF_{\rm satellite}$ in the logit scale, as we can see by the quadratic shape of the loess regression.

If the scatter plot shows evidence of non-linearity, we shall apply other methods to build the model, such as 2 or 3-power terms, fractional polynomials and spline function (Royston and Altman, 1994). As previously mentioned, in this work, we will not test interactions between the variables nor the addition of polynomial terms. Given this and also the fact that if we are more critical about the significance of the $\log_{10} \sigma_{\rm central}$ coefficient (for example, a level of significance of 99%), its coefficient would not be statistically significant in the model (see Table 6.7). We reinforce the fact that we must apply more restrictive p-values when these come from tests carried out with very large samples. Finally, as shown in Section 6.1, $\log_{10} \sigma_{\rm central}$ is highly correlated with $\log_{10} M_{\rm h}$, which may be an explanation for the weakness in its coefficient. All these arguments give us the confidence to delete the variable $\log_{10} \sigma_{\rm central}$ from the model.

The summary of the third trial model, now without the variable $\log_{10} \sigma_{\rm central}$, is present in Table 6.10. When removing the variable $\log_{10} \sigma_{\rm central}$, we see that the coefficient of $SF_{\rm central}$ loses strength. Again, due to the lack of belief in its p-value, we have an argument for removing $SF_{\rm central}$ from the model. Before that, we evaluate in Table 6.11 the percentage change in the magnitude of the coefficients from multivariate model 2 (Table 6.8) to multivariate model 3 (Table 6.10). No percentage change is greater than 20%, so the removal of $\log_{10} \sigma_{\rm central}$ did not significantly modify the magnitude of the other coefficients.

Now, we can build a new model without the variable $\log_{10} \sigma_{\rm central}$. The summary of the fourth trial model is presented in Table 6.12. All coefficients are statistically significant. Next, we will evaluate if there was any significant change in the magnitude of the coefficients. Table 6.13 shows that there is no change in the magnitude of the coefficients greater than 20%. Thus, we chose this model as the final model to study the $f_{\rm SF}$ at large clustercentric distances.

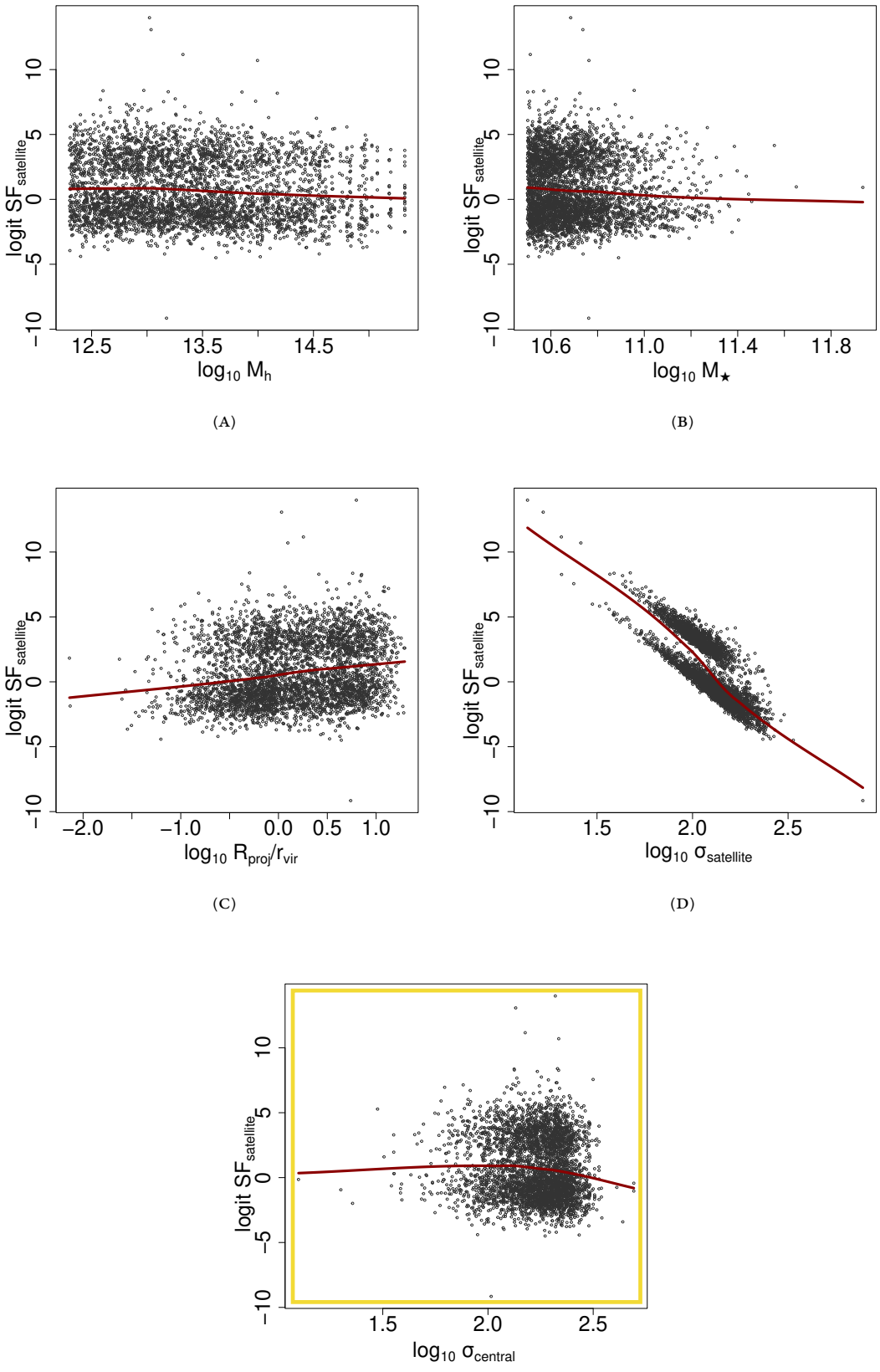

FIGURE 6.4: Smoother scatter plots of explanatory variables. The red lines are the loess regression. The panel highlighted in yellow indicates the variable that will be removed from the model.

TABLE 6.11: Percentage change in the coefficients ($\Delta\beta$ %) between the multivariate model 2 and the multivariate model 3.

Variable	$\Delta\beta$ %
Intercept	1.799 %
$\log_{10} M_{\mathrm{h}}$	15.175 %
$\log_{10} M_{\star}$	0.195 %
$\log_{10} R_{\mathrm{proj}}/R_{\mathrm{vir}}$	-1.166 %
$\log_{10} \sigma_{\mathrm{satellite}}$	0.001 %
$AGN_{\mathrm{satellite}}$	-0.011 %
SF_{central}	-1.117 %

TABLE 6.12: Summary of the fourth trial multivariate logistic model. The columns are, from left to right, the variable name, the coefficient estimate, the standard error, the z value, the confidence interval for the coefficient and the p-value associated with the z-test.

Variable	$\widehat{\beta}$	$\widehat{\sigma}(\beta)$	z value	Confidence interval	P-value
Intercept	-13.115	0.950	-13.81	[-14.977 — -11.253]	<2e-16
$\log_{10} M_{\mathrm{h}}$	-0.130	0.020	-6.38	[-0.169 — -0.091]	<2e-16
$\log_{10} M_{\star}$	3.655	0.100	36.67	[3.459 — 3.851]	<2e-16
$\log_{10} R_{\mathrm{proj}}/R_{\mathrm{vir}}$	0.276	0.027	10.27	[0.223 — 0.329]	<2e-16
$\log_{10} \sigma_{\mathrm{satellite}}$	-11.722	0.159	-73.88	[-12.034 — -11.41]	<2e-16
$AGN_{\mathrm{satellite}}$	2.880	0.043	67.34	[2.796 — 2.964]	<2e-16

TABLE 6.13: Percentage change in the coefficients ($\Delta\beta$ %) between the multivariate model 3 and the multivariate model 4.

Variable	$\Delta\beta$ %
Intercept	1.175 %
$\log_{10} M_{\mathrm{h}}$	-9.001 %
$\log_{10} M_{\star}$	0.023 %
$\log_{10} R_{\mathrm{proj}}/R_{\mathrm{vir}}$	1.702 %
$\log_{10} \sigma_{\mathrm{satellite}}$	0.018 %
$AGN_{\mathrm{satellite}}$	0.04 %

Chapter 7

Results

In this chapter, we present the results of our logistic regression and compare the best-fit models with observations. In Section 7.1, we describe the best-fit models obtained for the four datasets GSWLC01, GSWLC003, MPAJHU01, and MPAJHU003. We compare and discuss the predictions from these best-fit models with observations in Sections 7.2 and 7.3. Finally, to investigate the effects of the 2-halo term and pre-processing on the $f_{\rm SF}$, in Section 7.4, we explore how $f_{\rm SF}$ trends vary with the halo-mass threshold for assignment, $M_{\rm a}$.

7.1 Logistic model best-fits

In Chapter 6, using the GSWLC01 dataset, we built a parsimonious and robust logistic regression following a model-building strategy. The final model contains five explanatory variables: $\log_{10} M_{\rm h}$, $\log_{10} M_{\star}$, $\log_{10}(R_{\rm proj}/r_{\rm vir})$, $\log_{10} \sigma_{\rm e}$ [1] and AGN (note that, for simplicity, we will no longer mention whether the variable corresponds to satellites or central galaxies since all that were related to central galaxies are not part of the final model). As mentioned in Section 6.2, we used the GSWLC01 dataset to determine the best logistic model, but now we will *fit* this model using the other three datasets, GSWLC003, MPAJHU01 and MPAJHU003 since we are interested in seeing how the same set of variables under the same functional form will be adjusted to different samples. Table 7.1 shows the summaries of the four best-fit logistic models. We can see that in the four fits, almost all coefficients are statistically significant; only the coefficient for $\log_{10} M_{\rm h}$ is not significant for the sample with $z_{\rm max} = 0.03$.

Now, we want to determine the statistical importance of each explanatory variable in the four fits. In this context, the statistical importance is related to the effect of the variable in the model: more important variables have a greater effect on explaining the *SF* outcome. To evaluate this in GLM, the absolute value of the z value is frequently used (Quinlan, 1992). As already mentioned in Section 6.3, the z value measures the number of standard deviations $\widehat{\sigma}(\beta)$ that the estimated coefficient $\widehat{\beta}$ is away from zero. Figure 7.1 shows the importance variable for explanatory variables in the four fits. We can see that $\log_{10} \sigma_{\rm e}$ is the most important explanatory variable for three of the four samples, with the exception only for MPAJHU01 where AGN is the most important. Also, AGN is the second most

[1] Here we returned to the original notation for the stellar velocity dispersion, $\sigma_{\rm e}$, as there is no longer any confusion between the $\sigma_{\rm e}$ from satellites or central galaxies.

TABLE 7.1: Summary of the best-fit logistic model for each of the four datasets. The columns are, from left to right, the variable name, the coefficient estimate, the standard error, the z-value, the confidence interval for the coefficient, and the p-value associated with the z-test. The rows highlighted in yellow indicate a coefficient that is not statistically significant.

(A) Fit `GSWLC01`

Variable	$\widehat{\beta}$	$\widehat{\sigma}(\beta)$	z-value	Confidence interval	p-value
Intercept	-13.115	0.950	-13.81	[-14.977 — -11.253]	<2e-16
$\log_{10} M_{\rm h}$	-0.130	0.020	-6.38	[-0.169 — -0.091]	<2e-16
$\log_{10} M_{\star}$	3.655	0.100	36.67	[3.459 — 3.851]	<2e-16
$\log_{10} R_{\rm proj}/R_{\rm vir}$	0.276	0.027	10.27	[0.223 — 0.329]	<2e-16
$\log_{10} \sigma_{\rm e}$	-11.722	0.159	-73.88	[-12.034 — -11.41]	<2e-16
AGN	2.880	0.043	67.34	[2.796 — 2.964]	<2e-16

(B) Fit `GSWLC003`

Variable	$\widehat{\beta}$	$\widehat{\sigma}(\beta)$	z value	Confidence interval	P-value
Intercept	20.201	1.567	12.89	[17.13 — 23.272]	<2e-16
$\log_{10} M_{\rm h}$	-0.103	0.083	-1.25	[-0.266 — 0.06]	0.21175
$\log_{10} M_{\star}$	-1.144	0.132	-8.67	[-1.403 — -0.885]	<2e-16
$\log_{10} R_{\rm proj}/R_{\rm vir}$	1.030	0.085	12.17	[0.863 — 1.197]	<2e-16
$\log_{10} \sigma_{\rm e}$	-4.021	0.245	-16.43	[-4.501 — -3.541]	<2e-16
AGN	2.073	0.146	14.23	[1.787 — 2.359]	<2e-16

(C) Fit `MPAJHU01`

Variable	$\widehat{\beta}$	$\widehat{\sigma}(\beta)$	z value	Confidence interval	P-value
Intercept	12.844	1.241	10.35	[10.412 — 15.276]	<2e-16
$\log_{10} M_{\rm h}$	-0.120	0.024	-4.93	[-0.167 — -0.073]	<2e-16
$\log_{10} M_{\star}$	0.733	0.123	5.98	[0.492 — 0.974]	<2e-16
$\log_{10} R_{\rm proj}/R_{\rm vir}$	0.189	0.031	6.02	[0.128 — 0.25]	<2e-16
$\log_{10} \sigma_{\rm e}$	-10.078	0.182	-55.22	[-10.435 — -9.721]	<2e-16
AGN	2.512	0.034	73.58	[2.445 — 2.579]	<2e-16

TABLE 7.2: Fit `MPAJHU003`

Variable	$\widehat{\beta}$	$\widehat{\sigma}(\beta)$	z value	Confidence interval	P-value
Intercept	16.035	1.510	10.62	[13.075 — 18.995]	<2e-16
$\log_{10} M_{\rm h}$	-0.144	0.077	-1.87	[-0.295 — 0.007]	0.06157
$\log_{10} M_{\star}$	-0.827	0.137	-6.06	[-1.096 — -0.558]	<2e-16
$\log_{10} R_{\rm proj}/R_{\rm vir}$	1.081	0.080	13.46	[0.924 — 1.238]	<2e-16
$\log_{10} \sigma_{\rm e}$	-3.562	0.239	-14.91	[-4.03 — -3.094]	<2e-16
AGN	1.470	0.112	13.08	[1.25 — 1.69]	<2e-16

important variable for samples containing data from the GSWLC catalogue. Besides that, $\log_{10} M_{\rm h}$ is the least important variable in all four fits.

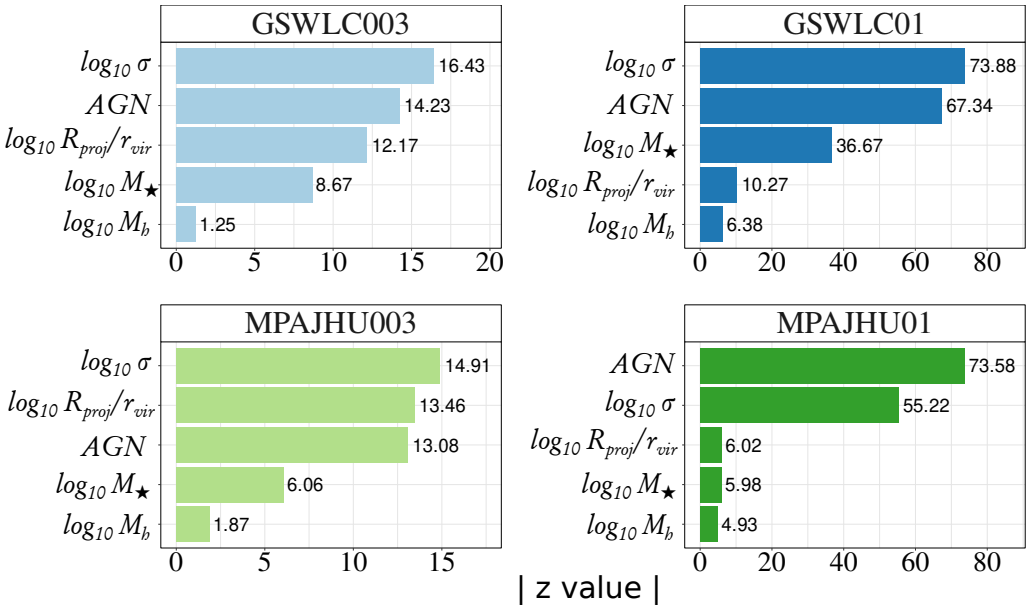

FIGURE 7.1: Importance of explanatory variables in the four fits. Each panel corresponds to one fit. The importance was measured using the absolute value of the z value.

Finally, we want to verify the effectiveness of our model in the four fits and check how well they are in classifying galaxies as star-forming or quiescent. First, we calculate an optimal cutoff for each fit based on the probability that gives minimum misclassification. Then, we classify the galaxies using the respective model fit for each sample and then compare the classification with the "true" response obtained through the classification into star-forming or quiescent made in Section 5.2. Table 7.3 shows the optimal cutoff and the confusion matrix for each fit. The confusion matrix is a table that displays the number of true negatives (TN), true positives (TP), false negatives (FN), and false positives (FP) produced by the fit of the data. In our context, these four categories mean:

- True negative (TN): The fit correctly indicates a quiescent galaxy as quiescent.
- True positive (TP): The fit correctly indicates a star-forming galaxy as star-forming.
- False negative (FN): The fit wrongly indicates a star-forming galaxy as quiescent.
- False positive (FP): The fit wrongly indicates a quiescent galaxy as star-forming.

We also calculate the true positive rate (TPR) and the true negative rate (TNR), where the TPR is the probability of the model identifying as star-forming a truly star-forming galaxy, and the TNR is the probability of the model identifying as quiescent a truly quiescent galaxy, so that

TABLE 7.3: Optimal cutoff and confusion matrix verification for each fit. From left to right: Sample dataset, optimal cutoff and percentage of true negatives, true positives, false negatives and false positives.

Fit	Optimal cutoff	TN	TP	FN	FP
GSWLC01	0.49	14497 (37%)	17810 (46%)	3936 (10%)	2543 (7%)
GSWLC003	0.49	557 (13%)	2915 (70%)	174 (4%)	528 (13%)
MPAJHU01	0.33	19359 (61%)	7583 (24%)	1593 (5%)	3162 (10%)
MPAJHU003	0.46	817 (22%)	2028 (55%)	235 (6%)	648 (17%)

TABLE 7.4: Model accuracy. From left to right: Sample dataset, fit accuracy, precision, as well as true positive and true negative rates.

Fit	ACC	Precision	TPR	TNR
GSWLC01	0.83	0.88	0.82	0.85
GSWLC003	0.83	0.85	0.94	0.51
MPAJHU01	0.85	0.71	0.83	0.86
MPAJHU003	0.76	0.76	0.90	0.56

$$\begin{aligned} \mathrm{TPR} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \\ \mathrm{TNR} &= \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}. \end{aligned} \tag{7.1}$$

Finally, we calculate the accuracy (ACC) of the fits, i.e., how often the fit prediction is correct, and the precision, i.e., out of all galaxies predicted as star-forming, the percentage that is truly star-forming, so that

$$\begin{aligned} \mathrm{ACC} &= \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}, \\ \mathrm{Precision} &= \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}. \end{aligned} \tag{7.2}$$

The results for the TPR, TNR, ACC and precision are shown in Table 7.4. We can see that all fits have great accuracy, with the fits of samples using the GSWLC catalogue being more precise. The fits for the $z_{\max} = 0.03$ samples have greater TPR, with the fit on GSWLC003 sample having a probability of 94% and on the MPAJHU003 sample of 90% to classify a truly star-forming galaxy as star-forming. However, the fits for the $z_{\max} = 0.1$ samples are those that have greater TNR, with the fit on GSWLC01 sample having a probability of 85% and on the MPAJHU01 sample of 86% to classify a truly quiescent galaxy as quiescent.

7.2 Analysis of the f_{SF} with GSWLC data

To analyse the variation of f_{SF} with galaxy properties and environment, we compare f_{SF} predicted by the logistic model with the fraction inferred from observation by binning

the galaxy data. In Figure 7.2, we show this comparison for the GSWLC01 satellite galaxies assigned to halos with masses greater than $\log_{10}(M_a/\mathrm{M}_\odot) = 12.3$. In this figure, the binned galaxy data is represented by solid symbols (satellites) and open symbols (central galaxies). The error bars associated with the data points are calculated using 95% binominal confidence intervals. The lines are the logistic regression for satellite galaxies, which is computed by fixing all continuous variables to their median values in each panel, and only the clustercentric distance is allowed to vary over the range from 0.1 to 20 $r_{\rm vir}$. The shaded area around the lines is the model's 95% confidence interval. Square symbols and dot-dashed lines represent non-AGN galaxies, while circles and solid lines represent AGN galaxies. Different panels correspond to bins of velocity dispersion (along columns) and halo mass (along rows), and colours (blue and red) indicate two different bins of stellar mass.

For the lowest bin of velocity dispersion ($1 \leq \log_{10}(\sigma_{\rm e}/[\mathrm{km\,s^{-1}}]) < 1.9$, the first column in Figure 7.2), we can see that $f_{\rm SF}$ is very high ($\gtrsim 0.9$) for all stellar masses, halo masses and clustercentric distances, regardless of whether the galaxy is AGN or non-AGN. In the second bin of velocity dispersion ($1.9 \leq \log_{10}(\sigma_{\rm e}/[\mathrm{km\,s^{-1}}]) < 2.1$, the middle column in Figure 7.2), almost all AGN galaxies are star-forming, with $f_{\rm SF} \gtrsim 0.95$ for all stellar masses, halo masses and clustercentric distances. On the other hand, for non-AGN galaxies, $f_{\rm SF}$ shows a stronger dependence on stellar mass (on average ≈ 0.11 of variation between the two bins of $M_\star$, and the increase in $f_{\rm SF}$ with $R/r_{\rm vir}$ is more pronounced compared to that in the low-$\sigma_{\rm e}$ bin (low-$\sigma_{\rm e}$ has $\approx 0.01-0.02$ of variation with $R/r_{\rm vir}$ and intermediate-$\sigma_{\rm e}$ has ≈ 0.07) and for AGNs in the intermediate-$\sigma_{\rm e}$ bin (≈ 0.01 of variation with $R/r_{\rm vir}$). In the high-$M_\star$ bin ($10.7 \leq \log_{10}(M_\star/\mathrm{M}_\odot) < 12.5$), $f_{\rm SF}$ varies from ≈ 0.7 at large $R/r_{\rm vir}$ to ≈ 0.5 in the inner parts of the halo. The $f_{\rm SF}$ values in the low-$M_\star$ bin ($10.5 \leq \log_{10}(M_\star/\mathrm{M}_\odot) < 10.7$) are smaller compared to those observed in the high-$M_\star$ bin, going from ≈ 0.6 at large distances to ≈ 0.4 at small $R/r_{\rm vir}$. There appears to be little dependence on halo masses, with a decrease of only $\approx 0.04-0.06$ in non-AGN $f_{\rm SF}$ with increasing $M_{\rm h}$, i.e., from the low- to the high-$M_{\rm h}$ bins; for AGNs, the variation in $f_{\rm SF}$ with $M_{\rm h}$ is less than ≈ 0.01.

For the high-$\sigma_{\rm e}$ bin (i.e. $2.1 < \log_{10}(\sigma_{\rm e}/[\mathrm{km\,s^{-1}}]) < 3$), last column in Figure 7.2), we observe the lowest values of $f_{\rm SF}$ among all $\sigma_{\rm e}$ bins. Contrary to what is observed in the other $\sigma_{\rm e}$ bins, $f_{\rm SF}$ for AGN galaxies depends on stellar masses and clustercentric distances, varying from $f_{\rm SF} \approx 0.75$ (≈ 0.88) at small $R/r_{\rm vir}$ to $f_{\rm SF} \approx 0.88$ (≈ 0.93) at large distances in the low-$M_\star$ (high-$M_\star$) bin. Variations of AGN $f_{\rm SF}$ with halo mass are small ($\lesssim 0.04$ and $\lesssim 0.01$ for galaxies in the low- and high-$M_\star$ bins, respectively). For non-AGN galaxies in the low-$M_\star$ (high-$M_\star$) bin, we see a increase from $f_{\rm SF} \approx 0.09$ (≈ 0.14) in the inner regions to $f_{\rm SF} \approx 0.20$ (≈ 0.26) at large $R/r_{\rm vir}$. Again, the halo mass appears to have little effect on $f_{\rm SF}$, with a decrease smaller than ≈ 0.04 from the low- to the high-$M_{\rm h}$ bin.

The $f_{\rm SF}$ for central galaxies tend to be higher compared to that of satellites within $\sim 1\,r_{\rm vir}$, with some exceptions. In some panels, we have a small sample of centrals, making it difficult to analyse the overall trends concerning all variables.

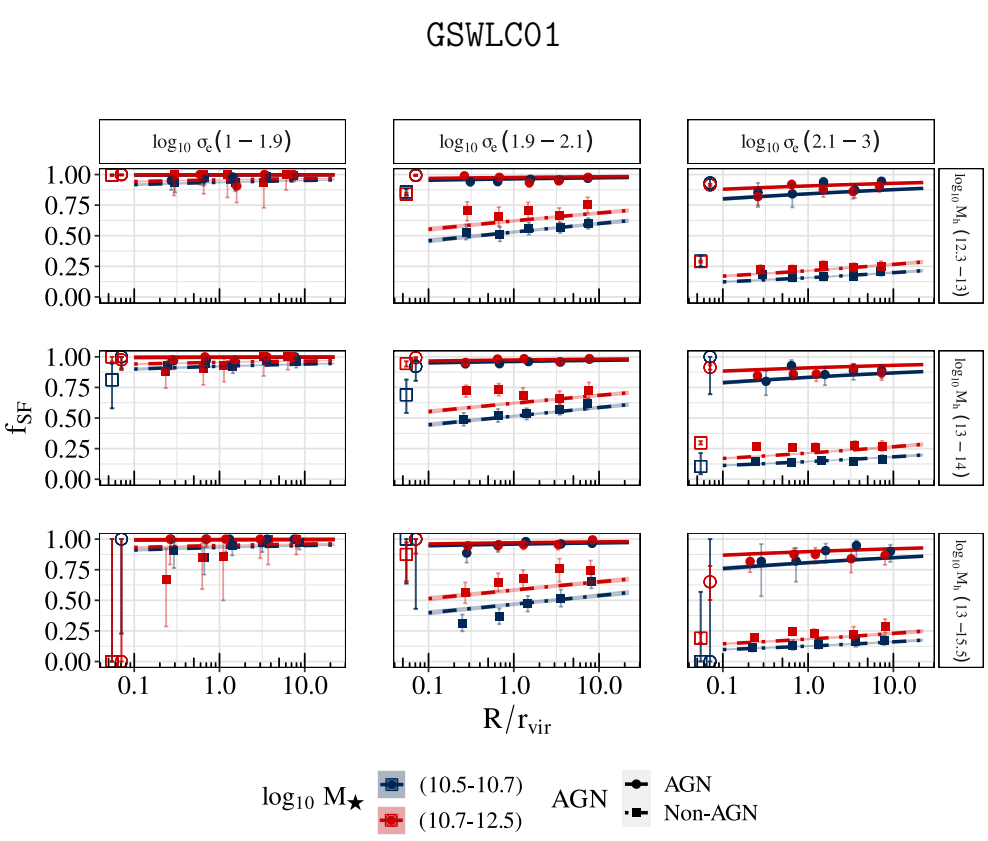

FIGURE 7.2: Values of f_{SF} predicted by the logistic model compared to binned data of galaxies in the GSWLC01 sample and using $\log_{10}(M_a/M_\odot) = 12.3$ for the membership assignment. Binned data of satellite galaxies are represented by solid symbols, and central galaxies are indicated as open symbols. The error bars associated with symbols are calculated using 95% binominal confidence intervals. The lines are the logistic model for satellite galaxies, and the shaded area around the lines is the model's 95% confidence interval. Squares and dot-dashed lines represent non-AGN galaxies, while circles and solid lines represent AGN galaxies. Different panels correspond to bins of velocity dispersion (along columns) and halo mass (along rows), and colours (blue and red) indicate two different bins of stellar mass.

To analyse f_{SF} for low mass galaxies ($\log_{10}[M_\star/M_\odot] < 10.5$), in Figure 7.3, we show f_{SF} predicted by the logistic model and inferred from observations for galaxies in sample GSWLC003 and assigned to halos more massive than $\log_{10}(M_a/M_\odot) = 12.3$. Due to the small sample size, we present the data and model for only two bins of velocity dispersion, $\log_{10}(\sigma_e/[\mathrm{km\,s^{-1}}]) < 2.1$ (low-σ_e bin) and > 2.1 (high-σ_e bin). For the other continuous variables, we consider their median values for the whole sample to compute f_{SF} as a function of R/r_{vir} from the best-fit model.

We can see that the trends of f_{SF} for low-mass galaxies are somewhat different from those observed for galaxies with $\log_{10}[M_\star/M_\odot] > 10.5$ in the GSWLC01 sample (Figure 7.2). The dependence of f_{SF} with clustercentric distance is stronger compared to that of high-mass galaxies. In the low-σ_e bin, f_{SF} for AGN (non-AGN) galaxies increases from $f_{SF} \approx 0.72$ (≈ 0.54) to $f_{SF} \approx 0.95$ (≈ 0.90) from the inner to the outer halo regions.

A significant variation of f_{SF} with R/r_{vir} is also observed in the high-σ_e bin, going from $f_{SF} \approx 0.26$ (≈ 0.04) to $f_{SF} \approx 0.73$ (≈ 0.26) for AGN (non-AGN) galaxies.

The central galaxies in the low-σ_e bin are more likely to be star-forming compared to satellites in the inner parts of the haloes. On the other hand, all high-σ_e central galaxies are quiescent; however, the number of AGN galaxies in the high-σ_e is small, making f_{SF} very uncertain.

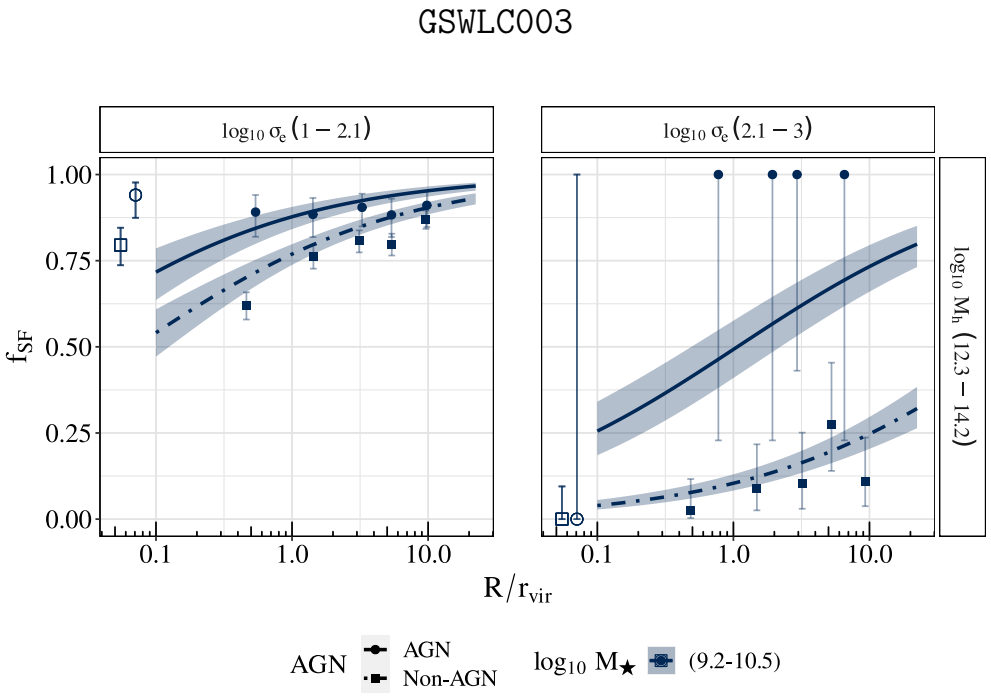

FIGURE 7.3: Values of f_{SF} predicted by the logistic model compared to binned data of galaxies in the GSWLC003 sample and using $\log_{10}(M_a/M_\odot) = 12.3$ for the membership assignment. Binned data of satellite galaxies are represented by solid symbols, and central galaxies are represented by open symbols. The error bars associated with symbols are calculated using 95% binominal confidence intervals. The lines are the logistic model for satellite galaxies, and the shaded area around the lines is the model's 95% confidence interval. Squares and dot-dashed lines represent non-AGN galaxies, while circles and solid lines represent AGN galaxies. The *left* and *right panels* correspond to the low- and high-velocity dispersion bins, respectively.

7.3 Analysis of the f_{SF} with MPA-JHU data

As discussed in Section 3.2, there are different ways to estimate SFRs, with each method having its own biases and limitations that can possibly lead to different values. Since our classification between star-forming or quiescent is based on this quantity, changes in SFRs estimates can directly affect our results and conclusions. Therefore, in Figures 7.4 and 7.5, we present the results obtained when using SFR and $M_\star$ estimates from MPA-JHU catalogue for our sample with $z_{max} = 0.1$ and $z_{max} = 0.03$ – i.e., the MPAJHU01 and MPAJHU003 datasets, respectively.

Comparing the results obtained for the GSWLC01 and MPAJHU01 samples (Figures 7.2 and 7.4), we can see that several trends are quantitatively similar. For AGN galaxies, f_{SF} is higher than that for non-AGNs in all bins of velocity dispersion, stellar mass, halo mass, and at all clustercentric distances. Besides, f_{SF} decreases with σ_e and increases with R/r_{vir}, and, as we observe in the results for the GSWLC01 sample, the dependence of f_{SF} on halo mass is very weak.

On the other hand, in the previous results, f_{SF} varies with stellar mass for non-AGN galaxies in the intermediate- and high-σ_e bins, while for MPAJHU01, this trend appears in the low-σ_e bin and is weaker (variation on average of ≈ 0.05, while in the GSWLC01 sample is on average ≈ 0.11). In addition, if we quantitatively compare the results obtained with each sample, we find that they differ significantly. Around 95% of the galaxies in the GSWLC01 low-σ_e bin are star-forming, but for MPAJHU01 we observe high f_{SF} values only for AGN galaxies ($f_{SF} \gtrsim 0.93$); for non-AGNs, the fraction varies between ≈ 0.50 and ≈ 0.75. The f_{SF} for non-AGN galaxies also present a stronger correlation with R/r_{vir} compared to the results with GSWLC01, with an increase of up to 10% from the inner to the outer parts of the haloes. The dependence on halo mass is also stronger (up to 10%).

For galaxies in the intermediate-σ_e bin, f_{SF} for AGNs varies from $f_{SF} \approx 0.68 - 0.75$ in the inner parts of the halo to $f_{SF} \approx 0.76 - 0.82$ in the outer parts, with larger fractions observed in the low-M_h bin. These fractions are much lower than those observed in the GSWLC01 results ($f_{SF} > 0.95$). The fraction of non-AGN star-forming galaxies is also smaller compared to the GSWLC01, varying between $f_{SF} \approx 0.13$ and 0.22 (for GSWLC01, $f_{SF} > 0.4$ in the intermediate-σ_e bin). On the other hand, in this σ_e bin, the dependency with M_h is similar for both GSWLC01 and MPAJHU01 (variation of ≈ 0.04).

Finally, in the high-σ_e bin, we observe that almost all non-AGN galaxies are quiescent, with $f_{SF} < 0.06$ for all bins of σ_e, $M_\star$, and M_h (for GSWLC01, f_{SF} for non-AGN galaxies can be as high as 0.26 in the halo outskirts within this σ_e bin). The fractions for AGN galaxies are also lower (f_{SF} between 0.38 and 0.55) than those obtained for GSWLC01 ($f_{SF} > 0.80$).

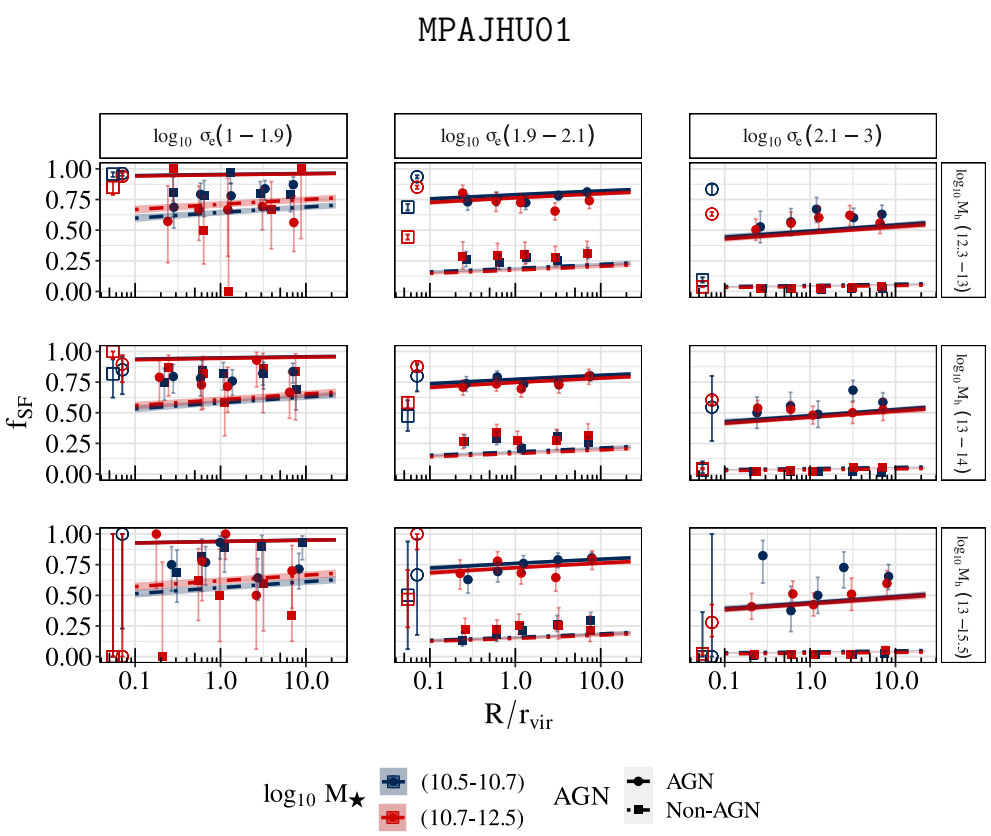

FIGURE 7.4: Values of f_{SF} predicted by the logistic model compared to binned data of galaxies in the MPAJHU01 sample and using $\log_{10}(M_a/M_\odot) = 12.3$ for the membership assignment. The notation is the same as in Figure 7.2.

Similarly, to analyse f_{SF} for low mass galaxies ($\log_{10}[M_\star/M_\odot] < 10.5$), we show in Figure 7.5 the f_{SF} predicted by the logistic model and inferred from observations for galaxies in sample MPAJHU003, for galaxies assigned to halos more massive than $\log_{10}(M_a/M_\odot) = 12.3$. $\log_{10}(\sigma_e/[\mathrm{km\,s^{-1}}]) < 2.1$ (low-σ_e bin) and > 2.1 (high-σ_e bin). For the other continuous variables, we consider their median values for the whole sample to compute f_{SF} as a function of R/r_{vir} from the best-fit model.

Comparing with the GSWLC003 fit (Figure 7.3), we can see that the trends of f_{SF} for low-mass galaxies are, just as they were for the GSWLC003, somewhat different from those observed for galaxies with $\log_{10}[M_\star/M_\odot] > 10.5$, especially with regard to dependence on R/r_{vir}, being much stronger compared to that of high-mass galaxies. Comparing the low-σ_e with that of the GSWLC003 sample, we can see that both AGN and non-AGN galaxies have, on average, smaller f_{SF}, with this difference decreasing ≈ 0.26 (≈ 0.18) for AGN (non-AGN) galaxies from one sample to the other in the centre of halos. At the outskirts of the halo, this difference between the two samples is smaller, decreasing by about ≈ 0.07 for both AGN and non-AGN from sample GSWLC003 to MPAJHU003. In the high-σ_e bin, we see that non-AGN galaxies also have lower f_{SF} than GSWLC003 (≈ 0.10 less thanGSWLC003 on the halo outer regions), but for AGN galaxies this difference is even higher, the difference is greater, reaching a decrease of ≈ 0.31 on the halo outer regions. Furthermore, the dependence on stellar mass is, on average, smaller when

compared to the GSWLC003 sample (about ≈ 0.04 less for the low-σ_e bin and ≈ 0.2 less for the high-σ_e). The central galaxies present the same trends as those in Figure 7.3.

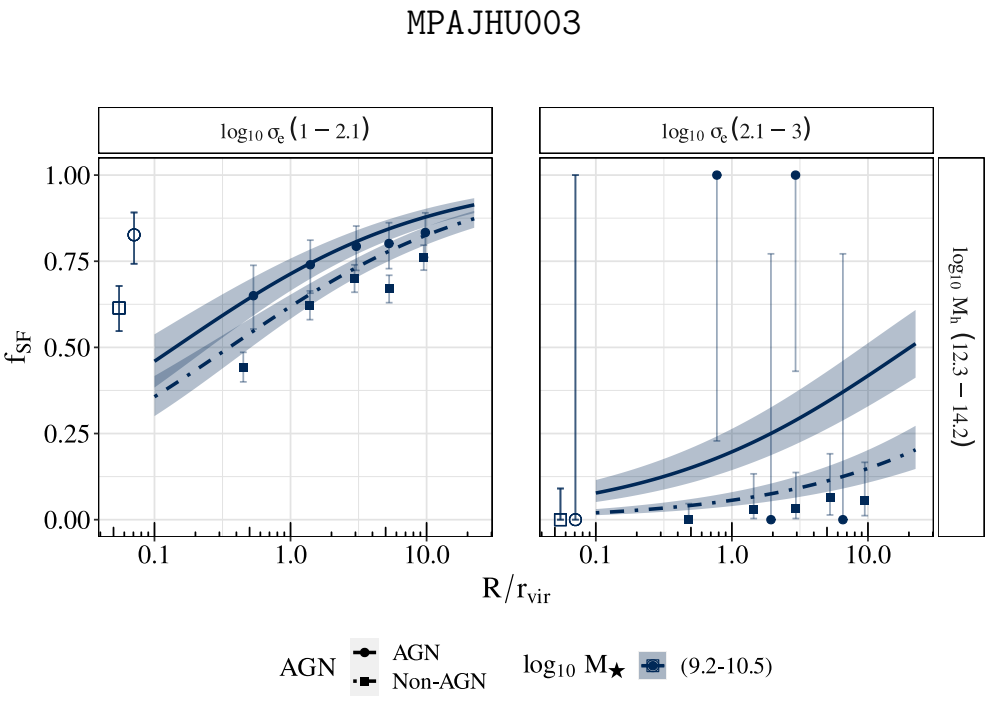

FIGURE 7.5: Values of f_{SF} predicted by the logistic model compared to binned data of galaxies in the MPAJHU003 sample and using $\log_{10}(M_a/M_\odot) = 12.3$ for the membership assignment. The notation is the same as in Figure 7.3.

7.4 Variations of f_{SF} trends with halo-mass threshold for assignment (M_a)

As explained in Chapter 4, with the assignment scheme adopted in this work, we are able to investigate how small haloes around larger ones affect the variation of f_{SF} with clustercentric distance (i.e., pre-processing) by changing the halo-mass threshold for assignment, M_a. In Figures 7.6 and 7.7, we show the logistic model and binned data for the f_{SF} galaxies in the samples GSWLC01 and MPAJHU01, respectively, but now we adopt $\log_{10}(M_a/M_\odot) = 13.0$. We chose not to show the figures for samples with $z_{max} = 0.03$ for galaxies assigned to halos with $\log_{10}(M_a/M_\odot) = 13.0$ because of their small sample size. Also, for comparison, we show the models for the f_{SF} galaxies assigned to halos with $\log_{10}(M_a/M_\odot) = 12.3$, shown in Figures 7.2 (sample GSWLC01) and 7.4 (sample MPAJHU01). In this case, the lighter blue lines indicate the lowest stellar mass bin and the lighter red lines indicate the highest stellar mass bin. Solid lines represent AGN galaxies, and dot-dash lines represent non-AGN galaxies.

For the GSWLC01 sample, in Figure 7.6, there is a very small difference in f_{SF} between the different values adopted for M_a, albeit, in general, when $\log_{10}(M_a/M_\odot) = 13.0$, f_{SF} is slightly higher. The only exception is for non-AGN galaxies in the intermediate σ_e bin,

where we can see that f_{SF} for $\log_{10}(M_a/M_\odot) = 13.0$ is strongly deviant, being much higher. In the least massive bins (blue lines), we see an increase in the f_{SF} of, on average, ≈ 0.07 from $\log_{10}(M_a/M_\odot) = 12.3$ to $\log_{10}(M_a/M_\odot) = 13.0$. In the most massive bins, this increase is, on average, $\approx$ 0.11. The trends of f_{SF} with R_{proj}/r_{vir} are similar between the two M_a choices.

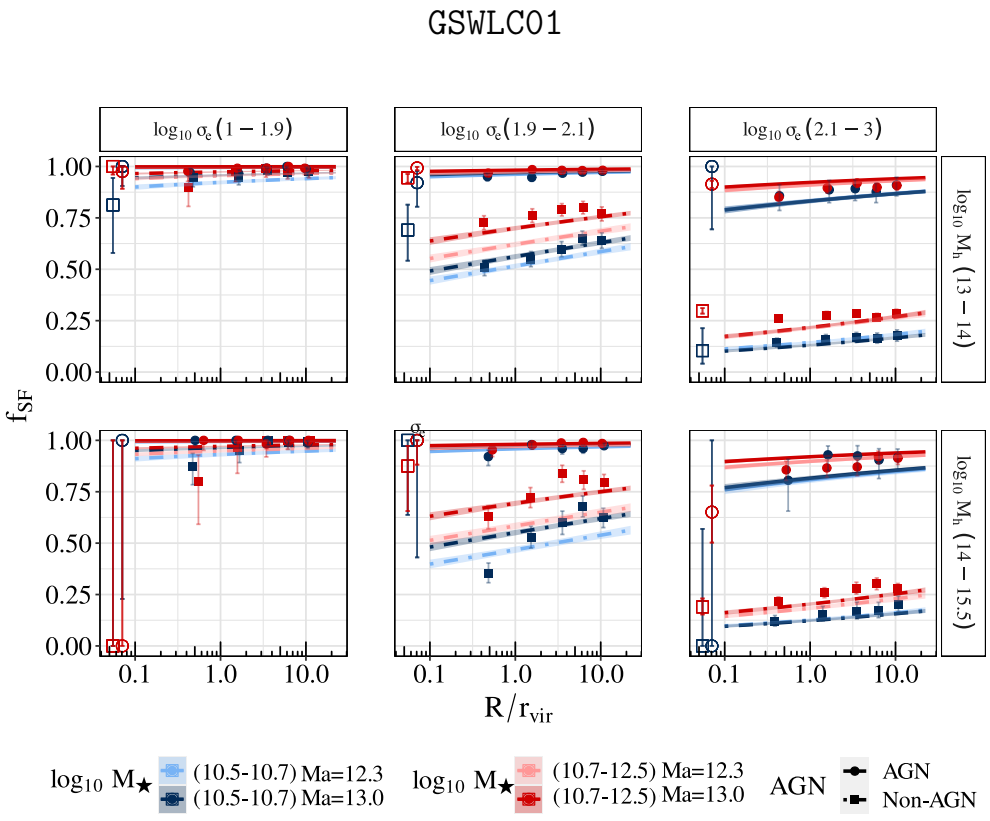

FIGURE 7.6: Values of f_{SF} predicted by the logistic model compared to binned data of galaxies in the GSWLC01 sample and using $\log_{10}(M_a/M_\odot) = 13$ for the membership assignment. The notation is the same as in Figure 7.2. For comparison, the predictions by the logistic model fitted to the sample using $\log_{10}(M_a/M_\odot) = 12.3$ (Figure 7.2) is shown as light-red and light-blue curves in all panels.

In Figure 7.7, we show the same analysis as in the previous figure, but this time for the MPAJHU01 sample. The observed trend that when $\log_{10}(M_a/M_\odot) = 13.0$ f_{SF} is generally higher is preserved, but this time this difference increases with R_{proj}/r_{vir}.

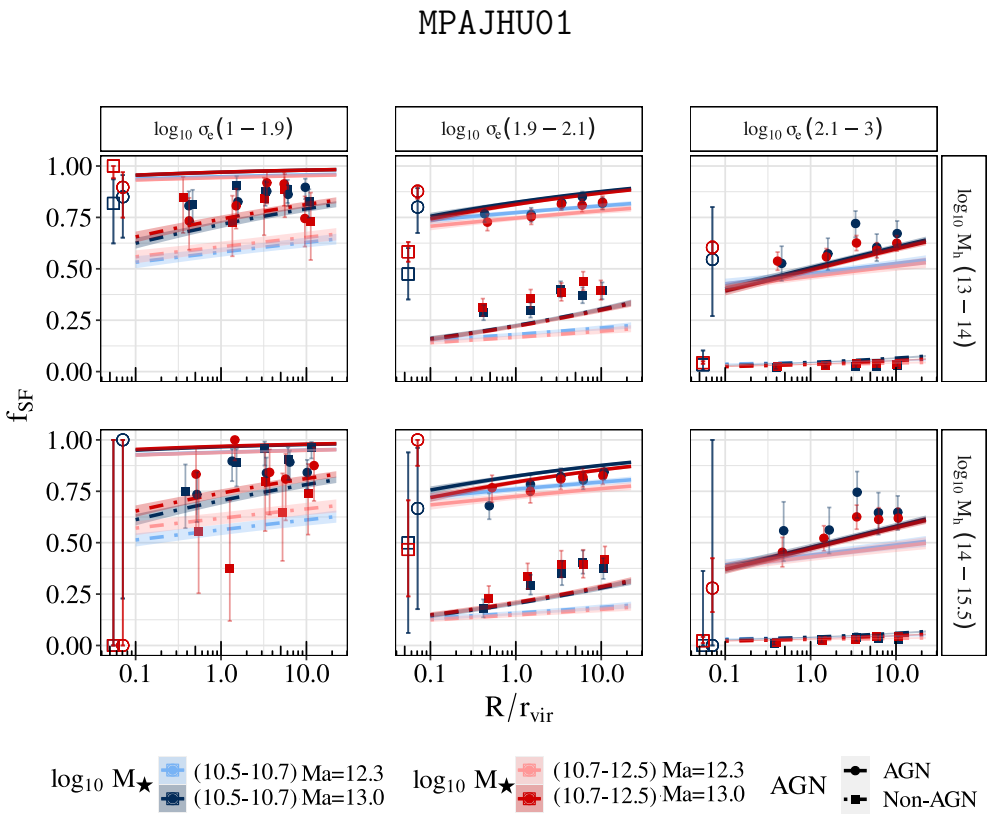

FIGURE 7.7: Values of f_{SF} predicted by the logistic model compared to binned data of galaxies in the MPAJHU01 sample and using $\log_{10}(M_a/M_\odot) = 13$ for the membership assignment. The notation is the same as in Figure 7.2. For comparison, the predictions by the logistic model fitted to the sample using $\log_{10}(M_a/M_\odot) = 12.3$ (Figure 7.4) is shown as light-red and light-blue curves in all panels.

In the context of *pre-processing*, one would expect the opposite scenario that is observed in Figures 7.6 and 7.7, i.e., a decrease in f_{SF} galaxies as we assign to larger halos galaxies from smaller halos (for instance, by increasing $\log_{10}(M_a/M_\odot)$ from 12.3 to 13.0). What we see is that larger halos are being populated with star-forming galaxies as f_{SF} increases as we go to a larger value of $\log_{10}(M_a/M_\odot)$. Motivated by this result, in the following, we reanalyse the effect of changing M_a but this time removing galaxies defined as centrals by the Lim *et al.*, 2017 galaxy catalogue.

In Figures 7.8 and 7.9, we show the logistic model and binned data for the f_{SF} galaxies in the samples GSWLC01 and MPAJHU01, respectively, adopting $\log_{10}(M_a/M_\odot) = 13.0$ and removing galaxies defined as centrals in Lim *et al.*, 2017. Once more, we do not show the figures for samples with $z_{max} = 0.03$ because of their small sample size. For both the GSWLC01 and MPAJHU01 samples, the effect of removing central galaxies of smaller halos being assigned to larger halos is a drastic decrease in the f_{SF} differences with the increase of M_a. Furthermore, regarding the GSWLC01 sample, by comparing Figure 7.8 with Figure 7.6, we immediately notice that the difference observed in the intermediary σ_e bin between the M_a choices disappears. Therefore, this suggests that galaxies that are likely to be central in small haloes are the most responsible for the increase in star formation at the halo's large

radii observed when increasing the M_a. Moreover, satellite galaxies in smaller halos do not produce a significant change in the overall f_{SF} trends when incorporated into larger halos.

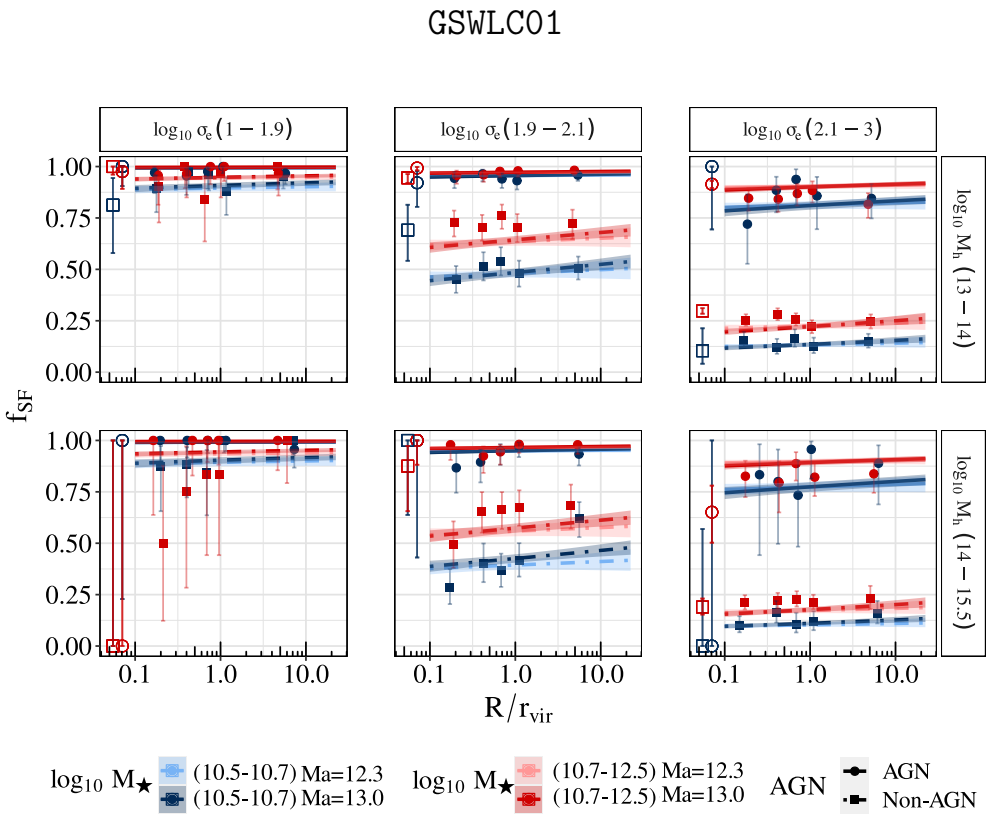

FIGURE 7.8: Values of f_{SF} predicted by the logistic model compared to binned data of galaxies in the GSWLC01 sample and using $\log_{10}(M_a/M_\odot) = 13$ for the membership assignment. The notation is the same as in Figure 7.6.

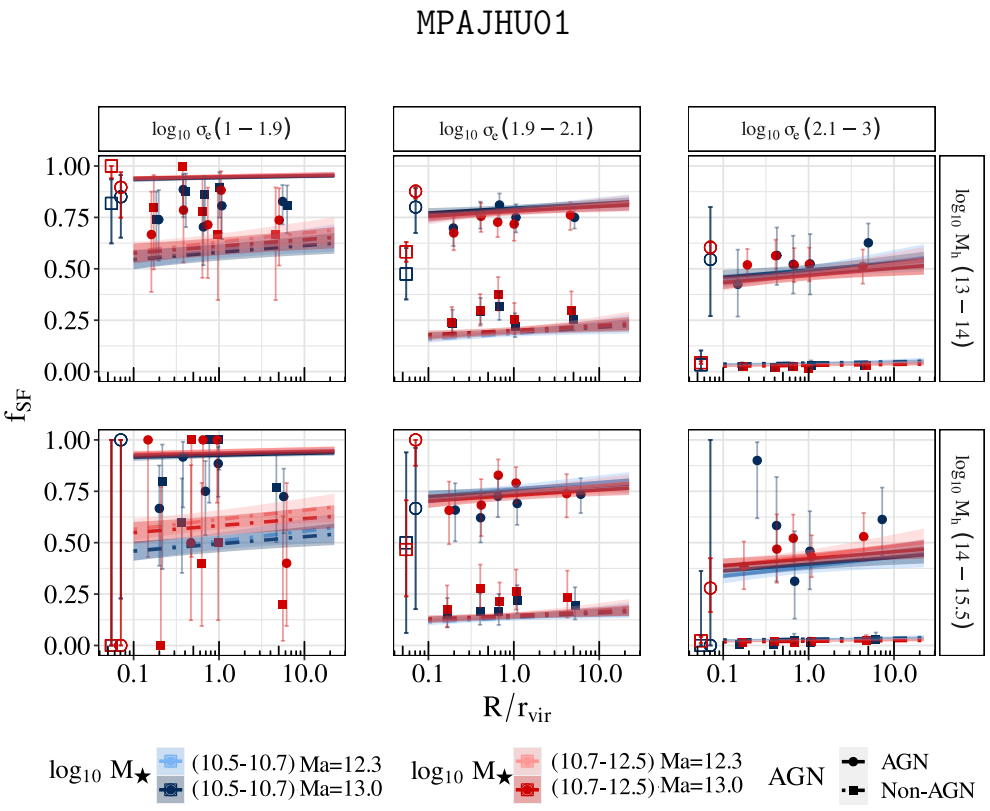

FIGURE 7.9: Values of f_{SF} predicted by the logistic model compared to binned data of galaxies in the `MPAJHU01` sample and using $\log_{10}(M_a/M_\odot) = 13$ for the membership assignment. The notation is the same as in Figure 7.8.

Chapter 8

Discussion

The models and results presented in Chapter 7 illustrate the complexity in the relations between $f_{\rm SF}$, galaxy properties and their environment. In this chapter, we discuss the implications of our findings.

8.1 The dependency with the clustercentric distance

A first glance at the Figures from 7.2 to 7.5 shows that the $f_{\rm SF}$ increases with clustercentric distance. This is in agreement with very well-known observations and results from several studies (e.g. von der Linden *et al.*, 2010; Haines *et al.*, 2015; Mahajan *et al.*, 2011). However, we see that such an increase is not homogeneous between different samples. For instance, one clear difference emerges from the analysis of variable importance shown in Figure 7.1, where we see that the variable $\log_{10}(R_{\rm proj}/r_{\rm vir})$ is more important in samples with $z_{\rm max}$ = 0.03 in comparison with the samples with $z_{\rm max}$ = 0.1. For a fixed catalogue (GSWLC or MPA-JHU), we see that the rank of $\log_{10}(R_{\rm proj}/r_{\rm vir})$ drops one position from $z_{\rm max}$ = 0.03 to $z_{\rm max}$ = 0.1. This can be due to the fact that the sample within $z_{\rm max}$ = 0.03 contains only galaxies with $\log_{10}(M_\star/{\rm M}_\odot) < 10.5$, and low-mass galaxies are more prone to be affected by the environment (Bluck *et al.*, 2020b and references therein).

Our analysis shows that trends of $f_{\rm SF}$ with $R/r_{\rm vir}$ are different depending on the SFR estimates that we use. The importance of variables shows that the $R/r_{\rm vir}$ is more important in the fits for MPA-JHU samples than those from GSWLC, for a fixed $z_{\rm max}$ (Table 7.1 and Figure 7.1). We discuss the differences between the results for MPA-JHU and GSWLC later on.

Finally, it is clear from our results that the relation between $f_{\rm SF}$ and clustercentric distance becomes weaker when the analysis includes other galaxy properties. For example, some studies show variations of more than 30% in $f_{\rm SF}$ going from the inner to the outer regions of the halo (e.g. Mahajan *et al.*, 2011; Haines *et al.*, 2015, see Figure 1.12). On the other hand, variations in our best-fit models are less than 20% if we consider galaxies with similar properties (i.e., in bins of stellar mass, velocity dispersion and AGN activity). This suggests that the absolute variations of $f_{\rm SF}$ with the environment are at least partially driven by the dependence of galaxy properties with the environment; i.e., the population of galaxies in different environments can be biased towards objects with specific properties.

8.2 The dependency with galaxy velocity dispersion

We find that the galaxy velocity dispersion is the most important property in our best-fit models (except for the MPAJHU01 sample, see Table 7.1 and Figure 7.1), in agreement with other studies (e.g. Brownson *et al.*, 2022; Bluck *et al.*, 2020b). In particular, for the low-σ_e bin in the GSWLC01 sample (Figure 7.2), we can see that almost all galaxies ($\gtrsim$ 95%) are star-forming according to our best-fit model, regardless of $M_\star$, M_h, and R/r_{vir}. On the other hand, for the other σ_e bins and for the MPA-JHU samples, the values of f_{SF} and their relation with R/r_{vir} depend mainly on whether the galaxy is classified as AGN or non-AGN.

In any case, we observe a clear decrease in f_{SF} with increasing σ_e for all samples. For GSWLC01 sample, only $\approx$ 20% of high-σ_e non-AGN galaxies are star-forming, with this fraction being even lower for the MPAJHU01 sample ($\lesssim$ 6%). The increase of the number of quiescent galaxies with velocity dispersion is usually interpreted as a probable signature of cumulative AGN feedback, particularly in the maintenance (radio) mode, being responsive for the quenching of high-mass galaxies (Bluck *et al.*, 2020b). This interpretation arises from the well-known relation between the velocity dispersion (or bulge mass) and the mass of the black hole (BH) (Ferrarese and Merritt, 2000; Gebhardt *et al.*, 2000) which indicates that high-σ_e galaxies host very massive BHs. Since the BH mass is a proxy for the amount of AGN activity that the galaxy had integrated over time, these high-σ_e are very likely to be affected by this process. Interestingly, even in the high-σ_e bin, we still see dependence of f_{SF} with R/r_{vir}, indicating that the AGN feedback is not the only mechanism behind the quenching of high-σ_e galaxies.

8.3 The dependency with stellar mass

In this work, we intended to investigate galaxies in a large mass range. However, due to sample completeness, we present the data in two redshift limits, which consequently have different mass coverage. For $z_{max} = 0.1$, the logarithmic stellar mass ranges from 10.5 to 12.5. For $z_{max} = 0.03$, the logarithmic stellar mass varies from 9.17 and 10.5. The $z_{max} = 0.1$ selection is by far the largest galaxy sample, with more than 32 thousand objects, while the low-z sample contains only $\approx$ 4000 galaxies. However, despite a smaller sample size, the $z_{max} = 0.03$ sample comprises low-mass galaxies in which star formation is more likely to be affected by the environment (Bluck *et al.*, 2020b and references therein). Furthermore, we divide the $z_{max} = 0.1$ samples into two mass bins. The more massive bin with galaxy masses between $10.7 \leq \log_{10}(M_\star/M_\odot) < 12.5$ and the least massive bin with masses between $10.5 \leq \log_{10}(M_\star/M_\odot) < 10.7$.

In the current paradigm, star formation is related to galaxy stellar mass, such as more massive galaxies tend to be more quiescent. However, the scatter observed in the trends of star formation with stellar mass indicates that the relationship may be more complex than usually assumed (Guo *et al.*, 2013; Schreiber *et al.*, 2015; Santini *et al.*, 2017; Matthee and Schaye, 2019; Berti *et al.*, 2021; Huang *et al.*, 2023). Results of our analysis show that, for the MPAJHU01 sample, there is no significant variation of the f_{SF} with mass (Figure 7.4), suggesting that σ_e is, in fact, the property that correlates with f_{SF}. Therefore, the variations of f_{SF} with $M_\star$ probably arise from the relation of $M_\star$ with σ_e, and not the stellar mass itself.

On the other hand, for the GSWLC01 sample, f_{SF} differences between the $M_\star$ bins appear for non-AGN galaxies in intermediate- and high-σ_e bins, while for AGN galaxies, these differences appear only for the high-σ_e bin (see Figure 7.2). Surprisingly, we observe that more massive galaxies present a slightly higher f_{SF} when compared to their least massive counterparts. In any case, the trends with stellar mass, when compared with trends based on other properties, such as σ_e, are considerably subtle. This can also be observed in Figure 7.1, where we show that the mass is the second-to-last important variable within the five properties studied in this work in 3 out of the 4 samples. Moreover, in the analysis of the sample within $z_{max} = 0.03$, despite the fact of not having enough sample size to justify a separation in $M_\star$ bins, the results in Figure 7.1 indicate that variations of f_{SF} with stellar mass are not important.

8.4 The dependency with *AGN* activity

The AGN activity, classified using the WHAN diagram, is among the most important variables in our best-fit models. For galaxies classified as AGN in our study, the f_{SF} is higher than that of non-AGN galaxies in all bins of σ_e, $M_\star$, M_h and R/r_{vir}. Such high fractions of AGNs hosted by star-forming galaxies can be related to the amount of gas in the inner parts of the galaxy, which would both feed the supermassive BH and lead to an increase in the SFR. Indeed, many studies show that the hosts of strong AGNs tend to be star-forming galaxies (Rosario *et al.*, 2013a; Rosario *et al.*, 2013b). However, even in this scenario, it is not expected to find such high f_{SF} for AGN galaxies since it is also well-known that many quiescent galaxies host an AGN (e.g. Silk and Rees, 1998; Cattaneo *et al.*, 2009; Fabian, 2012; Cicone *et al.*, 2014).

In addition, it is unlikely that such a strong correlation with current AGN activity, as indicated by our model, is a signature of the AGN feedback on the galaxy star-formation activity, given the timescales involved in these processes. The lifetime of AGN activity, which is about 10^5 to 10^8 years (e.g. Marconi *et al.*, 2004; Schawinski *et al.*, 2015; Yuan *et al.*, 2018), is much shorter than the quenching timescale (typically about 1 Gyr, Bell *et al.*, 2004; Blanton, 2006) so that it is not expected to find an instantaneous correlation between AGN and star-formation activity directly from the observational data. In fact, as already discussed in Section 8.2, there are strong indications that galaxy quenching is more likely to be related to the cumulative effect of AGN feedback integrated over time, especially for massive galaxies.

Identifying the source of ionisation in galaxies using diagnostic diagrams can be tricky. As argued by Fernandes *et al.* (2010), some widely used diagrams, such as the BPT diagram, can leave galaxies with weak emission out of statistical studies, and for this reason, we opted to use the WHAN diagram. However, many sources that are classified as AGN in the BPT diagram have EW(Hα) < 3 Å; i.e., using the WHAN diagram, they would be classified as retired or passive galaxies, illustrating the complexity of identifying a galaxy as an AGN host. Therefore, our AGN sample might be biased towards strong AGNs, which tend to be star-forming, and a more detailed classification must be considered (see the Perspectives in Chapter 9).

Nevertheless, it is interesting to notice that f_{SF} for AGNs has a strong dependence on the SFR estimates used (GSWLC or MPA-JHU). Using the GSWLC datasets, galaxies classified as AGN hosts are found to have higher f_{SF} values than when using MPA-JHU. This indicates that the SFR estimates of one (or both) catalogues might be affected by the

AGN emission. We discuss the differences between the results for MPA-JHU and GSWLC later on in this Chapter.

8.5 Effects of pre-processing

The variations of f_{SF} with clustercentric distance can be affected by the presence of other small haloes in the surroundings of a more massive one. The galaxies within these haloes are more likely to be quenched compared to field galaxies. Therefore, when we include the galaxies within these small haloes to compute f_{SF} (i.e., use a high halo-mass threshold, M_a, for membership assignment, see Fig. 4.3), one would expect to see a decrease in the fraction of star-forming galaxies, especially at large distances. However, contrary to what was expected, we see an *increase* in f_{SF} (Figs. 7.6 and 7.7). As already mentioned in Chapter 7, this increase is due to the fact that a large fraction of the haloes with $\log_{10}(M_h/M_\odot) < 13$ in the Lim *et al.* (2017) catalogue contains only the central galaxy. In addition, even if the original group membership definition made by Lim *et al.* (2017) results in more than one member galaxy, our cut in stellar mass to make our samples complete (Section 5.1) might exclude low-mass satellites from the analysis. Therefore, in practice, we include only the central galaxy of small haloes in our samples.

Several studies have shown that central galaxies tend to have different properties compared to satellites (Pasquali *et al.*, 2010; Bluck *et al.*, 2020b; Gallazzi *et al.*, 2021; Engler *et al.*, 2020; Spindler and Wake, 2017) including their sSFRs. This is also seen in our results, where we find that central galaxies are more likely to be star-forming than satellites in the same bin of σ_e and $M_\star$ (Figures 7.2 to 7.5), in agreement with other studies. Therefore, when we increase M_a and, in practice, add more central than satellite galaxies to the vicinities of a high-mass group, f_{SF} increases. This scenario is confirmed in Figures 7.8 and 7.9, where we show that when excluding all galaxies that are originally centrals in the Lim *et al.* (2017) catalogue, we no longer see an increase in f_{SF}. However, we do not observe a decrease in f_{SF} either; the relations of f_{SF} with R/r_{vir} are very similar for $\log_{10}(M_a/M_\odot) = 12.3$ and 13. These results suggest that the pre-processing in small groups along the cosmic web has not a strong effect on the fraction of star-forming galaxies in the outskirts of larger haloes.

8.6 Differences between GSWLC and MPA-JHU measurements

One important thing to note is that the MPA-JHU SFR estimation methodology relies on accurate AGN decontamination, as their measurements are based on Hα emission lines that need to be decoupled from the non-stellar ionising sources. Salim *et al.* (2007) argues that this led to the MPA-JHU excessively removing star-forming galaxies by misclassifying them as AGN. The GSWLC sample, on the other hand, is based on SED fitting to include the UV information in their estimates. This has the advantage of detecting star formation in dust-obscured galaxies but relies on the accuracy of stellar population models.

In general, f_{SF} estimated based on GSWLC present a higher fraction of star-forming galaxies than MPA-JHU. In contrast to measurements based on Hα, UV provides reliable SFRs for galaxies with weak or no Hα emission and when Hα is contaminated with

emission from an AGN. When we look at the f_{SF} estimates for non-AGN galaxies, the GSWLC sample also presents higher values of f_{SF} than the MPA-JHU, independently of other variables. Therefore, while it is expected that UV-based SFR estimates are somewhat biased by AGN activity, it can not be the sole reason for the higher f_{SF} values obtained in the GSWLC samples.

Hα emission is known to be related to AGN activity (Werner *et al.*, 2014). Therefore it is not surprising that non-AGN galaxies show decreased values of f_{SF} in the `MPAJHU01` sample, as its star formation estimate is based on Hα emission. On the other hand, Salim *et al.*, 2007 have shown that some galaxies with no Hα show star formation activity when it is estimated with the addition of UV information. Therefore, we find that it is possible that the f_{SF} results based on the `MPAJHU01` can be underestimated, especially for non-AGN galaxies.

Chapter 9

Conclusion and perspectives

In this work, we present a robust statistical model to predict the f_{SF} for satellite galaxies based on several observational variables and compare the model predictions with f_{SF} inferred from observations. Using galaxy observables such as stellar mass, velocity dispersion, halo mass, AGN activity and clustercentric distance alongside a logistic model, we are able to achieve a star-forming classification accuracy of, on average, more than 80%. This accuracy was estimated by comparing our model results with literature catalogues that employed the usual star formation tracers, such as UV and Hα emission. The high efficiency of our model indicates how a multivariate statistical analysis of certain galaxy properties is able to be employed as a very useful tool for studies of the quenching of galaxies in the ever-increasing world of large astronomical datasets. However, the choice of variables to be investigated is extremely important since a robust statistical model needs to be parsimonious, i.e., contain only variables that are really necessary to explain the studied outcome (in our case, the f_{SF} galaxies).

Although it is well known that the f_{SF} galaxies depend on galaxy properties and their environment in an intricate way, and there are many studies devoted to this topic, it is still not clear what drives the quenching. In this work, we show that the f_{SF} dependence with clustercentric distance is more reliant on AGN activity and σ_e; $M_\star$ is important in some circumstances, but overall plays a secondary role. In fact, σ_e plays the leading role in influencing the estimated fraction of star-forming galaxies. Moreover, AGN activity was also found to be very important in the model. However, the choice of AGN diagnostic diagram needs to be cautiously taken into consideration. For instance, our choice of using the WHAN diagram may bias our AGN sample towards galaxies in which the star formation is fuelled by a high fraction of gas in the central regions, which can both feed the central black hole and increase the SFR.

We do not find evidence of pre-processing. Through the employment of the assignment scheme from Trevisan *et al.* (2017b), we analysed the consequences of incorporating galaxies from smaller haloes into the outskirts of larger halos. We find that galaxies in smaller halos, classified as centrals by Lim *et al.* (2017), are actually responsible for an increase in the f_{SF} galaxies when incorporated into larger halos. Satellite galaxies from smaller groups play little to no role in changing the f_{SF} when considered as part of the outskirts of larger halos.

We find that, although general trends remain unchanged, the adopted SFR and $M_\star$ estimates to classify galaxies as star-forming and quiescent can significantly change the

results. In accordance with the literature, we find that UV-based estimates lead to a higher fraction of star-forming galaxies when compared to Hα-based estimates.

Our results show that the relation between $f_{\rm SF}$ and clustercentric distance becomes weaker when the analysis includes other galaxy properties, suggesting that the absolute variations of $f_{\rm SF}$ with the environment are at least partially driven by the dependence of galaxy properties with the environment; i.e., the population of galaxies in different environments can be biased towards objects with specific properties.

Perspectives

This section outlines the potential directions for future research, emphasising the broader context of the study and its implications.

1. Exploring additional galaxy properties: While the current study has focused on specific properties associated with quenching, there are other galaxy characteristics that warrant investigation. The density of the galaxy local environment and stellar mass surface density are two such properties that may be interconnected with the observed quenching of star formation. Additionally, we intend to include in our model estimates of the galaxy molecular gas fraction and examine its correlation with both stellar mass and velocity dispersion.

2. Higher-order and interaction terms in the logistic model: The current logistic model has successfully captured significant trends associated with the quenching of star formation. However, exploring higher-order terms and interaction terms within the model could provide a more nuanced representation of the underlying processes. This exploration should enclose properties that may not have been included in the final model due to their marginal significance, like $SF_{\rm central}$ and $\sigma_{\rm e,central}$.

3. Extension of the analysed galaxy stellar mass range: To broaden the scope of our study, it is interesting to incorporate a more diverse range of galaxies into the analysis. Specifically, the addition of low-mass galaxies ($\log_{10}(M_\star/M_\odot) < 10.5$) from deeper surveys can provide valuable insights into the differences between their quenching mechanisms and those affecting more massive galaxies.

4. Improved classification of AGN: The accurately classify galaxies based on AGN activity, we will combine different diagnostic diagrams depending on the signal-to-noise (S/N) and number of lines available in the spectra. We want to explore whether $f_{\rm SF}$ remains high with more refined classifications or whether, in fact, the WHAN diagram is skewing the AGN star-forming fraction of our data. A possible approach would be to test using BPT for galaxies that have the 4 emission lines ([OIII], Hβ, [NII] and Hα) with high signal-to-noise ratio and the WHAN diagram for those that do not.

5. Comparison with cosmological simulations: In order to compare and extend the findings from observational data, it is important to test the obtained results within the framework of cosmological simulations. These simulations allow us to observe the entry of galaxies into clusters and the subsequent effects on $f_{\rm SF}$.

In conclusion, this dissertation showcases the complexity of the galaxy evolution phenomena and the need for further investigation. However, the methodology hereby presented shows itself as a promising tool to efficiently make use of the large-scale data already available in the field and in forthcoming decades. Furthermore, the perspectives outlined above offer promising avenues for future research that can refine our results. We hope our results pave the way for future theoretical and statistical studies taking into

consideration these findings in light of the next generation of telescopes, such as EUCLID, JWST and Vera Rubin, among others.

Appendix A

MSG Query

```
SELECT
  p.objid, programname, instrument,
  s.specobjid, s.plate, s.mjd, s.fiberid, s.ra, s.dec,
  s.z, s.zErr, s.velDisp, s.velDispErr,
  p.petroMag_u - p.extinction_u as magPetro_u,
  p.petroMag_g - p.extinction_g as magPetro_g,
  p.petroMag_r - p.extinction_r as magPetro_r,
  p.petroMag_i - p.extinction_i as magPetro_i,
  p.petroMag_z - p.extinction_z as magPetro_z,
  p.petroR90_r as petroR90_r,
  p.petroR50_r as petroR50_r,
  p.petroRad_r as petroRad_r,
  g.bptClass,
  e.Flux_OIII_5006, e.Flux_Ha_6562, e.Flux_NII_6583, e.Flux_Hb_4861,
  e.EW_Ha_6562, e.EW_NII_6583,
  g.lgm_tot_p50, g.lgm_tot_p16, g.lgm_tot_p84,
  g.sfr_tot_p50, g.sfr_tot_p16, g.sfr_tot_p84

 FROM
    specobj as s,
    photoobj as p,
    galSpecExtra as g,
    emissionLinesPort as e

  WHERE
    s.bestobjid = p.objid AND
    s.specobjid = g.specobjid AND
    s.specobjid = e.specobjid AND
    programname = 'legacy' AND
    s.Class = 'GALAXY' AND
    s.z >= 0.01 AND s.z <= 0.1 AND
    (p.petroMag_r - p.extinction_r) < 17.77 AND
    primTarget & dbo.fPrimTarget('GALAXY') > 0

ORDER BY specobjid
```

Bibliography

1. M. G. Abadi *et al.*, *Monthly Notices of the Royal Astronomical Society* **308**, ADS Bibcode: 1999MNRAS.308..947A, 947–954, ISSN: 0035-8711, (2023; https://ui.adsabs.harvard.edu/abs/1999MNRAS.308..947A) (Oct. 1999).
2. K. Abazajian *et al.*, *The Astronomical Journal* **126**, ADS Bibcode: 2003AJ....126.2081A, 2081–2086, ISSN: 0004-6256, (2023; https://ui.adsabs.harvard.edu/abs/2003AJ....126.2081A) (Oct. 2003).
3. K. Abazajian *et al.*, *The Astronomical Journal* **128**, ADS Bibcode: 2004AJ....128..502A, 502–512, ISSN: 0004-6256, (2023; https://ui.adsabs.harvard.edu/abs/2004AJ....128..502A) (July 2004).
4. K. N. Abazajian *et al.*, *The Astrophysical Journal Supplement Series* **182**, ADS Bibcode: 2009ApJS..182..543A, 543–558, ISSN: 0067-0049, (2023; https://ui.adsabs.harvard.edu/abs/2009ApJS..182..543A) (June 2009).
5. A. Almeida *et al.*, *The Eighteenth Data Release of the Sloan Digital Sky Surveys: Targeting and First Spectra from SDSS-V*, arXiv:2301.07688 [astro-ph], Jan. 2023, (2023; http://arxiv.org/abs/2301.07688).
6. Y. M. Bahé *et al.*, *Monthly Notices of the Royal Astronomical Society* **430**, arXiv:1210.8407 [astro-ph], 3017–3031, ISSN: 1365-2966, 0035-8711, (2023; http://arxiv.org/abs/1210.8407) (Apr. 2013).
7. I. K. Baldry *et al.*, *Monthly Notices of the Royal Astronomical Society* **373**, 469–483, ISSN: 0035-8711, (2023; https://doi.org/10.1111/j.1365-2966.2006.11081.x) (Dec. 2006).
8. I. K. Baldry *et al.*, en, *The Astrophysical Journal* **600**, Publisher: IOP Publishing, 681, ISSN: 0004-637X, (2023; https://iopscience.iop.org/article/10.1086/380092/meta) (Jan. 2004).
9. J. A. Baldwin *et al.*, *Publications of the Astronomical Society of the Pacific* **93**, ADS Bibcode: 1981PASP...93....5B, 5–19, ISSN: 0004-6280, (2023; https://ui.adsabs.harvard.edu/abs/1981PASP...93....5B) (Feb. 1981).
10. N. M. Ball *et al.*, *Monthly Notices of the Royal Astronomical Society* **383**, 907–922, ISSN: 0035-8711, (2023; https://doi.org/10.1111/j.1365-2966.2007.12627.x) (Jan. 2008).

11. M. L. Balogh, S. L. Morris, *Monthly Notices of the Royal Astronomical Society* **318**, ADS Bibcode: 2000MNRAS.318..703B, 703–714, ISSN: 0035-8711, (2023; https://ui.adsabs.harvard.edu/abs/2000MNRAS.318..703B) (Nov. 2000).
12. M. L. Balogh *et al.*, *The Astrophysical Journal* **540**, ADS Bibcode: 2000ApJ...540..113B, 113–121, ISSN: 0004-637X, (2023; https://ui.adsabs.harvard.edu/abs/2000ApJ...540..113B) (Sept. 2000).
13. S. P. Bamford *et al.*, *Monthly Notices of the Royal Astronomical Society* **393**, ADS Bibcode: 2009MNRAS.393.1324B, 1324–1352, ISSN: 0035-8711, (2023; https://ui.adsabs.harvard.edu/abs/2009MNRAS.393.1324B) (Mar. 2009).
14. G. Barro *et al.*, **225**, Conference Name: American Astronomical Society Meeting Abstracts #225 ADS Bibcode: 2015AAS...22511107B, 111.07, (2023; https://ui.adsabs.harvard.edu/abs/2015AAS...22511107B) (Jan. 2015).
15. E. F. Bell *et al.*, *The Astrophysical Journal* **608**, ADS Bibcode: 2004ApJ...608..752B, 752–767, ISSN: 0004-637X, (2023; https://ui.adsabs.harvard.edu/abs/2004ApJ...608..752B) (June 2004).
16. R. B. Bendel, A. A. Afifi, *Journal of the American Statistical Association* **72**, Publisher: [American Statistical Association, Taylor & Francis, Ltd.], 46–53, ISSN: 0162-1459, (2023; https://www.jstor.org/stable/2286904) (1977).
17. C. L. Bennett *et al.*, *The Astrophysical Journal Supplement Series* **208**, ADS Bibcode: 2013ApJS..208...20B, 20, ISSN: 0067-0049, (2023; https://ui.adsabs.harvard.edu/abs/2013ApJS..208...20B) (Oct. 2013).
18. J. Berkson, *Journal of the American Statistical Association* **39**, Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.1944.10500699, 357–365, ISSN: 0162-1459, (2023; https://doi.org/10.1080/01621459.1944.10500699) (Sept. 1944).
19. J. C. Berrier *et al.*, *The Astrophysical Journal* **690**, ADS Bibcode: 2009ApJ...690.1292B, 1292–1302, ISSN: 0004-637X, (2023; https://ui.adsabs.harvard.edu/abs/2009ApJ...690.1292B) (Jan. 2009).
20. A. M. Berti *et al.*, *The Astronomical Journal* **161**, arXiv:2009.02405 [astro-ph], 49, ISSN: 0004-6256, 1538-3881, (2023; http://arxiv.org/abs/2009.02405) (Jan. 2021).
21. M. R. Blanton, *The Astrophysical Journal* **648**, arXiv:astro-ph/0512127, 268–280, ISSN: 0004-637X, 1538-4357, (2023; http://arxiv.org/abs/astro-ph/0512127) (Sept. 2006).
22. M. R. Blanton, S. Roweis, *The Astronomical Journal* **133**, ADS Bibcode: 2007AJ....133..734B, 734–754, ISSN: 0004-6256, (2023; https://ui.adsabs.harvard.edu/abs/2007AJ....133..734B) (Feb. 2007).
23. M. R. Blanton *et al.*, en, *The Astrophysical Journal* **594**, Publisher: IOP Publishing, 186, ISSN: 0004-637X, (2023; https://iopscience.iop.org/article/10.1086/375528/meta) (Sept. 2003).

24. M. R. Blanton *et al.*, en, *The Astrophysical Journal* **631**, Publisher: IOP Publishing, 208, ISSN: 0004-637X, (2023; https://iopscience.iop.org/article/10.1086/431416/meta) (Sept. 2005).

25. A. F. L. Bluck *et al.*, *Monthly Notices of the Royal Astronomical Society* **492**, ADS Bibcode: 2020MNRAS.492...96B, 96–139, ISSN: 0035-8711, (2023; https://ui.adsabs.harvard.edu/abs/2020MNRAS.492...96B) (Feb. 2020).

26. A. F. L. Bluck *et al.*, *Monthly Notices of the Royal Astronomical Society* **499**, 230–268, ISSN: 0035-8711, (2023; https://doi.org/10.1093/mnras/staa2806) (Oct. 2020).

27. A. F. L. Bluck *et al.*, *Monthly Notices of the Royal Astronomical Society* **462**, ADS Bibcode: 2016MNRAS.462.2559B, 2559–2586, ISSN: 0035-8711, (2023; https://ui.adsabs.harvard.edu/abs/2016MNRAS.462.2559B) (Nov. 2016).

28. A. F. L. Bluck *et al.*, *Astronomy and Astrophysics* **659**, ADS Bibcode: 2022A&A...659A.160B, A160, ISSN: 0004-6361, (2023; https://ui.adsabs.harvard.edu/abs/2022A&A...659A.160B) (Mar. 2022).

29. G. R. Blumenthal *et al.*, *Nature* **311**, ADS Bibcode: 1984Natur.311..517B, 517–525, ISSN: 0028-0836, (2023; https://ui.adsabs.harvard.edu/abs/1984Natur.311..517B) (Oct. 1984).

30. A. Bravais, *Analyse mathématique: sur les probabilités des erreurs de situation d'un point*, fr.

31. J. Brinchmann *et al.*, *Monthly Notices of the Royal Astronomical Society* **351**, ADS Bibcode: 2004MNRAS.351.1151B, 1151–1179, ISSN: 0035-8711, (2023; https://ui.adsabs.harvard.edu/abs/2004MNRAS.351.1151B) (July 2004).

32. S. Brownson *et al.*, *Monthly Notices of the Royal Astronomical Society* **511**, arXiv:2201.02484 [astro-ph], 1913–1941, ISSN: 0035-8711, 1365-2966, (2023; http://arxiv.org/abs/2201.02484) (Feb. 2022).

33. G. L. Bryan, M. L. Norman, *The Astrophysical Journal* **495**, ADS Bibcode: 1998ApJ...495...80B, 80–99, ISSN: 0004-637X, (2023; https://ui.adsabs.harvard.edu/abs/1998ApJ...495...80B) (Mar. 1998).

34. M. Cappellari, E. Emsellem, *Publications of the Astronomical Society of the Pacific* **116**, ADS Bibcode: 2004PASP..116..138C, 138–147, ISSN: 0004-6280, (2023; https://ui.adsabs.harvard.edu/abs/2004PASP..116..138C) (Feb. 2004).

35. M. Cappellari *et al.*, *Monthly Notices of the Royal Astronomical Society* **366**, 1126–1150, ISSN: 0035-8711, (2023; https://doi.org/10.1111/j.1365-2966.2005.09981.x) (Mar. 2006).

36. G. Casella, R. L. Berger, *Statistical Inference*, en, Google-Books-ID: FAUVEAAAQBAJ (Cengage Learning, Jan. 2001), ISBN: 978-0-357-75313-2.

37. A. Cattaneo *et al.*, en, *Nature* **460**, Number: 7252 Publisher: Nature Publishing Group, 213–219, ISSN: 1476-4687, (2023; https://www.nature.com/articles/nature08135) (July 2009).

38. C. Cicone *et al.*, *Astronomy & Astrophysics* **562**, arXiv:1311.2595 [astro-ph], A21, ISSN: 0004-6361, 1432-0746, (2023; http://arxiv.org/abs/1311.2595) (Feb. 2014).
39. W. S. Cleveland, en, *Journal of the American Statistical Association* (1979).
40. J. Cohen, eng, *Psychological Bulletin* **112**, 155–159, ISSN: 0033-2909 (July 1992).
41. D. R. Cox, en, *Journal of the Royal Statistical Society. Series B (Methodological)* **20**, 215–242, (http://www.jstor.org/stable/2983890) (1958).
42. D. J. Croton *et al.*, *Monthly Notices of the Royal Astronomical Society* **365**, 11–28, ISSN: 0035-8711, (2023; https://doi.org/10.1111/j.1365-2966.2005.09675.x) (Jan. 2006).
43. B. Darvish *et al.*, en, *The Astrophysical Journal* **825**, Publisher: The American Astronomical Society, 113, ISSN: 0004-637X, (2023; https://dx.doi.org/10.3847/0004-637X/825/2/113) (July 2016).
44. M. Davis, M. J. Geller, *The Astrophysical Journal* **208**, ADS Bibcode: 1976ApJ...208...13D, 13–19, ISSN: 0004-637X, (2023; https://ui.adsabs.harvard.edu/abs/1976ApJ...208...13D) (Aug. 1976).
45. G. de Vaucouleurs, *Handbuch der Physik* **53**, ADS Bibcode: 1959HDP....53..275D, 275, (2023; https://ui.adsabs.harvard.edu/abs/1959HDP....53..275D) (Jan. 1959).
46. A. Dekel, J. Silk, *The Astrophysical Journal* **303**, ADS Bibcode: 1986ApJ...303...39D, 39, ISSN: 0004-637X, (2023; https://ui.adsabs.harvard.edu/abs/1986ApJ...303...39D) (Apr. 1986).
47. A. Dressler, *The Astrophysical Journal* **236**, ADS Bibcode: 1980ApJ...236..351D, 351–365, ISSN: 0004-637X, (2023; https://ui.adsabs.harvard.edu/abs/1980ApJ...236..351D) (Mar. 1980).
48. S. P. Driver *et al.*, *Monthly Notices of the Royal Astronomical Society* **368**, 414–434, ISSN: 0035-8711, (2023; https://doi.org/10.1111/j.1365-2966.2006.10126.x) (May 2006).
49. M. Duarte, G. A. Mamon, *Monthly Notices of the Royal Astronomical Society* **458**, arXiv:1412.3364 [astro-ph], 1301–1301, ISSN: 0035-8711, 1365-2966, (2023; http://arxiv.org/abs/1412.3364) (May 2016).
50. C. Engler *et al.*, *Monthly Notices of the Royal Astronomical Society* **500**, arXiv:2002.11119 [astro-ph], 3957–3975, ISSN: 0035-8711, 1365-2966, (2023; http://arxiv.org/abs/2002.11119) (Dec. 2020).
51. A. Fabian, *Annual Review of Astronomy and Astrophysics* **50**, 455–489, (2023; https://doi.org/10.1146/annurev-astro-081811-125521) (2012).
52. J. J. Fang *et al.*, en, *The Astrophysical Journal* **776**, Publisher: The American Astronomical Society, 63, ISSN: 0004-637X, (2023; https://dx.doi.org/10.1088/0004-637X/776/1/63) (Sept. 2013).

53. R. Farouki, S. L. Shapiro, *The Astrophysical Journal* **243**, ADS Bibcode: 1981ApJ...243...32F, 32–41, ISSN: 0004-637X, (2023; https://ui.adsabs.harvard.edu/abs/1981ApJ...243...32F) (Jan. 1981).

54. R. Feldmann *et al.*, *The Astrophysical Journal* **709**, ADS Bibcode: 2010ApJ...709..218F, 218–240, ISSN: 0004-637X, (2023; https://ui.adsabs.harvard.edu/abs/2010ApJ...709..218F) (Jan. 2010).

55. R. C. Fernandes *et al.*, en, *Monthly Notices of the Royal Astronomical Society* **413**, 1687–1699, ISSN: 00358711 (May 2011).

56. R. C. Fernandes *et al.*, *Monthly Notices of the Royal Astronomical Society* **403**, 1036–1053, ISSN: 00358711 (Feb. 2010).

57. L. Ferrarese, D. Merritt, *The Astrophysical Journal* **539**, ADS Bibcode: 2000ApJ...539L...9F, L9–L12, ISSN: 0004-637X, (2023; https://ui.adsabs.harvard.edu/abs/2000ApJ...539L...9F) (Aug. 2000).

58. R. A. Fisher, *Philosophical Transactions of the Royal Society of London Series A* **222**, ADS Bibcode: 1922RSPTA.222..309F, 309–368, ISSN: 1364-503X0080-46140962-8436, (2023; https://ui.adsabs.harvard.edu/abs/1922RSPTA.222..309F) (Jan. 1922).

59. R. A. Fisher, *Journal of the Royal Statistical Society* **85**, Publisher: [Wiley, Royal Statistical Society], 597–612, ISSN: 0952-8385, (2023; https://www.jstor.org/stable/2341124) (1922).

60. A. R. Gallazzi *et al.*, *Monthly Notices of the Royal Astronomical Society* **502**, arXiv:2010.04733 [astro-ph], 4457–4478, ISSN: 0035-8711, 1365-2966, (2023; http://arxiv.org/abs/2010.04733) (Feb. 2021).

61. K. Gebhardt *et al.*, *The Astrophysical Journal* **539**, ADS Bibcode: 2000ApJ...539L..13G, L13–L16, ISSN: 0004-637X, (2023; https://ui.adsabs.harvard.edu/abs/2000ApJ...539L..13G) (Aug. 2000).

62. J. R. Gott III, M. J. Rees, *Astronomy and Astrophysics* **45**, ADS Bibcode: 1975A&A....45..365G, 365–376, ISSN: 0004-6361, (2023; https://ui.adsabs.harvard.edu/abs/1975A&A....45..365G) (Dec. 1975).

63. M. T. Graham *et al.*, *Monthly Notices of the Royal Astronomical Society* **477**, 4711–4737, ISSN: 0035-8711, (2023; https://doi.org/10.1093/mnras/sty504) (July 2018).

64. J. E. Gunn, J. R. Gott III, *The Astrophysical Journal* **176**, ADS Bibcode: 1972ApJ...176....1G, 1, ISSN: 0004-637X, (2023; https://ui.adsabs.harvard.edu/abs/1972ApJ...176....1G) (Aug. 1972).

65. J. E. Gunn *et al.*, *The Astronomical Journal* **131**, ADS Bibcode: 2006AJ....131.2332G, 2332–2359, ISSN: 0004-6256, (2023; https://ui.adsabs.harvard.edu/abs/2006AJ....131.2332G) (Apr. 2006).

66. Y. Guo *et al.*, *The Astrophysical Journal Supplement Series* **207**, 24, ISSN: 0067-0049, (2023; https://ui.adsabs.harvard.edu/abs/2013ApJS..207...24G) (Aug. 2013).

67. S. D. Gupta, en, *Psychometrika* **25**, 393–408, ISSN: 1860-0980, (2023; https://doi.org/10.1007/BF02289756) (Dec. 1960).

68. P. L. Gómez *et al.*, *The Astrophysical Journal* **584**, ADS Bibcode: 2003ApJ...584..210G, 210–227, ISSN: 0004-637X, (2023; https://ui.adsabs.harvard.edu/abs/2003ApJ...584..210G) (Feb. 2003).

69. E. Gómez-de Mariscal *et al.*, en, *Scientific Reports* **11**, Number: 1 Publisher: Nature Publishing Group, 20942, ISSN: 2045-2322, (2023; https://www.nature.com/articles/s41598-021-00199-5) (Oct. 2021).

70. C. P. Haines *et al.*, *The Astrophysical Journal* **806**, ADS Bibcode: 2015ApJ...806..101H, 101, ISSN: 0004-637X, (2023; https://ui.adsabs.harvard.edu/abs/2015ApJ...806..101H) (June 2015).

71. A. J. S. Hamilton, M. Tegmark, *Monthly Notices of the Royal Astronomical Society* **349**, 115–128, ISSN: 0035-8711, (2023; https://doi.org/10.1111/j.1365-2966.2004.07490.x) (Mar. 2004).

72. B. M. B. Henriques *et al.*, *Monthly Notices of the Royal Astronomical Society* **451**, ADS Bibcode: 2015MNRAS.451.2663H, 2663–2680, ISSN: 0035-8711, (2023; https://ui.adsabs.harvard.edu/abs/2015MNRAS.451.2663H) (Aug. 2015).

73. G. Hinshaw *et al.*, *The Astrophysical Journal Supplement Series* **208**, ADS Bibcode: 2013ApJS..208...19H, 19, ISSN: 0067-0049, (2023; https://ui.adsabs.harvard.edu/abs/2013ApJS..208...19H) (Oct. 2013).

74. P. F. Hopkins *et al.*, *Monthly Notices of the Royal Astronomical Society* **397**, 802–814, ISSN: 0035-8711, (2023; https://doi.org/10.1111/j.1365-2966.2009.14983.x) (Aug. 2009).

75. R. Huang *et al.*, *Monthly Notices of the Royal Astronomical Society* **520**, arXiv:2301.01995 [astro-ph], 446–460, ISSN: 0035-8711, 1365-2966, (2023; http://arxiv.org/abs/2301.01995) (Jan. 2023).

76. E. P. Hubble, *The Astrophysical Journal* **64**, ADS Bibcode: 1926ApJ....64..321H, 321–369, ISSN: 0004-637X, (2023; https://ui.adsabs.harvard.edu/abs/1926ApJ....64..321H) (Dec. 1926).

77. E. P. Hubble, *Realm of the Nebulae*, Publication Title: Realm of the Nebulae ADS Bibcode: 1936rene.book.....H, (2023; https://ui.adsabs.harvard.edu/abs/1936rene.book.....H).

78. E. Hubble, *Proceedings of the National Academy of Science* **15**, 168–173, ISSN: 0027-8424 (Mar. 1929).

79. M. L. Humason, *The Astrophysical Journal* **83**, ADS Bibcode: 1936ApJ....83...10H, 10, ISSN: 0004-637X, (2023; https://ui.adsabs.harvard.edu/abs/1936ApJ....83...10H) (Jan. 1936).

80. N. Häring, H.-W. Rix, *The Astrophysical Journal* **604**, ADS Bibcode: 2004ApJ...604L..89H, L89–L92, ISSN: 0004-637X, (2023; https://ui.adsabs.harvard.edu/abs/2004ApJ...604L..89H) (Apr. 2004).

81. G. Kauffmann *et al.*, *Monthly Notices of the Royal Astronomical Society* **264**, 201–218, ISSN: 0035-8711, (2023; https://doi.org/10.1093/mnras/264.1.201) (Sept. 1993).

82. G. Kauffmann *et al.*, *Monthly Notices of the Royal Astronomical Society* **341**, ADS Bibcode: 2003MNRAS.341...33K, 33–53, ISSN: 0035-8711, (2023; https://ui.adsabs.harvard.edu/abs/2003MNRAS.341...33K) (May 2003).

83. G. Kauffmann *et al.*, *Monthly Notices of the Royal Astronomical Society* **353**, ADS Bibcode: 2004MNRAS.353..713K, 713–731, ISSN: 0035-8711, (2023; https://ui.adsabs.harvard.edu/abs/2004MNRAS.353..713K) (Sept. 2004).

84. L. S. Kelvin *et al.*, *Monthly Notices of the Royal Astronomical Society* **444**, 1647–1659, ISSN: 0035-8711, (2023; https://doi.org/10.1093/mnras/stu1507) (Oct. 2014).

85. R. C. Kennicutt Jr., *The Astrophysical Journal* **272**, ADS Bibcode: 1983ApJ...272...54K, 54–67, ISSN: 0004-637X, (2023; https://ui.adsabs.harvard.edu/abs/1983ApJ...272...54K) (Sept. 1983).

86. R. C. Kennicutt, N. J. Evans, *Annual Review of Astronomy and Astrophysics* **50**, 531–608 (2012).

87. R. E. Kirk, en, *Educational and Psychological Measurement* **56**, Publisher: SAGE Publications Inc, 746–759, ISSN: 0013-1644, (2023; https://doi.org/10.1177/0013164496056005002) (Oct. 1996).

88. C. Knobel *et al.*, *The Astrophysical Journal* **800**, ADS Bibcode: 2015ApJ...800...24K, 24, ISSN: 0004-637X, (2023; https://ui.adsabs.harvard.edu/abs/2015ApJ...800...24K) (Feb. 2015).

89. P. Kroupa, *Monthly Notices of the Royal Astronomical Society* **322**, ADS Bibcode: 2001MNRAS.322..231K, 231–246, ISSN: 0035-8711, (2023; https://ui.adsabs.harvard.edu/abs/2001MNRAS.322..231K) (Apr. 2001).

90. F. La Barbera *et al.*, *Monthly Notices of the Royal Astronomical Society* **408**, 1313–1334, ISSN: 0035-8711, (2022; https://doi.org/10.1111/j.1365-2966.2010.16850.x) (Nov. 2010).

91. C. Lacey, S. Cole, en, *Monthly Notices of the Royal Astronomical Society* **262**, 627–649, ISSN: 0035-8711, 1365-2966, (2023; https://academic.oup.com/mnras/article-lookup/doi/10.1093/mnras/262.3.627) (June 1993).

92. R. B. Larson *et al.*, *The Astrophysical Journal* **237**, ADS Bibcode: 1980ApJ...237..692L, 692–707, ISSN: 0004-637X, (2023; https://ui.adsabs.harvard.edu/abs/1980ApJ...237..692L) (May 1980).

93. R. B. Larson, *Monthly Notices of the Royal Astronomical Society* **169**, 229–245, ISSN: 0035-8711, (2023; https://doi.org/10.1093/mnras/169.2.229) (Nov. 1974).

94. I. Lewis *et al.*, *Monthly Notices of the Royal Astronomical Society* **334**, 673–683, ISSN: 0035-8711, (2023; https://doi.org/10.1046/j.1365-8711.2002.05558.x) (Aug. 2002).

95. S. H. Lim *et al.*, *Monthly Notices of the Royal Astronomical Society* **470**, ADS Bibcode: 2017MNRAS.470.2982L, 2982–3005, ISSN: 0035-8711, (2022; https://ui.adsabs.harvard.edu/abs/2017MNRAS.470.2982L) (Sept. 2017).

96. M. Lin *et al.*, *Information Systems Research* **24**, Place: US Publisher: Institute for Operations Research & the Management Sciences (INFORMS), 906–917, ISSN: 1526-5536 (2013).

97. E. L. Lokas, G. A. Mamon, *Monthly Notices of the Royal Astronomical Society* **321**, arXiv:astro-ph/0002395, 155–166, ISSN: 0035-8711, 1365-2966, (2023; http://arxiv.org/abs/astro-ph/0002395) (Feb. 2001).

98. T. Lu *et al.*, *Monthly Notices of the Royal Astronomical Society* **420**, 126–140, ISSN: 0035-8711, (2023; https://doi.org/10.1111/j.1365-2966.2011.20008.x) (Feb. 2012).

99. A. V. Macciò *et al.*, *Monthly Notices of the Royal Astronomical Society* **391**, 1940–1954, ISSN: 0035-8711, (2022; https://doi.org/10.1111/j.1365-2966.2008.14029.x) (Dec. 2008).

100. J. Magorrian *et al.*, *The Astronomical Journal* **115**, ADS Bibcode: 1998AJ....115.2285M, 2285–2305, ISSN: 0004-6256, (2023; https://ui.adsabs.harvard.edu/abs/1998AJ....115.2285M) (June 1998).

101. S. Mahajan *et al.*, *Monthly Notices of the Royal Astronomical Society* **416**, arXiv:1106.3062 [astro-ph], 2882–2902, ISSN: 00358711, (2023; http://arxiv.org/abs/1106.3062) (Oct. 2011).

102. G. A. Mamon *et al.*, en, *Astronomy & Astrophysics* **520**, Publisher: EDP Sciences, A30, ISSN: 0004-6361, 1432-0746, (2022; https://www.aanda.org/articles/aa/abs/2010/12/aa13948-09/aa13948-09.html) (Sept. 2010).

103. G. A. Mamon *et al.*, *Monthly Notices of the Royal Astronomical Society* **492**, arXiv:1912.06522 [astro-ph], 1791–1811, ISSN: 0035-8711, 1365-2966, (2023; http://arxiv.org/abs/1912.06522) (Feb. 2020).

104. C. Maraston, G. Strömbäck, *Monthly Notices of the Royal Astronomical Society* **418**, ADS Bibcode: 2011MNRAS.418.2785M, 2785–2811, ISSN: 0035-8711, (2023; https://ui.adsabs.harvard.edu/abs/2011MNRAS.418.2785M) (Dec. 2011).

105. A. Marconi *et al.*, *Monthly Notices of the Royal Astronomical Society* **351**, 169–185, ISSN: 0035-8711, (2023; https://doi.org/10.1111/j.1365-2966.2004.07765.x) (June 2004).

106. A. Marconi, L. K. Hunt, *The Astrophysical Journal* **589**, ADS Bibcode: 2003ApJ...589L..21M, L21–L24, ISSN: 0004-637X, (2023; https://ui.adsabs.harvard.edu/abs/2003ApJ...589L..21M) (May 2003).

107. D. C. Martin *et al.*, en, *The Astrophysical Journal* **619**, arXiv:astro-ph/0411302, L1–L6, ISSN: 0004-637X, 1538-4357, (2022; http://arxiv.org/abs/astro-ph/0411302) (Jan. 2005).

108. J. Matthee, J. Schaye, *Monthly Notices of the Royal Astronomical Society* **484**, arXiv:1805.05956 [astro-ph], 915–932, ISSN: 0035-8711, 1365-2966, (2023; http://arxiv.org/abs/1805.05956) (Mar. 2019).

109. S. L. McGee *et al.*, *Monthly Notices of the Royal Astronomical Society* **400**, arXiv:0908.0750 [astro-ph], 937–950, ISSN: 00358711, 13652966, (2023; http://arxiv.org/abs/0908.0750) (Dec. 2009).

110. M. L. McHugh, *Biochemia Medica* **23**, 143–149, ISSN: 1330-0962, (2023; https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900058/) (June 2013).

111. R. M. Mickey, S. Greenlan, *American Journal of Epidemiology* **129**, 125–137, ISSN: 0002-9262, (2023; https://doi.org/10.1093/oxfordjournals.aje.a115101) (Jan. 1989).

112. B. Moore *et al.*, *Nature* **379**, ADS Bibcode: 1996Natur.379..613M, 613–616, ISSN: 0028-0836, (2023; https://ui.adsabs.harvard.edu/abs/1996Natur.379..613M) (Feb. 1996).

113. S. I. Muldrew *et al.*, en, *Monthly Notices of the Royal Astronomical Society* **419**, 2670–2682, ISSN: 00358711, (2023; https://academic.oup.com/mnras/article-lookup/doi/10.1111/j.1365-2966.2011.19922.x) (Jan. 2012).

114. J. F. Navarro *et al.*, *The Astrophysical Journal* **490**, arXiv:astro-ph/9611107, 493–508, ISSN: 0004-637X, 1538-4357, (2023; http://arxiv.org/abs/astro-ph/9611107) (Dec. 1997).

115. J. F. Navarro *et al.*, *The Astrophysical Journal* **462**, ADS Bibcode: 1996ApJ...462..563N, 563, ISSN: 0004-637X, (2022; https://ui.adsabs.harvard.edu/abs/1996ApJ...462..563N) (May 1996).

116. A. Oemler Jr., *The Astrophysical Journal* **194**, ADS Bibcode: 1974ApJ...194....1O, 1–20, ISSN: 0004-637X, (2023; https://ui.adsabs.harvard.edu/abs/1974ApJ...194....10) (Nov. 1974).

117. A. Pasquali *et al.*, *Monthly Notices of the Royal Astronomical Society* **407**, arXiv:0912.1853 [astro-ph], 937–954, ISSN: 00358711, (2023; http://arxiv.org/abs/0912.1853) (Sept. 2010).

118. S. G. Patel *et al.*, *The Astrophysical Journal* **766**, arXiv:1208.0341 [astro-ph], 15, ISSN: 0004-637X, 1538-4357, (2023; http://arxiv.org/abs/1208.0341) (Mar. 2013).

119. P. J. E. Peebles, *The large-scale structure of the universe*, Publication Title: Large-Scale Structure of the Universe by Phillip James Edwin Peebles. Princeton University Press ADS Bibcode: 1980lssu.book.....P, (2023; https://ui.adsabs.harvard.edu/abs/1980lssu.book.....P).

120. Y. Peng *et al.*, en, *Nature* **521**, Number: 7551 Publisher: Nature Publishing Group, 192–195, ISSN: 1476-4687, (2023; https://www.nature.com/articles/nature14439) (May 2015).

121. Y.-j. Peng *et al.*, en, *The Astrophysical Journal* **721**, 193–221, ISSN: 0004-637X, 1538-4357, (2022; https://iopscience.iop.org/article/10.1088/0004-637X/721/1/193) (Sept. 2010).

122. S. Perlmutter *et al.*, en, *The Astrophysical Journal* **517**, 565, ISSN: 0004-637X, (2023; https://dx.doi.org/10.1086/307221) (June 1999).

123. V. Petrosian, *The Astrophysical Journal* **210**, ADS Bibcode: 1976ApJ...209L...1P, L53, ISSN: 0004-637X, (2023; https://ui.adsabs.harvard.edu/abs/1976ApJ...209L...1P) (Dec. 1976).

124. Planck Collaboration *et al.*, *Astronomy and Astrophysics* **641**, A6, ISSN: 0004-6361 (Sept. 2020).

125. W. H. Press, P. Schechter, *The Astrophysical Journal* **187**, ADS Bibcode: 1974ApJ...187..425P, 425–438, ISSN: 0004-637X, (2023; https://ui.adsabs.harvard.edu/abs/1974ApJ...187..425P) (Feb. 1974).

126. V. Quilis *et al.*, *Science* **288**, arXiv:astro-ph/0006031, 1617–1620, ISSN: 0036-8075, 1095-9203, (2023; http://arxiv.org/abs/astro-ph/0006031) (June 2000).

127. J. R. Quinlan, presented at the.

128. A. L. B. Ribeiro *et al.*, en, *Astronomy & Astrophysics* **505**, Number: 2 Publisher: EDP Sciences, 521–528, ISSN: 0004-6361, 1432-0746, (2023; https://www.aanda.org/articles/aa/abs/2009/38/aa12591-09/aa12591-09.html) (Oct. 2009).

129. D. J. Rosario *et al.*, *The Astrophysical Journal* **771**, ADS Bibcode: 2013ApJ...771...63R, 63, ISSN: 0004-637X, (2023; https://ui.adsabs.harvard.edu/abs/2013ApJ...771...63R) (July 2013).

130. D. J. Rosario *et al.*, *Astronomy and Astrophysics* **560**, ADS Bibcode: 2013A&A...560A..72R, A72, ISSN: 0004-6361, (2023; https://ui.adsabs.harvard.edu/abs/2013A&A...560A..72R) (Dec. 2013).

131. P. Royston, D. G. Altman, *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **43**, Publisher: [Wiley, Royal Statistical Society], 429–467, ISSN: 0035-9254, (2023; https://www.jstor.org/stable/2986270) (1994).

132. S. Salim *et al.*, *The Astrophysical Journal* **859**, 11, ISSN: 0004-637X, (2023; https://ui.adsabs.harvard.edu/abs/2018ApJ...859...11S) (May 2018).

133. S. Salim *et al.*, *The Astrophysical Journal Supplement Series* **227**, 2, ISSN: 0067-0049 (Nov. 2016).

134. S. Salim *et al.*, en, *The Astrophysical Journal Supplement Series* **173**, arXiv:0704.3611 [astro-ph], 267–292, ISSN: 0067-0049, 1538-4365, (2022; http://arxiv.org/abs/0704.3611) (Dec. 2007).

135. P. Santini *et al.*, *The Astrophysical Journal* **847**, ADS Bibcode: 2017ApJ...847...76S, 76, ISSN: 0004-637X, (2023; https://ui.adsabs.harvard.edu/abs/2017ApJ...847...76S) (Sept. 2017).

136. M. Sarzi *et al.*, *Monthly Notices of the Royal Astronomical Society* **366**, ADS Bibcode: 2006MNRAS.366.1151S, 1151–1200, ISSN: 0035-8711, (2023; https://ui.adsabs.harvard.edu/abs/2006MNRAS.366.1151S) (Mar. 2006).

137. K. Schawinski *et al.*, *Monthly Notices of the Royal Astronomical Society* **451**, 2517–2523, ISSN: 0035-8711, (2023; https://doi.org/10.1093/mnras/stv1136) (Aug. 2015).

138. K. Schawinski *et al.*, *Monthly Notices of the Royal Astronomical Society* **382**, 1415–1431, ISSN: 0035-8711, (2023; https://doi.org/10.1111/j.1365-2966.2007.12487.x) (Dec. 2007).

139. C. Schreiber *et al.*, *Astronomy & Astrophysics* **575**, arXiv:1409.5433 [astro-ph], A74, ISSN: 0004-6361, 1432-0746, (2023; http://arxiv.org/abs/1409.5433) (Mar. 2015).

140. R. K. Sheth *et al.*, *The Astrophysical Journal* **650**, ADS Bibcode: 2006ApJ...650L..25S, L25–L28, ISSN: 0004-637X, (2023; https://ui.adsabs.harvard.edu/abs/2006ApJ...650L..25S) (Oct. 2006).

141. J. Silk, M. J. Rees, *Astronomy and Astrophysics* **331**, ADS Bibcode: 1998A&A...331L...1S, L1–L4, ISSN: 0004-6361, (2023; https://ui.adsabs.harvard.edu/abs/1998A&A...331L...1S) (Mar. 1998).

142. L. Simard *et al.*, *The Astrophysical Journal Supplement Series* **196**, arXiv:1107.1518 [astro-ph], 11, ISSN: 0067-0049, 1538-4365, (2023; http://arxiv.org/abs/1107.1518) (Sept. 2011).

143. R. A. Skibba *et al.*, *Monthly Notices of the Royal Astronomical Society* **399**, arXiv:0811.3970 [astro-ph], 966–982, ISSN: 00358711, 13652966, (2023; http://arxiv.org/abs/0811.3970) (Oct. 2009).

144. R. Smith *et al.*, *Monthly Notices of the Royal Astronomical Society* **405**, 1723–1735, ISSN: 0035-8711, (2023; https://doi.org/10.1111/j.1365-2966.2010.16545.x) (July 2010).

145. A. Spindler, D. Wake, *Monthly Notices of the Royal Astronomical Society* **468**, arXiv:1701.07069 [astro-ph], 333–345, ISSN: 0035-8711, 1365-2966, (2023; http://arxiv.org/abs/1701.07069) (June 2017).

146. I. Strateva *et al.*, en, *The Astronomical Journal* **122**, Publisher: IOP Publishing, 1861, ISSN: 1538-3881, (2023; https://iopscience.iop.org/article/10.1086/323301/meta) (Oct. 2001).

147. M. A. Strauss *et al.*, *The Astronomical Journal* **124**, arXiv:astro-ph/0206225, 1810–1824, ISSN: 00046256, 15383881, (2023; http://arxiv.org/abs/astro-ph/0206225) (Sept. 2002).

148. G. M. Sullivan, R. Feinn, *Journal of Graduate Medical Education* **4**, 279–282, ISSN: 1949-8349, (2023; https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3444174/) (Sept. 2012).

149. M. E. C. Swanson *et al.*, *Monthly Notices of the Royal Astronomical Society* **387**, 1391–1402, ISSN: 0035-8711, (2023; https://doi.org/10.1111/j.1365-2966.2008.13296.x) (July 2008).

150. S. Tacchella *et al.*, *Science* **348**, Publisher: American Association for the Advancement of Science, 314–317, (2023; https://www.science.org/doi/10.1126/science.1261094) (Apr. 2015).

151. S. Tacchella *et al.*, *Monthly Notices of the Royal Astronomical Society* **457**, arXiv:1509.02529 [astro-ph], 2790–2813, ISSN: 0035-8711, 1365-2966, (2023; http://arxiv.org/abs/1509.02529) (Apr. 2016).

152. H. Teimoorinia *et al.*, *Monthly Notices of the Royal Astronomical Society* **457**, ADS Bibcode: 2016MNRAS.457.2086T, 2086–2106, ISSN: 0035-8711, (2023; https://ui.adsabs.harvard.edu/abs/2016MNRAS.457.2086T) (Apr. 2016).

153. D. Thomas *et al.*, *Monthly Notices of the Royal Astronomical Society* **412**, ADS Bibcode: 2011MNRAS.412.2183T, 2183–2198, ISSN: 0035-8711, (2023; https://ui.adsabs.harvard.edu/abs/2011MNRAS.412.2183T) (Apr. 2011).

154. J. L. Tinker *et al.*, en, *The Astrophysical Journal* **686**, Publisher: IOP Publishing, 53, ISSN: 0004-637X, (2023; https://iopscience.iop.org/article/10.1086/589983/meta) (Oct. 2008).

155. C. A. Tremonti *et al.*, *The Astrophysical Journal* **613**, ADS Bibcode: 2004ApJ...613..898T, 898–913, ISSN: 0004-637X, (2023; https://ui.adsabs.harvard.edu/abs/2004ApJ...613..898T) (Oct. 2004).

156. M. Trevisan *et al.*, *Monthly Notices of the Royal Astronomical Society* **464**, 4593–4610, ISSN: 0035-8711, (2022; https://doi.org/10.1093/mnras/stw2588) (Feb. 2017).

157. M. Trevisan *et al.*, *Monthly Notices of the Royal Astronomical Society: Letters* **471**, arXiv:1706.02463 [astro-ph], L47–L51, ISSN: 1745-3925, 1745-3933, (2023; http://arxiv.org/abs/1706.02463) (Oct. 2017).

158. M. S. Turner *et al.*, *Physical Review Letters* **52**, Publisher: American Physical Society, 2090–2093, (2023; https://link.aps.org/doi/10.1103/PhysRevLett.52.2090) (June 1984).

159. P. G. van Dokkum *et al.*, *The Astrophysical Journal* **813**, ADS Bibcode: 2015ApJ...813...23V, 23, ISSN: 0004-637X, (2023; https://ui.adsabs.harvard.edu/abs/2015ApJ...813...23V) (Nov. 2015).

160. J. van de Sande *et al.*, *Nature Astronomy* **2**, ADS Bibcode: 2018NatAs...2..483V, 483–488, ISSN: 2397-3366, (2023; https://ui.adsabs.harvard.edu/abs/2018NatAs...2..483V) (Apr. 2018).

161. J. van de Sande *et al.*, *Monthly Notices of the Royal Astronomical Society* **472**, 1272–1285, ISSN: 0035-8711, (2023; https://doi.org/10.1093/mnras/stx1751) (Dec. 2017).

162. G. M. Voit, *Reviews of Modern Physics* **77**, arXiv:astro-ph/0410173, 207–258, ISSN: 0034-6861, 1539-0756, (2023; http://arxiv.org/abs/astro-ph/0410173) (Apr. 2005).

163. A. von der Linden *et al.*, *Monthly Notices of the Royal Astronomical Society* **404**, 1231–1246, ISSN: 0035-8711, (2022; https://doi.org/10.1111/j.1365-2966.2010.16375.x) (May 2010).

164. S. M. Weinmann *et al.*, *Monthly Notices of the Royal Astronomical Society* **372**, ADS Bibcode: 2006MNRAS.372.1161W, 1161–1174, ISSN: 0035-8711, (2023; https://ui.adsabs.harvard.edu/abs/2006MNRAS.372.1161W) (Nov. 2006).

165. N. Werner *et al.*, *Monthly Notices of the Royal Astronomical Society* **439**, 2291–2306, ISSN: 0035-8711, (2023; https://doi.org/10.1093/mnras/stu006) (Apr. 2014).

166. A. R. Wetzel *et al.*, *Monthly Notices of the Royal Astronomical Society* **424**, 232–243, ISSN: 0035-8711, (2023; https://doi.org/10.1111/j.1365-2966.2012.21188.x) (July 2012).

167. S. D. M. White, M. J. Rees, *Monthly Notices of the Royal Astronomical Society* **183**, ADS Bibcode: 1978MNRAS.183..341W, 341–358, ISSN: 0035-8711, (2023; https://ui.adsabs.harvard.edu/abs/1978MNRAS.183..341W) (May 1978).

168. J. Woo *et al.*, *Monthly Notices of the Royal Astronomical Society* **428**, 3306–3326, ISSN: 0035-8711, 1365-2966, (2023; http://arxiv.org/abs/1203.1625) (Feb. 2013).

169. J. Woo *et al.*, *Monthly Notices of the Royal Astronomical Society* **464**, ADS Bibcode: 2017MNRAS.464.1077W, 1077–1094, ISSN: 0035-8711, (2023; https://ui.adsabs.harvard.edu/abs/2017MNRAS.464.1077W) (Jan. 2017).

170. J. Woo *et al.*, *Monthly Notices of the Royal Astronomical Society* **448**, ADS Bibcode: 2015MNRAS.448..237W, 237–251, ISSN: 0035-8711, (2023; https://ui.adsabs.harvard.edu/abs/2015MNRAS.448..237W) (Mar. 2015).

171. S. Wuyts *et al.*, en, *The Astrophysical Journal* **742**, Publisher: The American Astronomical Society, 96, ISSN: 0004-637X, (2023; https://dx.doi.org/10.1088/0004-637X/742/2/96) (Nov. 2011).

172. X. Yang *et al.*, *The Astrophysical Journal* **671**, ADS Bibcode: 2007ApJ...671..153Y, 153–170, ISSN: 0004-637X, (2022; https://ui.adsabs.harvard.edu/abs/2007ApJ...671..153Y) (Dec. 2007).

173. F. Yuan *et al.*, *The Astrophysical Journal* **857**, ADS Bibcode: 2018ApJ...857..121Y, 121, ISSN: 0004-637X, (2023; https://ui.adsabs.harvard.edu/abs/2018ApJ...857..121Y) (Apr. 2018).

174. G. U. Yule, *Journal of the Royal Statistical Society* **75**, Publisher: [Wiley, Royal Statistical Society], 579–652, ISSN: 0952-8385, (2023; https://www.jstor.org/stable/2340126) (1912).

175. A. I. Zabludoff, J. S. Mulchaey, *The Astrophysical Journal* **496**, ADS Bibcode: 1998ApJ...496...39Z, 39–72, ISSN: 0004-637X, (2023; https://ui.adsabs.harvard.edu/abs/1998ApJ...496...39Z) (Mar. 1998).

176. I. Zehavi *et al.*, *The Astrophysical Journal* **630**, ADS Bibcode: 2005ApJ...630....1Z, 1–27, ISSN: 0004-637X, (2023; https://ui.adsabs.harvard.edu/abs/2005ApJ...630....1Z) (Sept. 2005).

177. Z. Zhang, *Annals of Translational Medicine* **4**, III, ISSN: 2305-5839, (2023; https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4828741/) (Mar. 2016).