Statistical Shape Analysis of Galactic H II Regions

by

Justyn Campbell-White

Thesis

Submitted to the University of Kent

for the degree of

Doctor of Philosophy

School of Physical Sciences

April 2019

Abstract

H II regions are diffuse nebulae of ionised hydrogen, excited by the extreme ultraviolet emission from massive stars. Due to the embedded nature of massive star formation, there are many observational difficulties involved when investigating such stars. H II regions, however, are readily observed via their infrared and radio emission. As such, they highlight the location of their massive star sources. Furthermore, H II region properties are directly resultant of their progenitors and environment. The overall aim of the work presented herein, is to determine whether statistical shape analysis of observational and numerically modelled H II region data can be used to probe the associated astrophysical properties.

Radio continuum and computer simulated synthetic images of H II regions were analysed using the shape extraction and statistical comparison methods constructed in this work. For the radio data, six morphological groups were identified. Visual inspection and quantitative ordinance techniques confirmed that the shape analysis and grouping procedure were working as intended. It was found that in the first Galactic quadrant, location is mostly independent of group, with a small preference for regions of similar Galactic longitudes to share common morphologies. The shapes are homogeneously distributed across Galactocentric distance and latitude. One group contained regions that are all younger than 0.5 Myr and ionised by relatively low- to intermediate-mass sources. Those in another group are all driven by intermediate- to high-mass sources. One group was distinctly separated from the other five and contained regions at the surface brightness detection limit for the survey. The hierarchical procedure employed was most sensitive to the spatial sampling resolution used, which is determined for each region from its heliocentric distance.

The numerical H II region data was the result of photoionisation and feedback of a 34 $M_\odot$ star, in a 1000 $M_\odot$ cloud. Synthetic observations (SOs) were provided, comprising four evolutionary snapshots (0.1, 0.2, 0.4 and 0.6 Myr), and multiple viewing projection angles. The shape analysis results provided conclusive evidence of the efficacy of the numerical simulations. When comparing the shapes of the synthetic regions to their observational counterparts, the SOs were grouped in amongst the Galactic H II regions by the hierarchical procedure. There was also an association between the evolutionary distribution of regions of the respective samples. This suggested that this method could be further developed for classification of the observational regions by using the synthetic data, with its well defined parameters.

Declarations

The content herein was composed by the author, and has not been submitted for the purposes of a qualification at any other institution or for any other degree.

The content comprising Chapters 3 & 4 was adapted and extended from work that has been published as Campbell-White et al. (2018).

All data is the author's own, unless explicitly stated otherwise. All instances where use has been made of other work has been cited.

Acknowledgments

Firstly, I would like to thank my supervisors, *Dr Dirk Froebrich* and *Dr Alfred Kume.* The enthusiasm you both have for your respective areas of research meant that I was keen to do this project justice. You both provided inspiration, support, and guidance throughout. Dirk also taught me how to apply all of this to the practical aspect of astronomy, using the Beacon Observatory. I am sure that I will continue to use these skills and develop my astro-statistics knowledge in my future career. I would also like to acknowledge *the School of Physical Sciences* and *the School of Mathematics, Statistics and Actuarial Sciences* for providing the funding for this project; and *Dr Ahmad Ali* for providing the numerical data used in this research.

Further thanks go to *Dr Tim Kinnear*, who is one of the most eagerly helpful people I have ever met. My 'Ph.D Sister', *Dr Sally Makin*; you joked that if you could do this, I definitely could. However, the truth is, you set the bar extremely high, which gave me something to aim towards. Your help throughout this project has been amazing. I would also like to thank *(soon-to-be-Dr) Sam Billington*, for our ongoing useful discussions, and for constantly reminding me that he is better at Python than me. That was the real reason that I had to learn R! Further thanks go to *Dr Michael Woods* for taking care of all administrative aspects of this project, and for making me laugh by saying that the Bills were going to beat the Patriots.

I would also like to thank *Dr James Urquhart.* James taught me everything I know about radio astronomy and was a superb away-day advisor during conferences in Heidelberg and Cardiff. And to set the record straight, that whole Airbnb

situation was Sam's fault!

I don't think I would have made it to this stage without *Lynsey*. You wrote in your acknowledgements that you hoped you could provide the same support for me when it was my turn. You definitely went above and beyond this hope, providing me with tremendous amounts of support, reassurance and insightful guidance. You are truly an inspiration and I am extremely grateful. I would also like to thank Lynsey's family, *Jan, Ash* and *Ashely* for making me feel like a part of it, and for 20 years of holidays. Jan - yes, it's done now! Ash - here's to another 20 years! And Ashely - when are you going home?

Last and by no means least, thanks to my Mum, *Carole*, and her partner, *Ron*. You both always believed that I was capable of achieving what ever I set out to do; with heartfelt encouragement and support throughout my education. Especially Mum, you have taken this journey with me. From your love of Star Wars and Star Trek to Discovery Science documentaries. I appreciate you taking the time to learn so much about what I'm doing, even what a parsec is.

P.S. Thanks for not naming me Jean-Luc!

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 HII Regions, a Brief History

Extended regions of ionised hydrogen (HII) emission were first observed in the night sky by Struve and Elvey (1938), using the nebula spectrograph equipped to the McDonald Observatory in Texas, USA. This type of nebulosity had not before been observed around bright field stars. Since the occurrence ceased abruptly at high Galactic latitudes, they suggested that the source of the ionisation was the integrated ultraviolet (UV) emission from the most massive of main sequence stars. Such extreme UV radiation from massive stars can photoionise the surrounding hydrogen cloud, from which they were born. This is the process of removing the electron from the nucleus, resulting in a cloud of ionised plasma. This type of nebula came to be known as an HII region. The analytical concept of an HII region was then described by Strömgren (1939). For stars with luminosities of sufficient emission energy to ionise hydrogen, the size of the diffuse HII region nebula can be approximated from the stellar ionising flux. The initial expansion of the HII region takes approximately 10^5 years. Typical sizes of the final region are in the order of tens of parsecs, however, much smaller compact and ultra-compact HII regions (smaller than 0.5 and 0.01 parsecs, respectively) are also found in the Galaxy (Wood and Churchwell, 1989).

Massive stars capable of exciting an HII region have lifetimes up to ~ 10 million years. Whilst this may seem like an eternity to us, this is a relatively short astronomical length of time. By contrast, our Sun has a lifetime of ~ 10 billion years, and it was ~ 2 million years ago that the human genus first emerged on Earth. Throughout the lifetime of massive stars, a significant amount of feedback is delivered to the surrounding interstellar medium (ISM), through tremendous processes such as the HII regions, outflows, stellar winds, and ultimately supernovae explosions. Supernovae are also the predominant source of heavy elements in the Cosmos. Massive stars are hence responsible for the enrichment of the ISM, allowing for complex organisms such as us to exist. The importance of studying massive stars and their influences is therefore key in understanding the evolution of our Galaxy. The formation of massive stars is a contentious subject in astrophysics. This is due to the fact that they are difficult to observe, forming deeply embedded within molecular clouds. HII regions provide an indicator of where in the Galaxy massive star formation is happening. Due to the clustered nature of massive star formation, insight into the physical properties of HII regions could allow for inferences to be made as to the initial conditions of massive cluster formation.

HII regions are located in the Galactic disk, with the majority at great distances from our Solar system. This leads to their optical and UV emission being obscured by the intervening interstellar material. However, infrared (IR) and radio wavelength emission from the HII regions are able to reach our detectors. Although astronomy is regarded as the oldest of the sciences, with its application for prediction dating back to ancient civilisations, multi-wavelength astronomy is a relatively young field of study. Development in hardware and imaging techniques throughout the 20th and 21st century thus allowed for high resolution imaging of HII regions to be carried out at IR and radio wavelengths. One source of the IR emission from HII regions is interstellar dust, which re-radiates optical and UV emission in the IR range. Additionally, IR emission from molecular material on the edge of HII regions brilliantly highlights the interaction of the ionised regions with their surroundings. Radio continuum emission from HII regions is the result of free accelerating electrons, within the ionised plasma, interacting with the electric field of the ions. Each of the concepts introduced here are explained in much greater detail in the subsequent chapter.

The work presented in this Thesis takes advantage of the homogeneity of high resolution radio images of HII regions, to statistically compare their shapes. Statistical shape analysis is the analysis of the geometrical properties of a given set of shapes, by quantitative methods. Mathematical shape provides an unbiased characteristic of an object, which can be readily compared through various statistical means. A successful application of shape analysis to astronomical data is that of the Galaxy Zoo project (Lintott et al., 2008), which has led to the morphological classification of hundreds of thousands of galaxies. The overarching aim of the work presented herein, is to investigate possible associations between HII region shapes (from both observational and simulated data) and the physical parameters/initial conditions of massive star cluster formation.

1.2 Thesis Aims & Structure

In order to determine whether shape can be considered an intrinsic property of observational HII region data, the Thesis aims and structure are as follows:

Theory, Observations & Simulations (Chapter 2)

> This chapter covers the background theory pertaining to HII regions, their massive star progenitors, and the interstellar medium in which they are found. Following this, an overview of Galactic Plane surveys is provided. The MAGPIS observational data, from which the HII region radio continuum data was taken, is detailed here. Finally, numerical simulations of HII regions are introduced, covering the physical processes modelled in order to simulate the environment, massive stars, and resulting nebula. This is with the aim of using the shape analysis method developed here to quantitatively test the efficacy of the resulting synthetic observations.

Shape Analysis & Statistical Methods (Chapter 3)

> This chapter introduces the concept of shape from both an astronomical and mathematical perspective. Full details of how the HII region shapes were systematically extracted and quantified are explained here. Also introduced and explained are the various statistical measures that are employed in this work, in conjunction with the shape analysis. The aim of this is to decide upon a suitable grouping measure for the comparison of the shape data.

Radio Continuum Observations (Chapter 4)

> This chapter details the application of the shape analysis and statistical clustering for a selection of Galactic HII regions from the MAGPIS radio continuum data. It is investigated here whether physical properties of the HII regions

and/or their ionising source(s), such as age and mass, correlate with the identified shape and shape groupings.

Synthetic Observations (Chapter 5)

This chapter details the application of the shape analysis methodology to a selection of HII regions from numerical simulations. The shapes of these synthetic HII regions are then compared to the MAGPIS sample, using the shape analysis method. This is to test the efficacy of the simulations, i.e. - determine if they are producing results that are representative of their observational counterparts. Following this, is an investigation into how the initial conditions and physical parameters of the simulation influence the resulting HII region shape. The ultimate aim here is to determine whether the numerical simulations can be used to construct a training set of HII region shapes, for use in a morphological supervised classification scheme.

Summary & Conclusions (Chapter 6)

This chapter includes a detailed summary of the Thesis and overall conclusions from the analysis and results. Future work and applications are also outlined.

Chapter 2

Theory, Observations & Simulations

This chapter provides the background theory relating to HII regions, massive stars and the interstellar medium. An overview of Galactic Plane surveys that feature HII regions is then provided. The radio continuum MAGPIS observational data is detailed within. The final part of this chapter introduces numerical simulations of HII regions.

2.1 The Interstellar Medium

Before establishing the theory of HII regions, the places in which their host stars form and evolve must first be understood. The interstellar medium (ISM) is the name given to all of the material between stars within the Galaxy. It is comprised of ∼70% hydrogen, ∼28% helium, with the remaining ∼2% constituting heavier elements[1] (Spitzer, 1978). The local ISM densities and temperatures are highly variable, depending on the constituents and environment. In fact, the vast majority of the ISM within the Galaxy can be considered empty space, with particle number

[1] Which are collectively deemed as 'metals' by astrophysicists.

densities as low as $1\,\mathrm{cm}^{-3}$. However, the ISM matter accounts for $\sim 10 - 15\%$ of the total mass of the Galactic disk (Mihalas and Binney, 1981). This means that interstellar mass is concentrated in dense regions. These are known as clouds. Clouds occupy only $\sim 1 - 2\%$ of the interstellar volume, yet account for approximately half of the interstellar mass (Shu et al., 1987).

From Ferrière (2001), there are three main types of clouds. The most abundant are molecular clouds, which are essentially made of extremely cold ($\mathrm{T} \sim 10 - 20\,\mathrm{K}$) molecular gas; the majority of which is molecular hydrogen (H_2). A large range of densities ($n \sim 10^2 - 10^6\,\mathrm{cm}^{-3}$) have been observed in molecular clouds. Sources of this variance are turbulence and the different feedback mechanisms from stars (Larson, 2003). It is well accepted that molecular clouds are the main birth place of stars in the Galaxy (Blitz, 1993); which we will return to later in this section. Next are diffuse clouds that consist of cold ($\mathrm{T} \sim 100\,\mathrm{K}$) atomic gas, the majority being neutral hydrogen (HI). These are at significantly lower densities than molecular clouds ($n \sim 20 - 50\,\mathrm{cm}^{-3}$). The third type are translucent clouds, which are similar to diffuse clouds but contain a mixture of atomic and molecular gas. The rest of the interstellar material, spread out between the clouds and around stars and star clusters, exist in three different forms: warm atomic, warm ionised and hot ionised (where warm in an astronomical sense is a temperature of $\sim 10^4\,\mathrm{K}$, and hot is $\sim 10^6\,\mathrm{K}$). Observations of different aspects of the ISM, along with stellar observations, yield insight to the structure of our Galaxy.

2.1.1 Galactic Structure

A collection of stars and the ISM, bound by gravity, is what defines a galaxy. Since we cannot view our Galaxy from the outside-in, our understanding of the Milky Way structure has come from observations of its contents. We now know that the Galaxy comprises a thin disk, with a radius of $\sim 25 - 30\,\mathrm{kpc}$ and an effective thickness of $\sim 400 - 600\,\mathrm{pc}$[2]. The inner region of the Galaxy is composed of a bulge with a radius of $\sim 3\,\mathrm{kpc}$, which hosts the supermassive black hole, Sagittarius A*. A Galactic halo then extends out to more than 30 kpc. The Sun is located within the Galactic disk, 12.5 pc above the midplane (with respect to the midplane itself, which was recently re-evaluated from HII region observations, Anderson et al., 2019); and 8.34 ± 0.16 kpc away from the centre of the Galaxy, around which it rotates with a velocity of $240 \pm 8\,\mathrm{km\,s}^{-1}$ (Reid et al., 2014). The stars in the Galactic disk

[2] pc is short for parsec, a common unit for astronomical distances. A parsec is defined as the distance at which one astronomical unit (AU, the average distance between the Earth and the Sun) subtends an angle of one arcsecond (″). One parsec is equal to 3.26 light years, or 3.0857×10^{16} m.

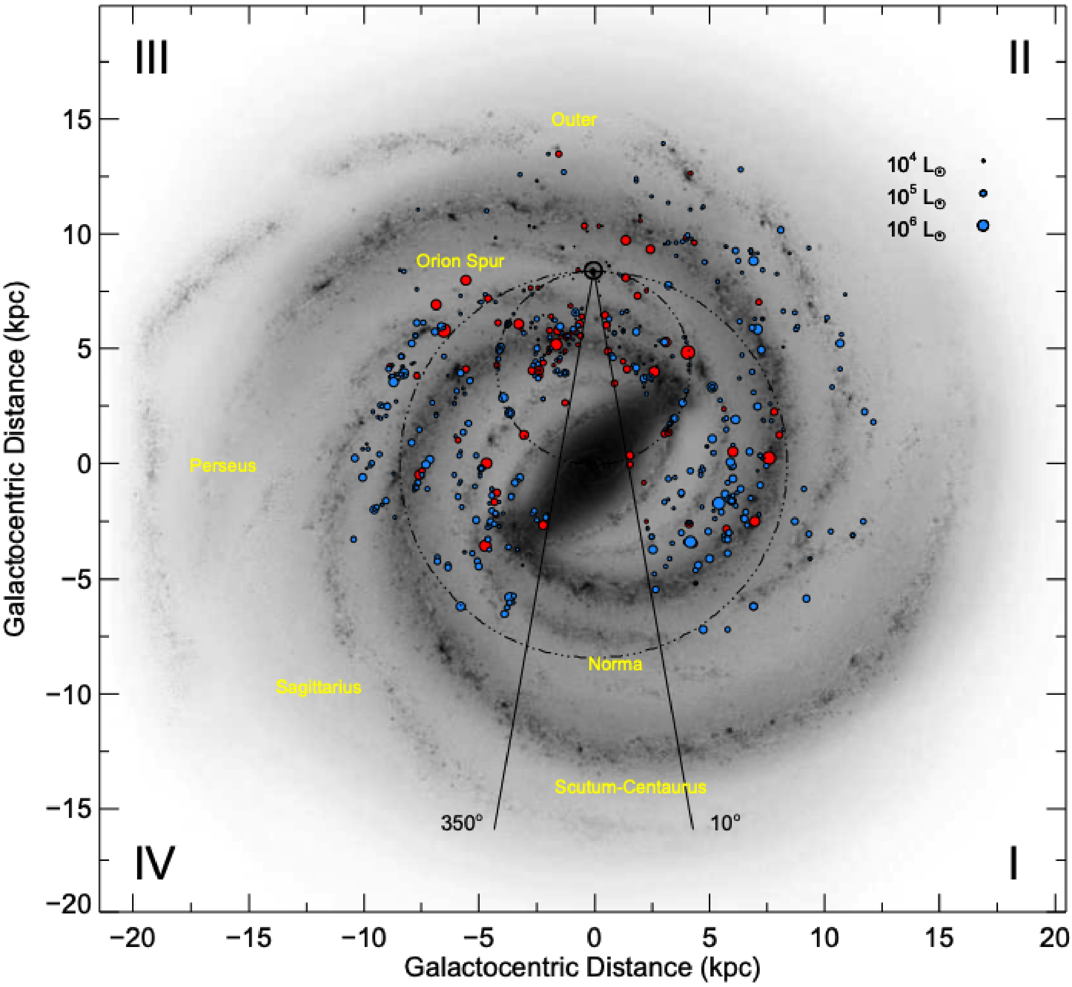


Figure 2.1: Top down sketch of the Milky Way. Figure 6 from Urquhart et al. (2014). Galactic distribution of all massive young stellar objects and HII regions from the RMS survey with bolometric luminosities greater than $10^4\,L_\odot$. Complexes and individual sources are shown by the red and blue points, respectively. The background image is a sketch of the Galaxy produced by R. Hurt of the Spitzer Science Center in consultation with R. Benjamin.

have nearly circular orbits, with angular orbital rates that decrease with increasing radial distance. Radio wavelength observations of the neutral hydrogen within the ISM revealed that the Milky Way possesses a spiral structure (Kerr, 1969), similar to that seen in numerous external galaxies. Local optical observations provide an accurate outline of the three closest spiral arms to the Sun, placing us between the inner Sagittarius arm and the outer Perseus arm, near the inner edge of the local Orion-Cygnus arm (Mihalas and Binney, 1981).

Figure 2.1, which is from Urquhart et al. (2014), shows a sketch of what our Galaxy is thought to look like from above. The position of the Sun is shown by a small circle above the Galactic centre. The larger of the dashed circles shows the Solar circle, which is the Sun's orbit around the centre of the Galaxy. Spiral arms and Galactic quadrants are labelled. The positions of the spiral arms were mapped out by observations of HII regions by Georgelin and Georgelin (1976). As we will see later in this chapter, HII regions trace active star formation. Included in Fig. 2.1 are the positions of all massive young stellar objects (MYSOs) and HII regions from the Red MSX Source (RMS) survey (Hoare et al., 2005) that have bolometric luminosities greater than $10^4\,L_\odot$. We can see that there is indeed a correlation between the massive stars and the spiral arms, with a stronger preference for the complexes (red dots) than for the individual sources (blue dots). The average scale height perpendicular to the Galactic disk of the MYSOs and HII regions for the inner part of the Galaxy is $\sim 25\,$pc. The observational data considered in this work hence focuses on those from Galactic Plane surveys, and will be detailed later in this chapter.

2.1.2 Giant Molecular Clouds

With this understanding of how our Galaxy is structured, we can return to the regions within the ISM that star formation takes place, Giant molecular clouds (GMCs). GMCs have masses up to $10^6\,M_\odot$, with sizes ranging from several parsecs to several tens of parsecs. Typical densities of GMCs are between $\sim 1 - 3 \times 10^2\,\mathrm{cm}^{-3}$, with approximately constant temperatures of $\sim 10\,$K (Goldsmith, 1987). At $\sim 10\,$K, H_2 does not emit radiation. Therefore, other molecules located within the GMCs are used to probe the structure of GMCs. Isotopes of carbon monoxide (CO) are regularly used as such a tracer (e.g. Roman-Duval et al., 2010). CO observations have revealed that GMCs have a clumpy, inhomogeneous and filamentary structure. These molecular clumps have masses between $10^3 - 10^4\,M_\odot$, sizes between $2 - 5\,$pc and particle densities between $10^2 - 10^3\,\mathrm{cm}^{-3}$ (Shu et al., 1987). Within molecular clumps there are then regions of localised high density known as molecular cores.

Cores possess the following properties: n $\sim 10^6 - 10^9\,\mathrm{cm}^{-3}$; T $\sim 10 - 200\,\mathrm{K}$; M $\sim 10 - 1000\,\mathrm{M}_\odot$; and sizes of $\sim 0.3 - 1\,\mathrm{pc}$ (Bergin and Tafalla, 2007). It is the collapse of such cores that can theoretically lead to the formation of stars (Shu, 1977).

Collapse of dense cores within GMCs, under the influence of purely gravity would be expected to occur on timescales of the order of the free-fall time, t_{ff} (Spitzer, 1978). This is the time taken for the cloud to fall into a point as a result of the gravity of the cloud, excluding hydrodynamical effects. The gravitational potential at a point located within a cloud for a given time is a function of the radial position from the centre of mass of that point, and the total mass enclosed by a sphere. Integrating the equations of motion for the gas gives:

$$t_{ff} = \sqrt{\frac{3\pi}{32G\rho}} \tag{2.1}$$

where $\rho = M/\left(\frac{4}{3}\pi R^3\right)$ is the initial mean density of the cloud, with mass, M, and radius, R; G is the Gravitational constant. Whilst this gives an estimate for the required timescale for collapse, as stated, this does not account for any thermodynamical changes undergone by the core during its collapse. Nor does it account for any support mechanisms within the GMC. Blitz and Shu (1980) showed that molecular clouds have characteristic timescales up to 30 times longer than t_{ff}. Furthermore, if this cloud collapse occurred uninhibited, the Galactic star formation rate (SFR) would be much larger than the observed rate of $\sim 3\ \mathrm{M}_\odot \mathrm{yr}^{-1}$ (McKee and Ostriker, 2007).

The reason that molecular clouds survive for periods longer than their free-fall times is due to the internal support mechanisms that allow the cloud to resist the collapse. The first of these support mechanisms is thermal pressure within the clouds, arising from the kinetic temperatures of the gas. Whilst this is low on a global scale, local increases in pressure due to feedback from massive stars can contribute to the thermal pressure support. Magnetic fields are another factor, however, these could only support against collapse in the direction perpendicular to the magnetic field lines. The most significant contributor to prevent cloud collapse is thought to be turbulence. Turbulence within a molecular cloud is the internal velocity dispersion of the molecular material within the clouds. Observations of GMCs at different spatial scales reveal filamentary structure across various size scales, showing that turbulence is present locally and globally, with different stellar feedback mechanisms suggested as the driver of turbulence at these varying scales (André et al., 2014).

Figure 2.2 shows a colour composite image of the Orion Molecular Cloud (OMC) complex. This a relatively nearby region of active star formation, featuring MYSOs,

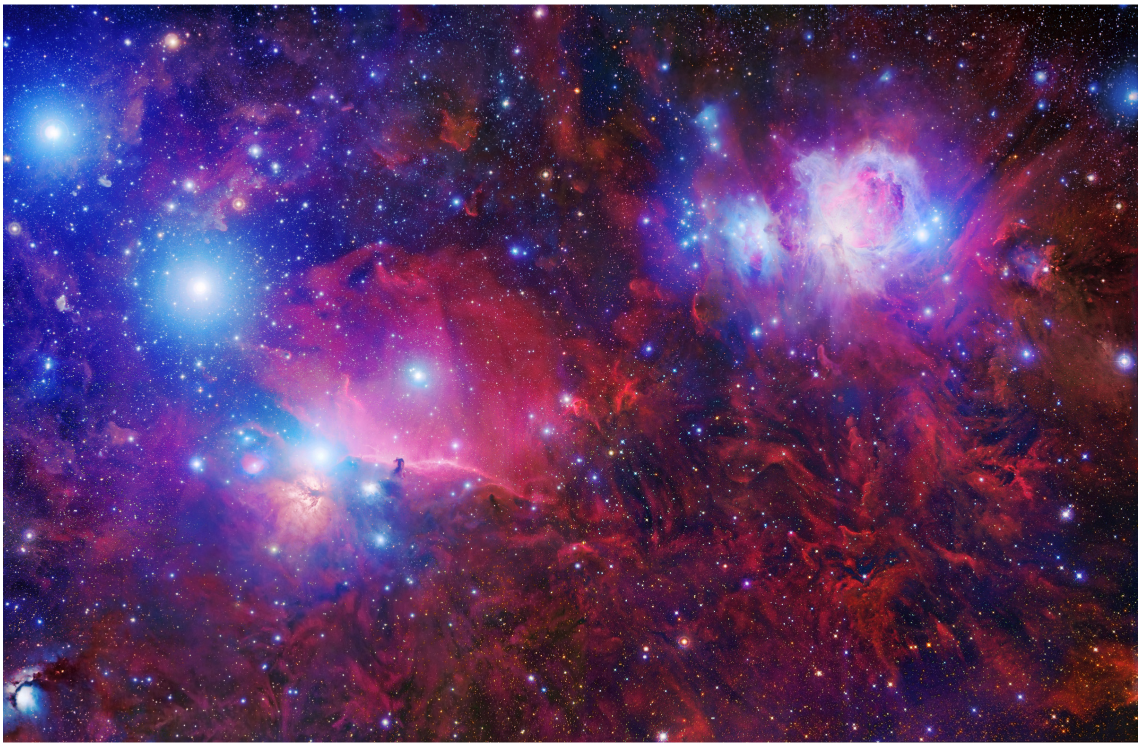


Figure 2.2: Image of the Orion molecular complex. The red channel comprises Hα observations, highlighting the dense gas. The green and blue channels are broadband visual of the respective colours. The brightest three stars on the left are the Orion's Belt stars. The Orion Nebula, M42, is seen in the top right. Image credit: Digitized Sky Survey, R. Gendler, R. Colombari & F. Pelliccia

Hii regions, open star clusters, dense clumps and cores, and large scale filamentary structure within the GMC. The three brightest stars along the left side of the image are the Belt stars. To the left and right of the bottom most Belt star, in this orientation, is the Flame Nebula Hii region and the Horsehead nebula, respectively. The Orion Nebula, M42, can be seen as the extended ring in the top right. Although the structure of the OMC complex can be seen in deep optical images, since it is nearby, other GMCs in the Galaxy at farther distances are masked by interstellar dust grains. However, such dust grains are also associated with the GMCs themselves, and can provide a way of observing the molecular clouds.

The description of the ISM thus far has considered only gasses. However, about 0.5 - 1% of the interstellar matter is in the form of solid dust grains. The dust comprises mainly graphites, silicates and amorphous carbons, which can chain together to form composite grains, with sizes ranging between 0.01 and 0.25 μm (Mathis and Whiffen, 1989). This dust can scatter and absorb optical and UV radiation (with wavelengths of similar size to the grains), which is referred to as interstellar reddening and extinction. The amount of extinction is wavelength dependent, with the dust becoming more and more transparent towards longer wavelengths. Thus, the issue of extinction by dust grains is overcome by using radio or sub-millimetre observations. Similarly, the amount of reddening is also wavelength dependent, blue light (shorter wavelength, λ) is scattered more than red light (longer λ), thus the background stars appear more red than their actual intrinsic colour. For stellar observations, the difference in magnitudes between two colours is known as the colour index. The amount of interstellar reddening can thus be calculated by determining the colour excess, which is the difference between the colour index of a star as it is observed, and the colour index of an idealised object of the same spectral type.

As previously mentioned, there is a strong correlation of dust within molecular clouds, with about 1% of the mass fraction of GMCs belonging to the dust (Schlegel et al., 1998). It is thought that the dust grains can actually enhance the H_2 density, far beyond that expected from random HI collisions (Savage et al., 1977; Hasegawa and Herbst, 1993). The dust provides a site for the HI to collect, then also absorbs the binding energy given off when the stable molecule forms. This energy goes into heating the dust and releasing the newly formed H_2. This occurs not only for hydrogen, but also heavier molecules such as CO, which as described earlier, enable us to probe the molecular regions. Furthermore, the dust grains themselves actually highlight the locations of molecular gas via 'thermal dust emission'. The dust can absorb stellar photons, which heats the grains to a temperature of $\sim 20\,\mathrm{K}$ (Dwek et al., 1997). Subsequent cooling of the dust grains is via re-emission of a cascade of photons at longer wavelengths than those originally absorbed, almost exclusively in the infrared. Since the breakthrough of space-based IR observatories in the early 1980's, it has been possible to observe the thermal dust emission and use it as a diagnostic tool to study the properties of interstellar dust and, by extension, molecular clouds.

2.2 Massive Stars

The final topic to cover before considering the theory of HII regions is that of their source: massive stars. The term 'massive' is given to any star $> 8\,\mathrm{M}_\odot$. These correspond to the O and B spectral class of stars (commonly referred to as OB stars, Fig. 2.3). The main difference between massive stars and low-mass stars is their luminosities (which is plotted on a log scale on the right hand side of Fig. 2.3). The luminosity of a star increases rapidly with mass, following $L = M^\alpha$ (Kuiper, 1938),

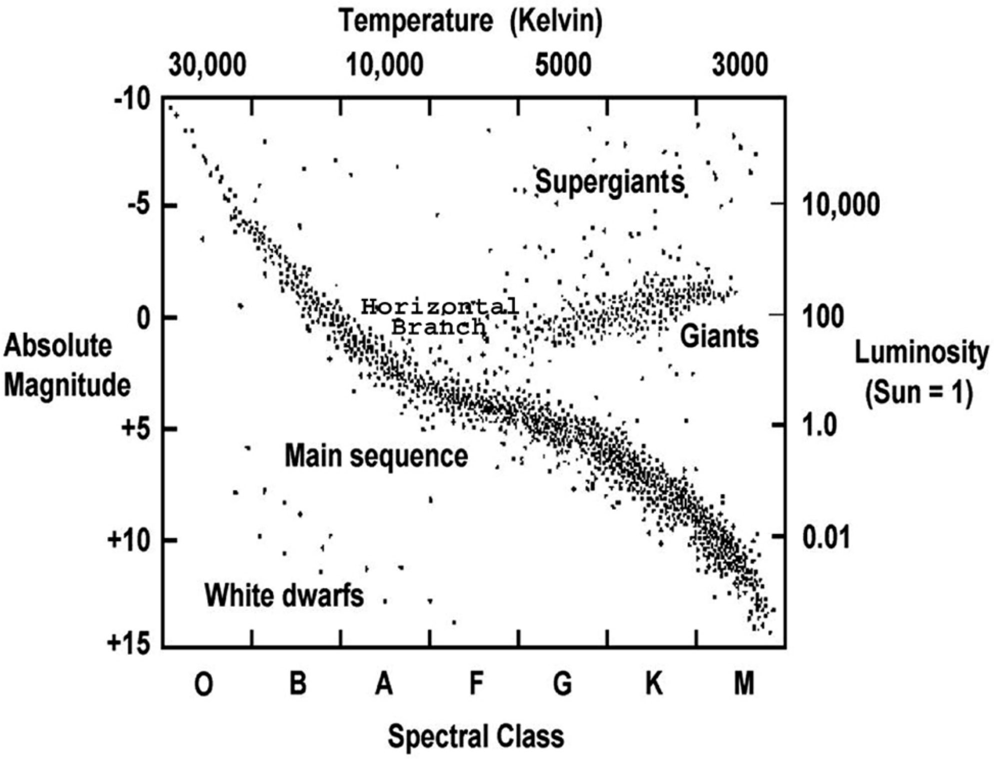


Figure 2.3: Hertzsprung-Russell (H-R) diagram of stellar populations. Temperature and spectral class are plotted against luminosity and absolute magnitude. Image source: http://chandra.harvard.edu/graphics/edu/formal/variable_stars/HR_diagram.jpg

with α varying for the respective stellar mass range. For massive stars, $\alpha = 2.76$ (Vitrichenko et al., 2007). With great luminosity comes great power. The photon energy distribution is dependent on the surface temperature of the star, which is directly proportional to the luminosity. Therefore, stars with high luminosities emit high energy photons. Photons with energy greater than $h\nu = 13.6\,\mathrm{eV}$ ($\lambda \leqslant 91.2\,\mathrm{nm}$) are able to ionise hydrogen. Approximately $8\,\mathrm{M}_\odot$ is the minimum mass for a star to have a high enough luminosity to ionise hydrogen, creating an HII region.

The more massive the star, the faster it uses its fuel supply. Massive stars therefore have a short main sequence lifetime, of the order of a few million years (Woosley et al., 2002). Massive stars are also comparatively rare throughout the Galaxy. They comprise $< 1\%$ of all stars. Whilst this is a small fraction, it is still a large number, considering the hundreds of billions of stars in the Galaxy. This fraction follows from the stellar initial mass function (IMF) (Salpeter, 1955; Kroupa, 2001). The IMF is an empirically derived relationship that describes the initial distribution of masses for a population of stars. It is widely accepted that stars form in clusters, and that the bulk of Galactic field stars result from dissolving and dispersed star clusters (Lada and Lada, 2003). A relationship is also observed between the most massive star within a star cluster and the mass of the cluster (Weidner et al., 2010).

The combination of being rarer, shorter lived and located at greater distances than low-mass stars all contribute to the issue that high-mass star formation is less well understood than the formation of low-mass stars. Another contributor to this is the fact that massive young stellar objects (MYSOs) are more heavily embedded and obscured by the GMCs within which they form (Wynn-Williams, 1982). For massive star formation, the time required to reach thermal equilibrium, known as the Kelvin-Helmholtz timescale, $t_{KH} = (GM_*^2/R_*L_*)$ is much faster than for low-mass stars. However, the infall time for collapse for both is similar. This means that high-mass stars continue to accrete mass from the free-falling stellar envelope after reaching the main sequence (e.g. Haemmerlé et al., 2016).

An issue arises here, such that once the massive star reaches the main sequence and begins to fuse hydrogen, the luminosity of the star becomes sufficiently large as to prevent further accretion (Eddington, 1916; Larson and Starrfield, 1971). It is therefore accepted that massive star formation requires a different manner of collapse to that of low-mass star formation. Bonnell et al. (1998) produced a model of high-mass star formation whereby low- and intermediate-mass stars in densely populated star clusters could coalesce to form high-mass stars. McKee and Tan (2002) suggest that cores within a large cluster are highly turbulent, allowing for higher accretion rates. Zinnecker and Yorke (2007) suggest a hierarchical process. Kuiper et al. (2010) investigated the effects of self-shielding by the accretion disk, in order to allow for higher accretion rates. The issue of high-mass star formation remains a popular issue of debate amongst the community. A thorough review of different formation mechanisms, theoretical models and corresponding observational evidence is given in Tan et al. (2014).

In addition to being a significant contributor to the support mechanisms against collapse of molecular clouds (lowering the SFR), massive stars drive substantial amounts of energy into the surrounding ISM. This is via continuous streams of mass loss, known as stellar winds (Morton, 1967; Conti and Leep, 1974), the ionisation front propagating from the HII region, and ultimately supernovae explosions. Each of these mechanisms can potentially lead to cases of triggered star formation. This is the process by which some external factor directly instigates the formation of stars. Stellar winds can lead to compression of the ISM, increasing the local density, causing fragmentation of clumps (Foster and Boss, 1996). Supernovae can contribute either directly or as large source of turbulence (Cameron and Truran, 1977; Elmegreen, 2002). This has even been suggested as a possible formation mechanism for our

Solar system (Cameron and Truran, 1977; Gritschneder et al., 2012).

In the context of this work, HII regions have been extensively studied as the source of triggered star formation, via the concepts of radiative driven implosion (RDI) (Reipurth, 1983; Bertoldi, 1989) and collect & collapse (C&C) (Deharveng et al., 2005; Zavagno et al., 2006; Dale et al., 2007a). In RDI, the formation and subsequent evolution of an HII region results in the expansion of the ionisation front (which we will cover in the next section) up to the surfaces of pre-existing clumps within the GMC. For C&C, the triggering is a result of fragmentation of a dense shell that has been spherically swept up via the expanding ionisation front and/or stellar wind. The balance and significance of each mechanism is thought to be determined by the anisotropy of the surrounding ISM into which the HII region expands (Pomarès et al., 2009; Walch et al., 2012). Thompson et al. (2012) carried out a detailed statistical study into triggered star formation around HII regions and stellar bubbles, estimating that the fraction of Galactic massive stars formed via triggering could be between 14 and 30%.

With this understanding of the important role that massive stars have on the ISM and Galactic evolution, we will now turn our attention to the astrophysical subject of this Thesis.

2.3 HII Regions

Following the overview in Chap. 1, this section covers the physics of HII regions, from how they develop and evolve; to how they emit radiation, allowing us to observe them. Let us begin by taking a look at the spectrum of the most abundant element in the ISM. In its simplest form, hydrogen is composed of one proton and one electron. The Bohr model of the atom tells us that the electron can exist in a number of quantised energy states. If the hydrogen atom absorbs a photon, causing the electron to move up energy levels, this is shown as an absorption line in the spectra. Conversely, if the electron moves down an energy level, this results in the emission of a photon with the discrete energy value given by the transition, producing emission lines. This is represented diagrammatically in Fig. 2.4. We also see here that to remove the electron from the atom, the process of ionisation, requires 13.6 eV. Transitions to or from the first three energy levels are named the Lyman, Balmer and Paschen series, respectively. Emission from the Lyman series yields photons with UV wavelengths. The Balmer series corresponds to photons in the

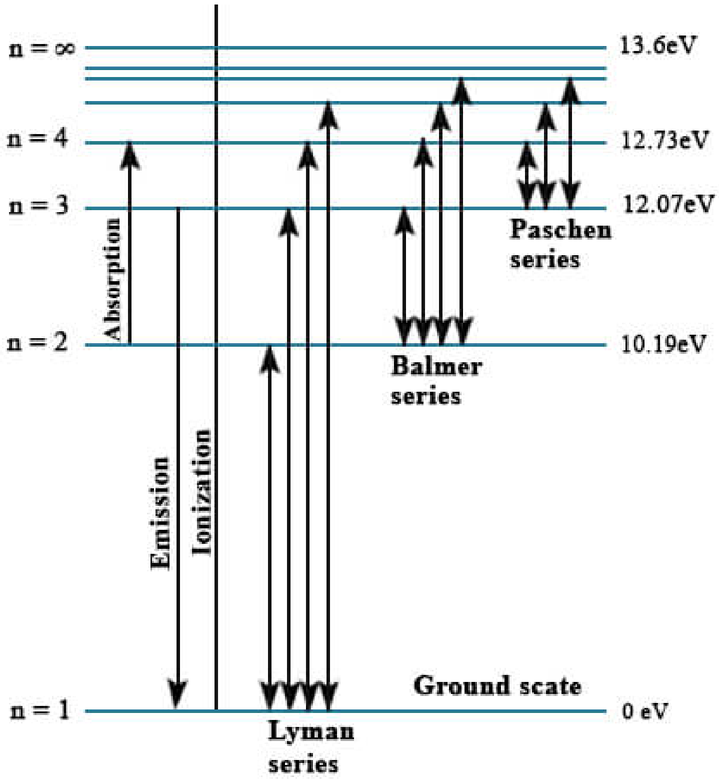


FIGURE 2.4: Hydrogen spectrum diagram showing the Lyman, Balmer and Paschen series of absorption and emission spectra. Absorption, emission and ionisation examples are indicated by the left most arrows. The photon energy required or emitted for different energy levels is given in eV on the right.

optical part of the electromagnetic spectrum. We have already seen an example of the Balmer 3-2 transition at 656.3 nm, which is more commonly referred to as the Hα line (the red channel in the OMC image, Fig. 2.2).

In the previous sections, we have seen that OB stars form embedded in dense molecular clouds. The energy required to disassociate the H_2 molecule is $\sim$4.4 eV. Therefore, as soon as an OB star starts to emit UV radiation, the surrounding hydrogen is dissociated and ionised. There are three main processes that govern the structure, properties and evolution of the resulting HII region around massive stars. The first is photoionisation equilibrium. This determines the structure of the nebula and the spatial distribution of ionised material. Next is the thermal balance between heating and cooling processes. This determines the temperature of the region and the amount of emission that is able to escape from the nebula. The third process is the hydrodynamics. This includes supersonic shock fronts, outflows and winds from the embedded stars. The following subsections will examine these processes and how they allow us to observe and analytically describe HII regions.

2.3.1 STRUCTURE

The structure of an HII region is primarily determined by the photoionisation balance in the nebula. Let us begin by investigating the ratio of ionised to neutral hydrogen within the region. For the purpose of introduction and overview, we will consider the case of a pure hydrogen nebula. Whilst this is an idealised example, the subsequent physical representation for all heavier atoms and their ionisation is incidental compared to the abundance of hydrogen in the vicinity of an OB star. For a given location within the HII region, the rate of photoionisation per unit volume is:

$$n_{\text{HI}} \int_{\nu_0}^{\infty} \frac{4\pi J_\nu}{h\nu} a_\nu d\nu \tag{2.2}$$

where n_{HI} is the number density of neutral hydrogen, J_ν is the flux of ionising photons and a_ν is the photoionisation cross section. The integral is taken over all photons with energies $h\nu \geqslant h\nu_0$, with $h\nu_0$=13.6 eV.

The rate of recombination per unit volume is:

$$n_e n_p \alpha\left(\text{HI}, T\right) \tag{2.3}$$

where n_e and n_p are the number densities of electrons and protons, respectively, and α is the recombination coefficient of neutral hyrdogen, which is weakly dependent on the electron temperature.

If we assume that all of the ionising photons are from a single ionising source, and neglect the minor contribution from the diffuse radiation within the nebula, at a distance r from the central star, the flux of photons at frequency ν can be expressed in terms of the stellar luminosity:

$$4\pi J_\nu = \left(\frac{4\pi R_*^2}{4\pi r^2}\right) \times \pi F_\nu(0) = \frac{L_\nu}{4\pi r^2} \tag{2.4}$$

where R_* is the stellar radius, $F_\nu(0)$ is the flux at the surface of the stellar photosphere and L_ν is the stellar luminosity for frequency ν.

Next, let us define the hydrogen neutral fraction, ξ, such that:

$$\begin{aligned} n_e = n_p &= (1-\xi)n_0 \\ n_{\text{HI}} &= \xi n_0 \end{aligned} \tag{2.5}$$

where n_0 is the ambient interstellar number density. Hence, $\xi = 0$ means that the hydrogen is fully ionised ($n_e = n_p = n_0$) and $\xi = 1$ fully neutral ($n_{\rm HI} = n_0; n_e = n_p = 0$).

We can now equate the photoionisation rate with the recombination rate to give the photoionisation equilibrium condition:

$$\xi n_0 \int_{\nu_0}^{\infty} \frac{L_\nu}{4\pi r^2} \frac{a_\nu}{h\nu} d\nu = (1-\xi)^2 n_0^2 \alpha\left(\mathrm{H\,I}, T\right) \tag{2.6}$$

In order to determine the fraction of neutral hydrogen, ξ, we must know some physical quantities. Let us assume a region of gas with $n_0 = 10\,\mathrm{cm}^{-3}$, and evaluate Eq. 2.6 at a distance of r=5 pc away from an O6.5V star with surface temperature T=40,000 K. From Panagia (1973), the number of ionising photons emitted by such a star per second is:

$$\int_{\nu_0}^{\infty} \frac{L_\nu}{h\nu} d\nu = N_{ly} = 6.6 \times 10^{48} \text{ photons/sec} \tag{2.7}$$

For a photoionisation cross section of $a_\nu = 6 \times 10^{-18}\,\mathrm{cm}^2$, this yields a number of ionisations per second of $\sim 10^{-8}\,\mathrm{s}^{-1}$. This implies a characteristic ionisation timescale of $\sim 10^8$ s, which is ~ 3 years. This is assuming the ionising flux has already reached the 5 pc radius.

For the recombination, using a recombination coefficient of $\alpha\left(\mathrm{H\,I}, T\right) \approx 4\times10^{-13}\,\mathrm{cm}^3\,\mathrm{s}^{-1}$ gives a recombination time for $n_{\rm H}$=10 cm^{-3} of $\sim 3\times10^{11}$ s, or $\sim 10^4$ years. Hence, in such an HII region, once the neutral hydrogen is ionised, it stays ionised for a long time before recombining, and when it does recombine, it is ionised again by the stellar radiation relatively quickly. Finally, substituting these values into Eq. 2.6 gives the fraction of neutral hydrogen to be:

$$\begin{gathered} \xi n_0 \times \left(10^{-8}\right) = (1-\xi)^2 n_0^2 \left(4 \times 10^{-13}\right) \\ \xi \approx 4 \times 10^{-4} \ll 1 \end{gathered} \tag{2.8}$$

Thus, the gas is nearly completely ionised.

In terms of the structure, the transition between the ionised and neutral material can be determined by considering the mean-free path of an ionising photon with $\nu = \nu_0$ at the location where $\xi = 0.5$. Here, $n_{\rm HI} = \xi n_0 = 0.5 \times 10 = 5\ \mathrm{cm}^{-3}$. The mean-free path is then:

$$\ell_{v_0} = \frac{1}{a_{v_0} n_{\rm HI}} \approx 3 \times 10^{16}\ \mathrm{cm} \approx 0.01\ \mathrm{pc} \tag{2.9}$$

which is only ∼0.2% of the size of the 5 pc region. Thus, the boundary of the ionised nebula has a definite 'edge'. The radius at which this occurs, for the idealised pure hydrogen region, is known as the Strömgren radius, R_s (Strömgren, 1939). This is the radius where the total number of ionising photons emitted per second is equal to the number of recombinations per second. Therefore, by evaluating the recombination rate over the radius of the nebula, from $r = 0$ to $r = R_s$; and taking $\xi \approx 0$, i.e. the region is fully ionised:

$$N_{ly} = \int_0^{R_s} n_0^2 \alpha \left(\mathrm{H\,I}, T\right) 4\pi r^2 dr \tag{2.10}$$

Carrying out the integration and rearranging therefore gives the equation of the Strömgren radius:

$$R_s = \left(\frac{3N_{ly}}{4\pi \; n_0^2 \; \alpha \left(\mathrm{H\,I}, T\right)} \right)^{1/3} \tag{2.11}$$

Whilst the Strömgren radius represents the idealised case, it is useful for deriving further properties, such as the dynamical age of the region, as we will see later in this section.

In the non-idealised case, one must consider the effects of the presence of helium and metals. Despite the fact that the number density of helium is $\sim 0.1\, n_{\mathrm{H}}$, the ionisation cross section for HeI is ∼10 times higher than the HI ionisation cross section. This means that photons with energies $\geqslant h\nu_{\mathrm{He\,I}} = 24.6\,\mathrm{eV}$ will primarily ionise the helium instead of the hydrogen. This is therefore dependent on the ionising source and whether the photons have sufficient energy. The result of this is that HII regions from stars later than O6 can have layers where both the hydrogen and helium are ionised, followed by layers where the hydrogen is ionised and the helium is neutral, since the recombination of He can result in photons with energy sufficient to ionise hydrogen. For stars earlier than O6, the entire ionised region consists of the mix of ionised hydrogen and helium.

The other effect to consider is that of metals and dust. Metals such as C, N, O and Si have low abundances relative to hydrogen and helium. Therefore, absorption by different ionisation stages of the metals does not significantly modify the radiation field of the nebula. It does, however, lead to various optical spectral emissions from the metals. The majority of this optical emission, for distant nebulae, does not reach us though, due to interstellar extinction by dust. As with the dust in GMCs, dust grains within HII regions can have a significant effect on the structure of the nebula. The dust absorbs the UV continuum photons, that would have otherwise gone into ionising the gas. The re-emitted photons by the dust are IR, hence not energetic enough to ionise hydrogen. This can hence decrease the size of the ionised zone. A countering effect to this is that the ionisation radiation field can photoevaporate the dust grains, which in turn, can liberate elements and molecules that are locked to the dust. The dust must therefore be carefully considered when numerically modelling HII regions. It also allows for us to detect HII regions via the IR emission.

2.3.2 RADIO EMISSION

Having described how HII regions are observable in the IR part of the continuum, with optical and UV emission heavily obscured, another excellent way of detecting HII regions is via radio continuum and line emission. In order to understand how the radio emission from an HII region arises, we will briefly consider the thermal structure of the ionised nebula. Let us assume that the main source of heating within the region is from the photoionisation. The different cooling mechanisms (that emit radiation) then include recombination, free-free continuum emission and line emission from collisionally excited ions.

Free-free (Bremsstrahlung) radiation is produced by the acceleration of electrons in the electric fields of positive ions. It is thus a continuum. The radiation intensity for a given frequency unit, I_ν, through unit solid angle, ds, is defined as follows:

$$\frac{dI_\nu}{ds} = j_\nu - \kappa_\nu I_\nu \tag{2.12}$$

where j_ν and κ_ν are the emission and absorption coefficients, respectively. For a system in local thermodynamic equilibrium (LTE), the coefficients are connected by the Kirchhoff relationship: $j_\nu = \kappa_\nu B_\nu(T)$, where $B_\nu(T)$ is the Planck function at the temperature of the system. For radio frequencies, where $h\nu/kT_e \ll 1$, we can use the Rayleigh-Jeans approximation to the Planck function: $B_\nu\left(T_e\right) \approx 2kT_e\nu^2/c^2$. The optical depth, τ, at frequency ν is then defined by: $d\tau_\nu = \kappa_\nu ds$. Eq. 2.12 is then rewritten in the form:

$$\frac{dI_\nu}{d\tau_\nu} = B_\nu\left(T_e\right) - I_\nu \tag{2.13}$$

This differential equation has the solution:

$$I_\nu = B_\nu\left(T_e\right)\left(1 - \mathrm{e}^{-\tau_\nu}\right) \tag{2.14}$$

having made the following assumptions: T_e is constant throughout the nebula, τ_ν is the total optical depth across the nebula and $I_\nu = 0$ at $\tau_\nu = 0$.

Thus, for an optically thin nebula, with $\tau_\nu \ll 1$, Eq. 2.14 has the approximate form:

$$I_\nu \approx B_\nu \left(T_e\right) \tau_\nu \tag{2.15}$$

Mezger and Henderson (1967) give the optical depth of radio continuum emission within an HII region to be:

$$\tau_\nu = 8.235 \times 10^{-2}\, T_e^{-1.35}\, \nu^{-2.1}\, n_e^2\, \ell \tag{2.16}$$

where ℓ is the mean-free path through the nebula. We thus arrive at the dependence of the radiation intensity on the frequency, since $B_\nu(T_e) \propto \nu^2$, it thus follows that for an optically thin nebula: $I_\nu \propto \nu^{-0.1}$.

Matsakis et al. (1976) then showed that by combining Eq. 2.10 (showing the number of ionising photons equal to the recombination rate), Eq. 2.15 (the radiation intensity for the optically thin nebula) and Eq. 2.16 (the expression for the optical depth), and combining all physical constants resulted in:

$$N_{ly} = 7.54 \times 10^{46} \left(\frac{S_\nu}{\mathrm{Jy}}\right) \left(\frac{D}{\mathrm{kpc}}\right)^2 \left(\frac{T_e}{10^4\mathrm{K}}\right)^{-0.45} \left(\frac{\nu}{\mathrm{GHz}}\right)^{0.1} \tag{2.17}$$

thus allowing for the number of ionising photons from the source to be calculated for a flux density $S_\nu = I_\nu \Omega_s$, where Ω_s is the source solid angle in steradians. This was derived using the assumptions outlined in Rubin (1968), concerning the temperature dependent factors and the emission measure, $\mathrm{EM} = \int_0^\ell n_e^2 d\ell$. We can therefore compare the observed value of N_{ly} from the radio continuum flux density to those tabulated in Panagia (1973) to determine the spectral type of the ionising star.

The other important cooling mechanism that gives rise to radio wave photons is recombination line emission. We saw previously in this section that for photoionisation equilibrium within the ionised plasma of an HII region, the number of recombinations equals that of the number of ionisations. Recombination usually occurs at weakly bound energy levels, with principle quantum numbers $n \gg 1$. Radio recombination line (RRL) emission is hence due to the cascade of the electron from a highly excited state, down through the energy levels to a low n. RRL is hence not due to the actual process of recombination itself.

Each different RRL has a well defined wavelength and frequency. Considering the general transition between levels of principle quantum numbers n and n'; the frequency of the transition, $\nu_{nn'}$ is given by:

$$\nu_{nn'} = R\left(1/n'^{\,2} - 1/n^2\right) \tag{2.18}$$

where R is the Rydberg constant expressed in frequency units (R = 3.3 $\times 10^{15}$ Hz). For $n' - n = \Delta n$, we can express Eq. 2.18 in terms of the wavelength: $\lambda_{nn'} \approx 4.5 \times 10^{-8}\, n'^{\,3}/\Delta n\,[\mathrm{m}]$. If, for example, we have $\Delta n = 1$ and $n' = 100$, then $\lambda_{nn'} = 4.5\,\mathrm{cm}$, which is in the radio range of the spectrum.

The detection of radio emission from HII regions will be described later in this chapter, in Sect. 2.4.

2.3.3 Dynamical Evolution

The final aspect of HII region theory that we will consider here is that of the dynamical evolution. This includes how the region expands, due to the photoionisation and hydrodynamical processes; and how we can hence determine the age of the region.

During the ionisation process, the temperature of the plasma is increased to the order of 10^4 K. Furthermore, the number of gas particles essentially doubles due to this process. As a result, the pressure within the nebula increases, which is not confined by the ISM, hence the nebula will expand. This is one of the main feedback mechanisms from massive stars, as both the HII region and surrounding ISM are pushed spherically outwards by the ionisation front (IF). Initially, the IF travels at supersonic speeds relative to the sound speed in the ambient ISM. This creates a shock wave. During this stage, both the shock wave and the IF can be described by a single radius, R and expansion velocity, dR/dt. This follows from showing that the ionisation boundary is relatively thin (Eq. 2.9).

The pressure behind the shock wave is assumed to be uniform, in both the neutral and ionised gas. We still assume photoionisation equilibrium throughout this stage, and that the neutral gas ahead of the shock is at rest. Hence, for a given HII region radius, R, the expansion rate dR/dt and corresponding expansion time can be derived from the region pressures.

The pressure in the shock wave, P_s, is equal to the pressure of the ionised gas, P_i. Taking c_s as the isothermal sound speed in the ionised gas, n_i the number density of ions and m_{H} the mass of the hydrogen atom, the latter is given by:

$$P_i = 2n_i kT_e \equiv n_i m_{\mathrm{H}} c_s^2 \tag{2.19}$$

From Dyson and Williams (1980), the pressure in the shock wave behind a supersonic shock is related to the shock velocity, V_s via:

$$P_s = \varepsilon \rho_0 V_s^2 \tag{2.20}$$

where ρ_0 is the ambient interstellar density and ε is the specific kinetic energy of the shock, which is taken as unity for an isothermal shock.

Since the velocity of the shock front is equal to the velocity of the IF expansion, we hence equate these Equations for $V_s = dR/dt = \dot{R}$ and $\rho_0 = n_0 m_{\mathrm{H}}$, giving:

$$\dot{R}^2 = c_s^2\,(n_i/n_0) \tag{2.21}$$

Assuming that the recombination rate for n_i is balanced by the number of ionising photons from the progenitor star, integrating Eq. 2.10 for a radius R gives:

$$N_{ly} = \frac{4}{3}\pi n_i^2 \alpha^2 R^3 \tag{2.22}$$

Then, rearranging and substituting for n_i in Eq. 2.21:

$$\dot{R}^2 = c_s^2 \left(\frac{3N_{ly}}{R^3\ 4\pi\ n_0^2\ \alpha} \right)^{1/2} \tag{2.23}$$

By factoring out the R^3 term from the parenthesis, we see that the expression remaining within is equal to the Strömgren radius, R_s, cubed (Eq. 2.11), hence:

$$\dot{R}^2 = c_s^2 \left(\frac{R_{\mathrm{s}}}{R} \right)^{3/2} \tag{2.24}$$

This means that for an observed HII region of radius, R, having determined the number of ionising photons, N_{ly} from the integrated radio continuum emission, we can hence determine the expansion rate of the IF and thus the time taken for the HII region to reach this radius. This is known as the HII region dynamical age, t_{dyn}. Dyson and Williams (1980) give the solution of this differential equation as:

$$t_{dyn} = \left(\frac{4R_s}{7c_s} \right) \left[\left(\frac{R}{R_s} \right)^{7/4} - 1 \right] \tag{2.25}$$

It is also shown in Dyson and Williams (1980) that pressure equilibrium between the HII region and the surrounding ISM would not be achieved within the main sequence lifetime of a massive star. Hence, the diffuse Galactic HII regions that we observe are all still expanding, assuming an ambient interstellar density of $10^3\,\mathrm{cm}^{-3}$. For higher values of ambient density, the expansion could be halted. Inhomogeneities in the ISM can also lead to non-spherical IFs.

The other dynamical factor to consider for HII regions is that of the stellar winds from the progenitor stars. Optical (Conti and Leep, 1974) and UV (Morton, 1967) observations of massive stars earlier than spectral type B2 show that these stars have strong stellar winds, with velocities in the range of $2000 - 3000\,\mathrm{km\,s^{-1}}$ and mass loss rates of the order $10^{-6}\,\mathrm{M_{\odot}\,yr^{-1}}$. Castor et al. (1975) found that such stellar winds cause a thin circumstellar shell to develop around the progenitor star, caused by the shock compression of the gas. This blows a large cavity, or 'bubble', into the ambient ISM. For massive stars with both an HII region and a stellar bubble, the stellar wind is confined within the ionised region, hence it is not as important for providing feedback to the ISM as the ionisation front. However, this theory does allow for later B type stars with strong stellar winds, but with insufficient luminosities for the creation of an HII region, to host a stellar bubble.

This theory was greatly built upon in Weaver et al. (1977), where a more detailed analytical description was provided and further observational consequences discussed. This theoretical work predicted the existence of molecular material on the edges of such stellar bubbles, whereby an expanding bubble into a dust cloud allows for the rapid formation of H_2 and more complex molecules. Escaping stellar flux and shock heating from the expanding region result in a photodissociation region (PDR) surrounding the HII region and stellar bubble. As with the dust associated with an HII region, the PDR emits in the IR range of the spectrum, due to the rotational and vibrational transitions of the molecules (Tielens and Hollenbach, 1985). This, consequently, led to many HII regions being identified visually from IR surveys. This is described in the next section, along with the radio continuum observations of HII regions.

2.4 Galactic Plane Surveys

This section introduces the Galactic Plane surveys that feature in this work. As already detailed in this chapter, star formation is concentrated in the Plane of the Galaxy, therefore, many surveys that aim to investigate this, along with the structure of the Galaxy, focus their coverage on a select area of the Plane. Table. 2.1 lists a selection of radio continuum surveys that cover this area. The observational data chosen for analysis in this work was from MAGPIS (the Multi-Array Galactic Plane Imaging Survey, Helfand et al., 2006), because of its high angular resolution (see Tab. 2.1 for comparisons), and the fact that it includes both interferometry and single dish observations. This means that it caters to both the large and small structure of the wavelength coverage. Table. 2.2 lists recent infrared Galactic Plane surveys, many of which have been used to construct HII region catalogues, via visual

Table 2.1: Radio Continuum Galactic Plane Surveys, their corresponding wavelengths, beam sizes and sky coverage. References: (1): Becker et al. (1994), (2): Hoare et al. (2012), (3): Beuther et al. (2016), (4): Helfand et al. (2006), (5): Condon et al. (1998), (6): Stil et al. (2006), (7): Taylor et al. (2003), (8): McClure-Griffiths et al. (2005), (9): Murphy et al. (2007)

Survey	Wavelength	Beam	Longitude	Latitude	Reference
MAGPIS	6 cm	4 - 9″	$350° < l < 42°$	$\|b\| < 0.4°$	1
CORNISH	6 cm	1.5″	$10° < l < 65°$	$\|b\| < 1°$	2
THOR	15 - 30 cm	20 - 40″	$15° < l < 67°$	$\|b\| < 1.25°$	3
MAGPIS	20 cm	6″	$5° < l < 48.5°$	$\|b\| < 0.8°$	4
NVSS	20 cm	45″	$66° < l < 247°$	$\|b\| < 8°$	5
VGPS	21 cm	60″	$18° < l < 66°$	$\|b\| < 1\text{-}2°$	6
CGPS	21 cm	60″	$66° < l < 175°$	$-3.5° < b < 5.5°$	7
SGPS	21 cm	120″	$253° < l < 358°$	$\|b\| < 1°$	8
MGPS-2	35 cm	45″	$245° < l < 360°$	$\|b\| < 10°$	9

inspection. The first part of this section focuses on detailing the MAGPIS image data. The latter part of the section summarises the IR surveys that were key to constructing the HII region catalogues used in this work.

2.4.1 MAGPIS

The observational data of HII regions in this work comes from the 20 cm, 1.4 GHz Multi-Array Galactic Plane Imaging Survey (MAGPIS, Helfand et al., 2006). MAGPIS combined Very Large Array (VLA) interferometry images with those from the 1.4 GHz, 100m Effelsberg telescope. The VLA (now known as the Karl G. Jansky Very Large Array, named for the discoverer of astronomical radio waves) is a centimetre wavelength radio astronomy observatory, located in New Mexico, USA. The array comprises 27 radio antennae, each with a dish diameter of 25 metres. The antennae are arranged in a Y shape configuration, with each arm 21 km long. The configuration of the array can be changed, via tracks for the telescopes to move along (see Fig. 2.5), which changes the angular resolution of the array. For example, with the array in the largest A configuration, this simulates a single dish that is 36 km in diameter; and results in a synthesised beam range of 24 to 0.043″ (for the respective frequency range of 74 MHz to 45 GHz; more on the radio beam later in this section). The MAGPIS survey collected data in the B, C, and D configurations, across $|b| < 0.8°$ and $5° < l < 48.5°$. The array was operated in a pseudo-continuum mode at 20 cm. This comprised two 25 MHz bandwidths centred at 1365 and 1435 MHz.

Even though data from the most compact VLA configuration was included to make the MAGPIS images, the resulting maps would have still suffered from missing

Figure 2.5: Photograph of the Karl G. Jansky Very Large Array. The tracks can be seen in the foreground of this image, whereby the positions of the radio antennae can be changed. Image credit: D. Lyon

fluxes from large scale structures ($>> 1'$). To rectify this, Helfand et al. (2006) combined the VLA images with those from the 1.4 GHz survey made with the Effelsberg 100 m telescope (Reich et al., 1990). The Effelsberg survey had an angular resolution of $\sim 10'$. To create the final MAGPIS images, the VLA and Effelsberg images were combined in Fourier space before converting back to the image plane. A restoring beam of $6.2'' \times 5.4''$ was used with a pixel scale of $2''$. The first catalogue from these data included over 3000 discrete radio sources (with an effective source detection threshold of 1.5 mJy), along with $\sim$400 regions of diffuse emission (Helfand et al., 2006).

Throughout this work, the radio continuum images from the MAGPIS survey are analysed by considering the signal from the HII regions as well as the background noise. We should therefore further examine the pixel values of the images: Jy/beam. Jy is short for the unit Jansky (again, named after Karl G. Jansky), and is a non-SI unit of flux density. Whilst we have already encountered the flux density, when detailing the radio continuum emission from HII regions; to properly arrive at this unit, we should first consider the emission source and the power that it radiates. The power, dP, received through a solid angle of $d\Omega$ for a frequency range of $\nu + d\nu$ is:

$$dP = I_\nu \cos\theta d\sigma d\Omega d\nu \tag{2.26}$$

where I_ν is the specific intensity of the emission through an infinitesimal surface area, $d\theta$, incident at an angle θ to the normal of the surface. For a source that subtends a well defined solid angle, we can measure the flux density, S_ν, as the power received by a detector of unit projected area. Rearranging Eq. 2.26 gives:

$$\frac{dP}{d\sigma d\nu} = I_\nu \cos\theta d\Omega \tag{2.27}$$

and then integrating over the solid angle subtended by the source results in:

$$S_\nu \equiv \int_{\text{source}} I_\nu(\theta, \phi) \cos\theta d\Omega \tag{2.28}$$

For source angular sizes <<1 rad, cos $\theta \approx 1$ and Eq: 2.28 simplifies to:

$$S_\nu \approx \int_{\text{source}} I_\nu(\theta, \phi) d\Omega \tag{2.29}$$

which is nearly always the case for astronomical sources. So, back to our original unit of the Jy; the SI units of flux density are W m^{-2} Hz^{-1}. However, astronomical flux densities are orders of magnitude lower than these standard units, hence the smaller unit of the Jansky was defined:

$$1 \text{ Jansky } = 1 \text{ Jy} \equiv 10^{-26} \text{ W m}^{-2} \text{ Hz}^{-1} \tag{2.30}$$

Having defined the flux density aspect of the pixel units, we shall now turn our attention to the beam. The beamwidth is essentially the resolution of the telescope. The beam can be thought of as the sensitivity of the instrument as a function of direction. This is the main disparity between radio telescopes, which detect waves, and e.g. optical telescopes, which detect photons. There will be maximum sensitivity in the direction that the radio telescopes are pointing, and the sensitivity drops off away from that direction. For most radio telescopes, the beams are nearly completely Gaussian. Figure 2.6 shows a representative main beam profile of affective aperture (the collecting area of the telescope or array) versus beam centre. As the angle from the central position increases, further side lobe beams are also incident on the telescope. The side lobes contribute to the noise pattern from interferometers, however, it is the main beam that constitutes what is effectively termed the beamwidth or beam size. This is specified by the angle θ_{HPBW}, which is the angle between the half power points of the Gaussian.

In order to calculate properties from the MAGPIS pixel values of Jy/beam, we must first define the solid angle of the beam, Ω_{A}. From Fig. 2.6, we can determine the beam solid angle by integrating over the solid angle subtended by the effective

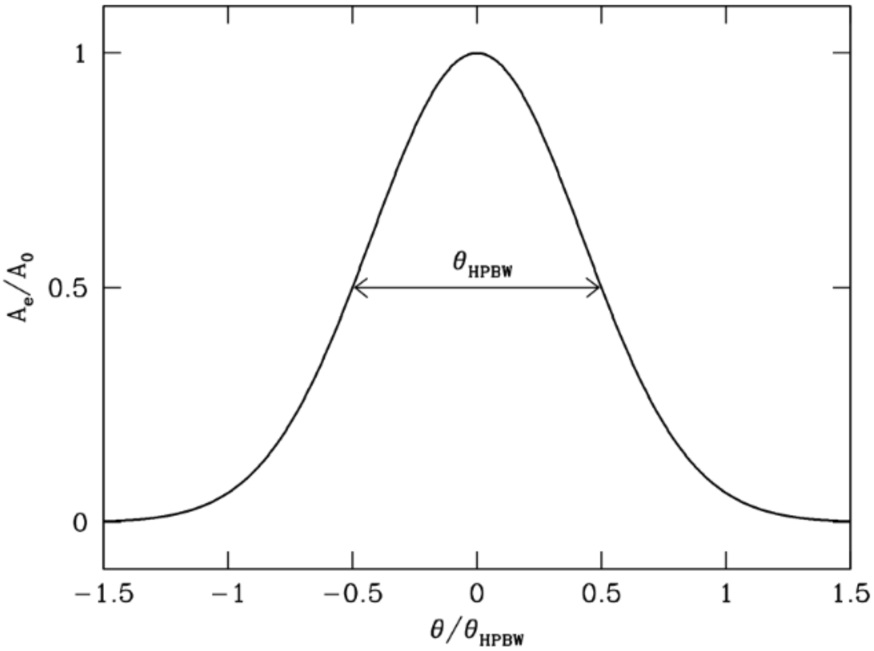


Figure 2.6: Example beam from a radio telescopes. Typically radio beams are nearly fully Gaussian, with their beamwidths specified by the angle θ_{HPBW} - between the half-power points. A_e is the effective aperture and A_0 the peak effective aperture.

aperture:

$$\Omega_\mathrm{A} \equiv \int_{4\pi} \frac{A_\mathrm{e}(\theta,\phi)}{A_0} d\Omega \tag{2.31}$$

Since the beam is nearly Gaussian, the ratio of effective aperture to peak effective aperture can be written as:

$$\frac{A_\mathrm{e}}{A_0} = \exp\left(-x\theta^2\right) \tag{2.32}$$

where θ is the angle from the beam centre. x is then a scaling factor such that $A_\mathrm{e}/A_0 = 1/2$ when $\theta = \pm\theta_{\mathrm{HPBW}}/2$ (Fig. 2.6). Substituting this in to Eq. 2.32 gives:

$$\frac{1}{2} = \exp\left[-x\left(\frac{\theta_{\mathrm{HPBW}}}{2}\right)^2\right] \tag{2.33}$$

and thus, rearranging for x gives:

$$x = \frac{4\ln 2}{\theta^2_{\mathrm{HPBW}}} \tag{2.34}$$

Substituting Eq. 2.34 into Eq. 2.32 then gives:

$$\frac{A_{\mathrm{e}}}{A_0} = \exp\left[-4\ln 2\left(\frac{\theta}{\theta_{\mathrm{HPBW}}}\right)^2\right] \tag{2.35}$$

Hence, returning to Eq. 2.31, we can now carry out the following integral:

$$\Omega_{\mathrm{A}} = \int_{\theta=0}^{\infty}\int_{\phi=0}^{2\pi} \exp\left[-4\ln 2\left(\frac{\theta}{\theta_{\mathrm{HPBW}}}\right)^2\right]\theta d\phi d\theta \tag{2.36}$$

by integrating first over ϕ, using the substitution $y = 4\ln 2\left(\theta/\theta_{\mathrm{HPBW}}\right)^2$:

$$\Omega_{\mathrm{A}} = 2\pi\left(\frac{\theta_{\mathrm{HPBW}}^2}{8\ln 2}\right)\int_{y=0}^{\infty}\exp(-y)dy \tag{2.37}$$

Therefore arriving at the final beam solid angle of:

$$\Omega_{\mathrm{A}} = \left(\frac{\pi}{4\ln 2}\right)\theta_{\mathrm{HPBW}}^2 \approx 1.133\ \theta_{\mathrm{HPBW}}^2 \tag{2.38}$$

We are now able to discern the amount of pixels contained in the beam for the MAGPIS images, thus enabling the extraction of the flux density values. For the MAGPIS images, $\theta_{\mathrm{HPBW}} \approx 5.8''$. Therefore, $\Omega_{\mathrm{A}} \approx 38.1''^2$. Dividing this value by the pixel size ($2''$) squared results in the number of pixels within the beam: ~ 9.5. Hence, for investigation of the flux density from a number of pixels, n:

$$S_\nu = \sum_i^n \left(\frac{\mathrm{Jy}}{beam}\right)_i \div 9.5 \tag{2.39}$$

HII regions in the MAGPIS images are readily observable from their radio continuum emission, however, they have a similar morphological structure to supernovae remnants (SNRs). Respective positions of both HII regions and SNRs are now well documented, from years of surveys and investigations. We will therefore now turn our attention to the infrared Galactic Plane surveys; and summarise the HII region catalogue constructions that are utilised in this work.

2.4.2 Infrared Observations

The Infrared Astronomical Satellite (IRAS) (Neugebauer et al., 1984) was the first of its kind to perform an all sky survey at infrared (IR) wavelengths from space. Since the principal limitation of IR astronomy is water vapour in the Earth's atmosphere, the images obtained by IRAS in 1983 were the highest resolution ever seen. Visual inspection of the IRAS images by van Buren and McCray (1988) revealed a selection of extended ring-like features (inferred to be bubbles in three dimensions) associated

Table 2.2: Infrared Galactic Plane Surveys, their corresponding wavelengths, resolutions and sky coverage. References: (1): Wright et al. (2010), (2): Benjamin et al. (2003), (3): Price et al. (2001), (4): Neugebauer et al. (1984), (5): Carey et al. (2009), (6): Doi et al. (2015), (7): Molinari et al. (2010)

Survey	Wavelength	Resolution	Longitude	Latitude	Reference
WISE	3 - 22 μm	6 - 12″	All Sky		1
GLIMPSE	4 - 8 μm	1 - 2″	-65°$< l <$ 65°	$\|b\| < 1°$	2
MSX	8 - 21 μm	18″	All	$\|b\| < 5°$	3
IRAS	12 - 100 μm	5 - 120″	All sky		4
MIPSGAL	24,70 μm	6,20″	-65°$< l <$ 65°	$\|b\| < 1°$	5
AKARI	50 - 200 μm	30 - 50″	All sky		6
Hi-GAL	70 - 500 μm	10 - 34″	All	$\|b\| < 1°$	7

with luminous stars. The source of this IR emission was deemed to be the dust in HII regions that reradiates the stellar flux and the molecular emission from the surrounding PDRs.

Stellar bubbles continued to be a common feature in subsequent IR surveys, e.g. the Midcourse Space Experiment (MSX; Price et al., 2001). The field of IR observations of bubbles was then revolutionised with the launch of the *Spitzer Space Telescope*, which led to the Galactic Legacy Infrared Mid-Plane Survey Extraordinaire (GLIMPSE, Benjamin et al., 2003) and the Multiband Infrared Photometer for *Spitzer* Galactic survey (MIPSGAL, Carey et al., 2009). Both of which provided ~ 10 times higher spatial resolution and ~ 100 times better sensitivity than MSX. Visual inspection of the original GLIMPSE survey by Churchwell et al. (2006, hereafter C06) revealed 322 full or partial rings, which were inferred to be three-dimensional stellar bubbles. A further 269 bubbles were then found within $|l| = 10°$ of the Galactic centre by Churchwell et al. (2007, hereafter C07). The authors noted in each paper that due to selection effects, the bubble catalogues were incomplete. These selection effects include: nearby bubbles favoured over distant ones; faint bubbles masked by background emission; ring morphology of small bright bubbles often not apparent due to diffuse emission filling the cavities; and small bubbles easily missed upon visual inspection. An estimate of at most 50% completeness was given in C06. A more realistic estimate however, would be less than 10% completeness, as we will see later in this section.

The bubbles were identified from false colour images of the Infrared Array Camera (IRAC, Fazio et al., 2004) wavebands 2 (4.5 μm , *blue*), 3 (5.8 μm , *green*), and 4 (8.0 μm , *red*) (Figure 2.7, Left). Although the bubbles are generally detectable in each of the IRAC bands, it is at the longer wavelengths in which the features are most prominent. This is due to the 7.7 μm and 8.6 μm emission peaks

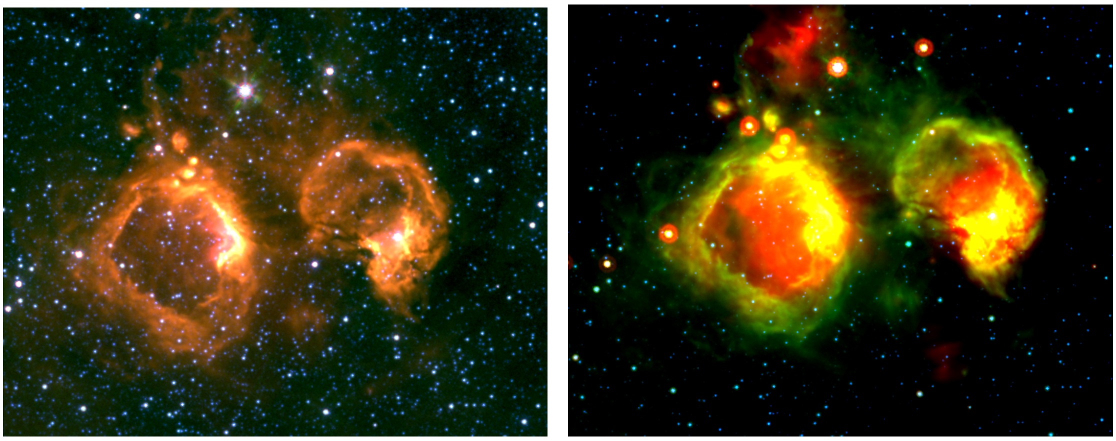


Figure 2.7: Comparison images of C07 bubbles CS44 and CS46. Left: GLIMPSE IRAC bands 4.5 μm (*blue*), 5.8 μm (*green*), and 8.0 μm (*red*). Right: IRAC and MIPSGAL bands 4.5 μm (*blue*), 8.0 μm (*green*) and 24.0 μm (*red*).

of Polycyclic Aromatic Hydrocarbons (PAHs) that are within the bandwidth of the 8.0 μm channel (Berné et al., 2009). The absence of 8.0 μm emission in the centre of the bubbles suggests that PAHs are present on the edge of the bubbles and in the photo-dissociation regions (PDRs) surrounding HII regions. The PAHs are either easily destroyed by the stellar radiation or are blown out by the stellar winds. Figure 2.7: Right shows an alternative false colour image using 4.5 μm and 8.0 μm IRAC bands (*blue* and *green*, respectively) and the MIPSGAL 24 μm band for *red*. The source of the 24 μm emission is thermal radiation from the mix of dust grains and ionised gas within the bubble (Watson et al., 2008). The 24 μm emission is usually concentrated at the centre of the bubbles, or within a torus inside the 8 μm shell, and is observed to be well spatially coincident with the radio continuum emission from HII regions (e.g. Deharveng et al., 2010). This spatial correlation decreases with increasing radial distance. Due to the differing temperatures of the dust, which is higher near the ionising source, this suggests that the dust grains are heated by Lyman continuum photons from the ionising source, rather than by the Lyman α radiation that is abundant throughout the ionised region. This shows that the dust is able to survive the intense radiation field of the ionising source and is also not evacuated by the stellar winds.

The 591 bubbles found in C06 and C07 were compiled as catalogues with the following observed parameters: Galactic longitude and latitude; semi-major and semi-minor axes of the inner and outer ellipses; average radius of the bubble; average thickness of the shell; and attributed morphological flag(s), assigned by the authors. Example images for each of the morphological flags can be found in Figues 2 and 1 of C06 and C07, respectively. It was found that the bubbles from C06 are strongly concentrated to the Galactic Plane, with an angular scale height of $0.63° \pm 0.03°$. This corresponds to a physical scale height of 44 pc at a distance of 4.2 kpc. The most common average angular diameter is between $1'$ and $3'$, with 88% smaller than $4'$. A large fraction of the bubbles have a large eccentricity, with an average value of $e = 0.65$.

The *Spitzer* surveys highlighted the abundance of stellar bubbles along the Galactic Plane, which was not apparent in earlier IR surveys due to sensitivity and spatial resolution limitations. C06 and C07 made a good start of investigating these bubble systems, noting that each catalogue was incomplete and further observational probes were still required in order to reveal the implications of bubble evolution on star formation and the ISM; and their relationship to diffuse HII regions.

Then, in 2012, the Milky Way Project (MWP) Catalogue (Simpson et al., 2012) arrived. This increased the number of identified IR bubbles from the *Spitzer* surveys by an order of magnitude. 5106 bubbles were catalogued by citizen science volunteers using the Zooniverse platform[3]. The users were shown an image of the Galactic Plane with the IRAC and MIPS colour stretch shown in Figure 2.7, Right. Users were then asked to draw ellipses around large enough bubbles; and to mark the positions of smaller bubbles (with outer diameter $< 0.64'$) or other objects of interest. The catalogue was then split into 3744 large bubbles and 1362 small bubbles. Each bubble was measured by at least five individuals, leading to a consensus catalogue comprising bubble positions, radius, thickness, and eccentricity.

The MWP catalogue rediscovered 85% of the C06 and C07 bubbles. It is not clear from the literature why the remaining 15% were not rediscovered, however it may be due to the construction of the MWP catalogue from the citizen science data. In order for a bubble to qualify for the final catalogue, it required a ‘hit rate’ – which is the ratio of number of times a bubble is identified to the number of times that region of space is shown to users – higher than 0.1. Since the bubbles discovered in C06 and C07 were identified by a select few experts, it could be that the missing 15% of Churchwell bubbles had a low hit rate and were thus not included in the final catalogue. The fraction HII regions from the Paladini et al. (2003) and the Anderson et al. (2011) catalogues (available at the time of publishing the MWP catalogue) that were positionally coincident with a MWP bubble was 86% and 96%, respectively. However, these matches were only 17% of the entire MWP bubble catalogue. We will see in the next subsection that the fraction of bubbles that actually correspond to an HII region is significantly higher.

The distributions of bubble properties from the MWP catalogue is generally

[3] https://www.zooniverse.org/

similar to those found by C06 and C07. The bubbles are concentrated on the Galactic plane, with a slight asymmetry towards lower latitudes. Many of the peaks in longitude distribution appear to derive from the larger scale structure of the Galaxy. There is a notable lack of bubbles in the region of the Galactic centre, which is likely due to confusion from background emission. The angular sizes of the MWP bubbles follow the same decreasing power law distribution as the Chuchwell bubbles, but with a slightly larger average for bubble thickness. This was an expected result due to the colour stretch and wavelengths used in the MWP images and an example of this is seen in the comparison image Figure 2.7. The MWP bubble peak eccentricity is ~ 0.35, a lower value than for the Churchwell bubbles. This can be explained by the fact that more smaller bubbles were identified in the MWP catalogue, which appear generally more circular. It was found that there is no correlation between bubble size and distance from the Galactic centre, and only a slight indication for bubbles further from the Sun to be larger, which is likely due to the corresponding selection effect. The MWP catalogue is a sufficient size to provide statistical evidence on the location of massive star formation and allows for inference into triggered star formation. This was investigated further in Kendrew et al. (2012), where a strong positional correlation of MYSOs and MWP bubbles was found. This independently concurred with the triggered star formation results obtained by Thompson et al. (2012).

The majority of focused morphological investigation of HII regions has been carried out via the IR analysis of bubbles (e.g. Ji et al., 2012; Xu and Ju, 2014; Dewangan et al., 2017; Topchieva et al., 2018). However, not all bubbles are the result of HII regions (Anderson et al., 2011), with some attributed to later B type stars with energies not sufficient for the ionisation of hydrogen, yet with strong enough stellar winds to evacuate the cavity in the ISM. Furthermore, the detailed mathematical shape analysis procedure conducted in this work (outlined in the Chap. 3), has not before been attempted for HII regions. For such a method, the radio continuum data were favoured over the IR data as a proof of concept. This was due to the homogeneity of the radio data, and it not suffering from point source contamination, as in the IR data. In order to confirm the selection of HII regions from the radio continuum images, to carry out the shape analysis work in this thesis, we next turn our attention to the largest compiled catalogue of HII regions.

2.4.3 The WISE Catalogue of Galactic HII Regions

The WISE Catalogue of Galactic HII regions (Anderson et al., 2014) comprises details of 8399 identified and candidate HII regions. It was constructed via visual

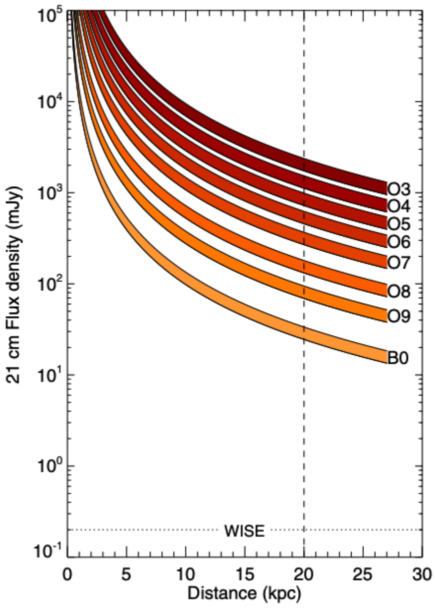


Figure 2.8: Figure 2 from Anderson et al. (2014). Model 21 cm radio flux densities for H ii regions ionised by single stars of spectral types O3-B0 as a function of heliocentric distance. The width of the curves corresponds to a range of nebulae emission temperatures from 5000 K to 10000 K. The vertical dashed line is the distance to the furthest Galactic H ii region. The horizontal dotted line is the corresponding sensitivity limit of the WISE 22 μm data.

and automated inspection of the WISE (Wide-field Infrared Survey Explorer) data, which had similar wavelength coverage to *Spitzer*, but covered the entire sky (see Tab. 2.2). Anderson et al. (2012) found that the WISE 22 μm flux was equal to that of the MIPSGAL 24 μm flux; having already shown how the MIPSGAL flux correlates with 21 cm radio flux with a conversion factor of ∼30. Therefore, the sensitivity of the WISE survey at 6 mJy was able to detect H ii regions with integrated 21 cm radio fluxes of ∼0.2 mJy, which is well below that required for Galactic H ii regions (see Fig. 2.8).

The authors carried out their visual inspection of the WISE IR images, searching for the characteristic morphology of MIR H ii regions, which was described in the previous subsection. Figure 2.9 shows an example image tile from Anderson et al. (2014). A selection of radio continuum surveys from Tab. 2.1 were then used to look for spatially coincident radio emission. This step also included matching the positions from the previously compiled radio continuum H ii region catalogues (e.g. Paladini et al., 2003; Giveon et al., 2005; Anderson et al., 2011; Bania et al., 2012). The automated section of their search involved matching the NVSS and MAGPIS 20 cm continuum data with the WISE point sources. A WISE colour criteria for the point sources was enforced, such that $[F_{12}/F_{22}] > 0.5$ (Anderson et al., 2012). This search only yielded a further 20 H II region candidates, suggesting that the visual search alone was sufficient to identify most Galactic H II region and H II region candidates. This WISE colour criteria could not reliably distinguish between H II regions and planetary nebulae (PNe). The automated search therefore resulted in hundreds of PNe candidates, which are identified via their high Galactic latitude distributions and lack of MIR nebulosity. Further literature searches were carried out by the authors on all of the H II region candidates to determine whether they could be PNe or external galaxies, removing hundreds of misclassified H II region candidates from the catalogue.

The final H II region catalogue comprises 8399 identified and candidate H II regions. Those classified as 'Known' include 1524 H II regions, which possess observed Hα spectroscopic and/or Radio-Recombination-Line (RRL) emission. The remaining regions are flagged as 'Group' (650 regions), if they belong to a known H II region complex; 'Candidates' (1986 regions), if they are spatially coincident with radio continuum emission but lack the Hα or RRL detection; and 'Radio-Quiet' (4124 regions) that were identified by their MIR morphology alone. There were a further 115 regions that lack the high-quality radio continuum observations of the candidate group (from historic observations). Parameters listed in the catalogue include: source name, classification, Galactic coordinates, approximate circular radius of MIR emission, properties of the Hα or RRL observations, LSR velocity and error, FWHM line width and error and corresponding references. In terms of correspondence with the MWP catalogue, the authors found that approximately half of all MWP bubbles are positionally correlated with a WISE H II region. The properties of the H II regions investigated in this thesis that were taken from the WISE catalogue are the Galactic coordinates and the determined distances. We should therefore take a closer look at those distances.

The WISE catalogue provided the distance of 1413 H II regions. There are three main ways to determine distances to H II regions: via association with a maser that has a measured parallax distance; spectroscopically; and kinematically. Maser parallaxes provide the most accurate distances, but are only available for a small subset of the sample. This is because masers generally exist before the diffuse H II region stage of massive stellar evolution. Once the H II region starts to expand, it destroys the masing material (Walsh et al., 1998). Spectroscopic distances require accurate identification of the ionising source, which can generally only be done for nearby bright stars that have low line-of-sight extinction. Hence, spectroscopic distances were not used for the WISE H II region distances due to their large uncertainties. Parallax distances were used where possible, for 62 of the H II regions. The remaining

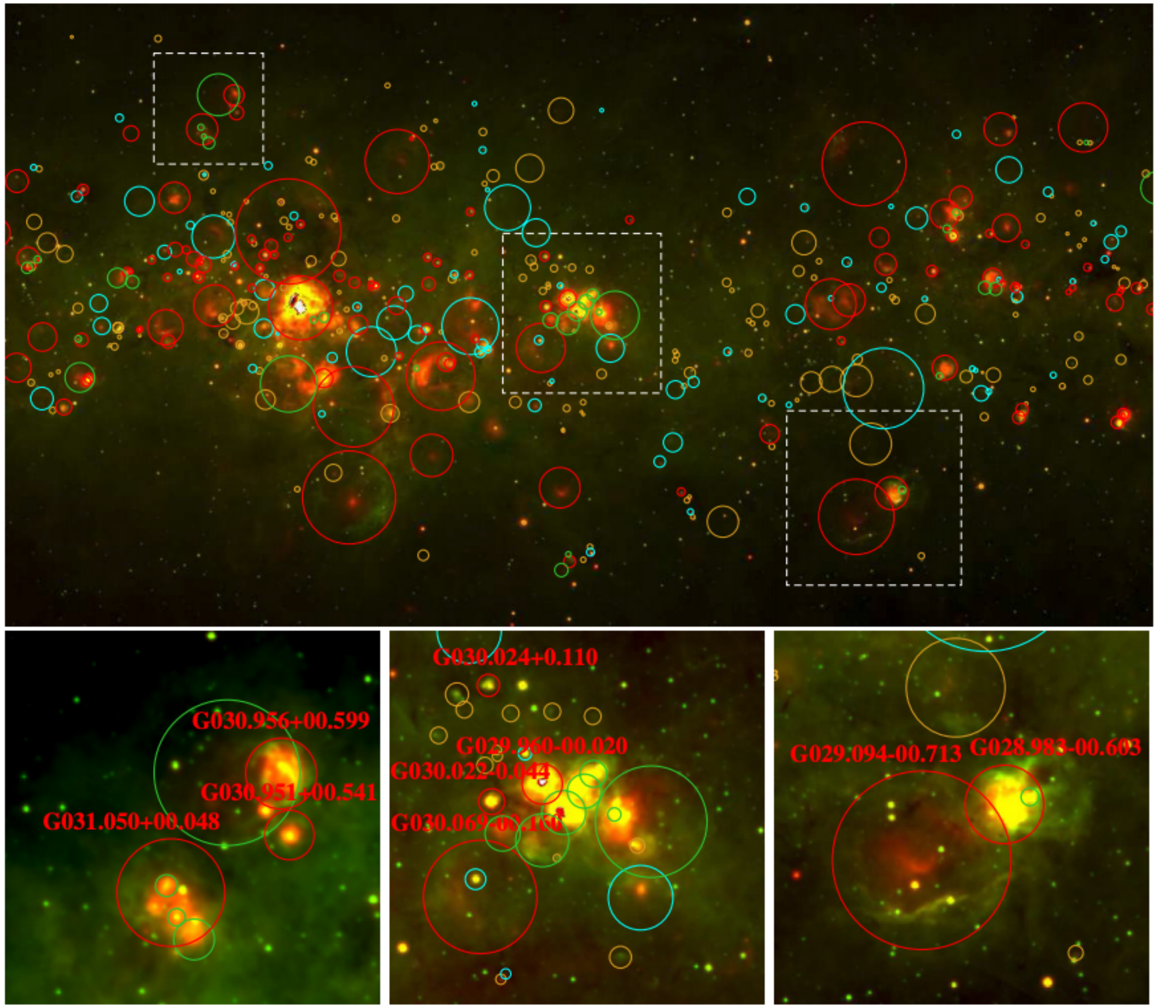


Figure 2.9: Figure 3 from Anderson et al. (2014). The background images are the WISE 22 μm data in red and WISE 12 μm data in green. The red, green, cyan, and yellow circles show the locations of the known, group, candidate, and radio quiet sources, respectively. The top image is 3°by 2°, centred at l=30°, b=0°.

distances were determined using the kinematic method.

Kinematic distances are, in principle, available for all HII regions with measured velocities. However, kinematic distances suffer from a number of uncertainties. Firstly, they require knowledge of the Solar orbital speed around the Galactic centre, and the Sun's distance from the Galactic centre. This is seen by referring back to Fig. 2.1, where the outer dashed circle represents the Solar circle and the Sun's position in the Galaxy is shown. Anderson et al. (2014) used the Brand et al. (1986) rotation curve model, with a Solar circular orbital speed of $\theta_0 = 220\,\mathrm{km\,s^{-1}}$ and a Sun – Galactic centre distance of $R_0 = 8.5\,\mathrm{kpc}$. More recent models provide slightly different values. The Reid et al. (2009) model uses maser parallaxes and proper motions to provide values of $\theta_0 = 254\,\mathrm{km\,s^{-1}}$ and $R_0 = 8.4\,\mathrm{kpc}$. More recently, Russeil et al. (2017) tested various models and found that a power law fit of $\theta(R)/\theta_0 = 1.022\,(R/R_0)^{0.0803}$, with θ_0=240 km s^{-1} and R_0=8.34 kpc is the best fit to the data considered in that study. The choice of orbital model typically only has a small systematic affect on the final determined distances. Next, the source LSR velocities are obtained from the Doppler shifted spectra emitted from the rotational transitions of molecular lines, which is a relatively straightforward step (e.g. Urquhart et al., 2011a,b). Using this LSR velocity of the object and the Solar orbital model, one can determine the galactocentric and heliocentric distances. However, returning to the uncertainties; for sources whose LSR velocity places them within the Solar circle, there are two possible solutions for each radial velocity, corresponding to two radial distances, known as the near and far distance. Objects located either beyond the Solar circle, or along the tangential points are exempt from this effect (again, from Fig. 2.1, the smaller dashed circle is the locus of the tangent points). This effect is known as the kinematic distance ambiguity (KDA).

In order to resolve the KDA, auxiliary data are required to determine whether the near or far distance is correct for a given HII region. There are three main techniques for resolving the KDA: using HI Emission/Absorption (HI E/A); using H_2CO (Formaldehyde) absorption; and using HI self-absorption (HI SA). In the HI E/A method, the radio continuum HII emission is absorbed by foreground neutral HI. HI will only absorb the thermal continuum emission if the brightness temperature of the HI is less than that of the HII region at 1.4 GHz. Since the continuum emission spans a range of frequencies there is always the potential for foreground HI to absorb part of the continuum. Radial velocities increase to a maximum at the tangent point along any line of sight, therefore HI absorption at velocities higher than that of the HII region implies that the HII region is at the far distance. If, however, the radial velocities of the HI are smaller than the object's velocity, the object is located at the near distance. This method is analogous to the H_2CO absorption method, with cold molecular clouds containing H_2CO acting as the foreground absorber. The HI SA method is useful for resolving the KDA for molecular clouds and, in this context, relies upon the association between molecular clouds and HII regions (e.g. Anderson and Bania, 2009). Cold foreground HI will absorb warmer background HI at the same velocity. Thus HI within a cold molecular cloud (T ∼ 10 K), that is associated with an HII region at the near distance will absorb the warmer HI from the ISM (T ∼ 100 K) at the far distance. Whereas, the HI in a molecular cloud at the far distance shows no such self absorption, since there is no background HI at the same velocity. Whilst still an indicator to resolve the KDA, the HI SA method is considerably less reliable than the former two methods (Anderson and Bania, 2009).

For the distance assignments of the WISE HII regions that were subject to the KDA, Anderson et al. (2014) prioritised using the H_2CO far distance, where possible. This is because background fluctuations in HI can cause false absorption signals. However, for the near distances, priority shifts to the HI E/A method, since a lack of H_2CO absorption at the near distance may simply be due to the lack of H_2CO. The HI is much more abundant. Finally, if neither HI nor H_2CO E/A is known, they then used the HI SA method. All sources with LSR velocities within 10 km s^{-1} of the tangent point velocities were assigned the tangent point distance. The overall uncertainty of the distance assignments for the entire catalogue (where distances are determined) is computed from the following combined considerations: The Galactic rotation curve choice, allowing for different θ_0 values. Streaming motions of 7 km s^{-1}, this is where the gas in the spiral arms rotates at a different speed to the spiral arm. And tangent point distance uncertainties that consider the near distance uncertainty and the difference between the near and the far distances. These considerations result in an average distance uncertainty of ∼15% for the first quadrant of the Galaxy – the quadrant that corresponds to the MAGPIS survey coverage. However, we should still remember that if the wrong KDA assignment was made, this uncertainty is significantly higher.

The information presented in this section pertains mainly to Chap. 4, where the MAGPIS HII region data are analysed. The final section in this chapter provides an overview of the numerical simulations of HII regions, that are the subject of Chap. 5.

2.5 Numerical Simulations of Hii Regions

This section provides an introduction and overview of numerical simulations of Hii regions. The goals of numerical simulations are to analytically recreate astrophysical observations in an attempt to explain the underlying physics that is not always apparent from the observations themselves. An example of this is the observed star formation efficiency (SFE). Galactic observations put this at a few percent (Lada and Lada, 2003). Numerical simulations of SF via the collapse of cores within GMCs have to include various feedback mechanisms to drive down the SFE to a comparable rate; for example, by introducing thermal feedback that can prevent fragmentation of the gas into clumps (Krumholz et al., 2007; Bate, 2009a,b). Gas dispersal is another mechanism for altering the SFE of molecular clouds. As we have seen throughout this chapter – both of these support mechanisms are resultant from Hii regions on their surroundings. Dale et al. (2007a,b) showed via numerical models of clouds irradiated by ionising stars, both internally and externally, that feedback caused some stars to form earlier, compared to the control runs without feedback. Furthermore, evidence for triggered star formation was noted. In this instance, the overall SFE of the cloud was increased by the ionisation feedback. On the other hand, simulations by Walch et al. (2013) found that although triggering was effective on small timescales, larger timescales resulted in a reduced SFE due to the dispersal of the gas, as the Hii regions evolved. More recently, the ionising radiation models of Geen et al. (2017) displayed a low SFE that was consistent with the Galactic observations.

Investigating the different feedback mechanisms and their combinations is hence an ongoing pursuit of theorists. Another aspect of this is producing numerical simulations that are fully representative of what we observe in the Cosmos. Synthetic observations of numerical simulations can be generated and compared to their observational counterparts. This is the main investigation of the numerical simulations in this work, by employing the statistical shape analysis method (detailed in Chap. 3) to provide insight to this comparison. The remainder of this section outlines the physical processes modelled by the numerical simulations of Ali et al. (2018), whose data is analysed in Chap. 5. The objective here is to provide a working understanding of how the synthetic Hii regions from these models are generated. Analogous to the outline given thus far - for the the MAGPIS Hii region images.

The Ali et al. (2018) numerical simulations use the Monte Carlo (MC) radiative transfer (RT) and hydrodynamics (HD) code `TORUS`, as described in Harries (2015).

In TORUS, the MCRT is coupled with the HD module, allowing for a self-consistent radiation-hydrodynamical simulation. The MCRT model is based on the model developed by Lucy (1999), which is widely used in the theoretical community, not just for modelling ionising feedback. The basis for this model is as follows: at the beginning of a simulation time-step duration of Δt, the total stellar luminosity, L, is split into a total of N **packets**. Each packet represents a bundle of photons of a particular frequency. Each ith packet has an energy:

$$\epsilon_i = w_i \frac{L\Delta t}{N} \tag{2.40}$$

with packets weighted by a factor w_i (the sum of which is normalised to 1) depending on the frequency of the photons. This ensures that packets containing photons of high frequency effectively have fewer photons and vice versa. A word on the simulation nomenclature, which will be defined as we go: the simulation **grid** is the boundary in which the entire simulation takes place. The grid is split into discrete **cells**, with defined properties based on their constituents. It is within cells where **events** take place. Events can be absorption, emission, scattering or cell crossings by the packets.

Packets propagate through the simulation grid with randomly sampled path lengths, ℓ, between events. After absorption/emission events, packets are re-emitted with a frequency sampled from a probability density function constructed from the cell emissivities and appropriately re-weighted with a new w_i. For a path length time of $\delta t = \ell/c$, the total time-step Δt contributes an energy of $\epsilon_i \delta t/\Delta t = \epsilon_i \ell / c\Delta t$ to the radiation field. The total energy density, u, is then summed over all events as:

$$\mathrm{d}u_v = \frac{1}{c\Delta t V} \sum \epsilon \; \ell \tag{2.41}$$

where V is the volume within the grid. This is proportional to the mean radiation intensity I_v via:

$$\mathrm{d}u_v = \frac{4\pi I_v}{c} \mathrm{d}v \tag{2.42}$$

Combining and rearranging these equations gives:

$$I_v dv = \frac{1}{4\pi} \frac{\epsilon}{\Delta t} \frac{1}{V} \sum_{dv} \ell \tag{2.43}$$

Finally, we arrive at the corresponding MC estimator from the Lucy (1999) model for radiation intensity, I, where the summation, now over all cells in V, regardless of frequency is thus:

$$I = \frac{1}{4\pi} \frac{\epsilon}{\Delta t} \frac{1}{V} \sum \ell \tag{2.44}$$

This MC estimator from the path length algorithm is then used for applications of radiative equilibrium and photoionisation equilibrium. It allows for any equation that requires the radiation intensity, I_v, to be estimated by via the photon packets in the simulation. For example, the photoionisation balance obtained by equating the ionisation and recombination rates gives the ratio of number densities, n, of successive ionisation states, i, of species X, (Osterbrock and Ferland, 2006):

$$\frac{n\left(X^{i+1}\right)}{n\left(X^{i}\right)} = \frac{1}{n_e \alpha\left(X^i, T\right)} \int_{v_I}^{\infty} \frac{a_\nu\left(X^i\right) 4\pi I_v}{hv} \mathrm{d}v \tag{2.45}$$

where n_e is the electron number density, α is the total recombination coefficient to all levels, a_v is the absorption cross section and v_I is the ionisation frequency for species X^i. The MC estimator in Eq. 2.44 can then be used to replace the mean radiation intensity, I_v:

$$\frac{n\left(X^{i+1}\right)}{n\left(X^{i}\right)} = \frac{1}{n_e \alpha\left(X^i\right)} \frac{1}{V \Delta t} \sum_{v_I}^{\infty} \frac{a_v\left(X^i\right) \epsilon \ell}{hv} \tag{2.46}$$

We thus have the ratio of number densities in the simulation from the packet energies, ϵ and the sampled path lengths, ℓ. Similar approaches are taken for the thermal balance, the radiation pressure and the far-UV interstellar radiation field. The explicit steps for these further physical representations will not be repeated here, but are described in full in Ali et al. (2018). As is the hydrodynamics handling, which takes place on alternating steps with the radiative transfer throughout the simulation. The initial conditions for the numerical simulation are covered in full in Chap. 5. This, therefore, concludes the overview of the physical processes that are modelled by the simulations, with the knowledge of how the numerical HII region was constructed.

In the following chapter, the shape analysis method is fully explained, along with the various statistical measures employed, in conjunction with the shape analysis, to investigate the morphologies of HII regions.

Chapter 3

Shape Analysis & Statistical Methods

This chapter introduces the concept of shape from both an astronomical and mathematical perspective. Details of how the HII region shapes were extracted and quantified are explained herein. Also introduced and explained are the various statistical measures that are employed in this work in conjunction with the shape analysis, in order to systematically compare the shapes of HII regions.

3.1 Shape Analysis

3.1.1 Astronomical Shape

The concept of shape in astronomy is a relatively recent field of study. Whilst we are now used to seeing spectacular, colourful images of many kinds of astronomical phenomena, it is only thanks to the advances in both telescopes and imaging techniques that we are able to appreciate how much shape and structure the Cosmos has. Before astronomers had telescopes, nearly everything in the night sky appeared to be point sources of light, with the few exceptions being the Moon, Milky Way, Magellanic Clouds, Andromeda and perhaps some comets (examples shown in Fig. 3.1).

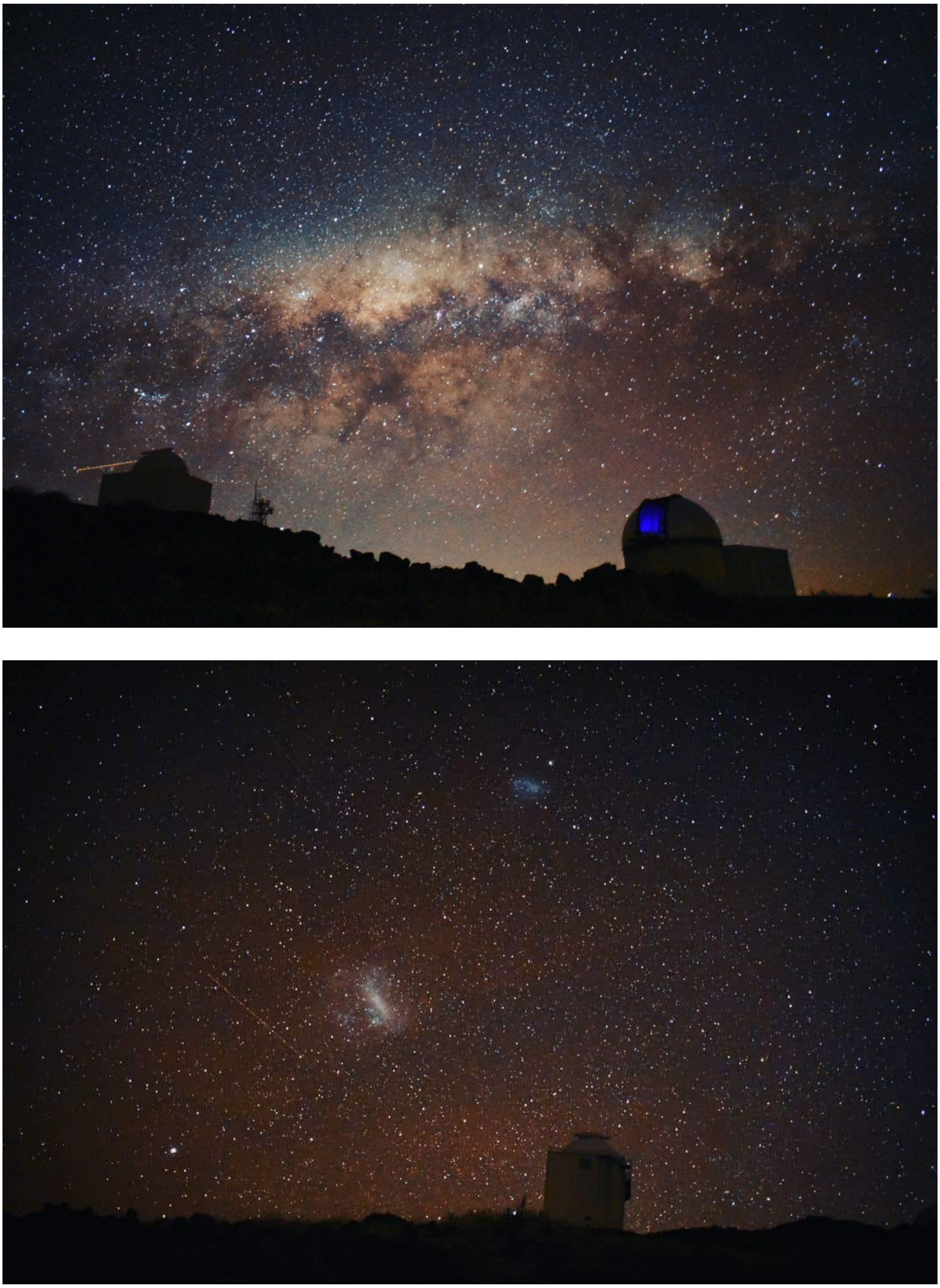


Figure 3.1: Examples of shape in astronomy. Top: Centre of the Milky Way Galactic Plane. Bottom: Large and Small Magellanic Clouds. Image credit: B. Rojas-Ayala, taken at the European Southern Observatory site, La Silla, Chile.

Figure 3.1: Continued from previous page. Night sky view of the Andromeda Galaxy. Image credit: T. Van (EarthSky), Montana, USA.

The planets were the first of these point sources revealed to have a definite shape, even with a magnification of just 8x that Galileo's telescope originally possessed. An example of where telescope and imaging advancement rapidly improved was with the misnomer of planetary nebulae. When the English astronomer William Herschel and French astronomer Antoine Darquier de Pellepoix first independently observed these nebulae, they each commented on the resemblance with the recently resolved shapes of the gas- and ice-giant planets. It was not until spectroscopic investigation of planetary nebulae that their true nature was revealed. Until then they were thought to be stars collapsing and condensing into planets, rather than stars reaching the end of their lifetimes and shedding their outer layers. Images from the Hubble Space Telescope (see for example: Fig. 3.2) in the 1990's revealed the stunning and complex structures within the layered shells of planetary nebulae, showing there was more complex structure than simply a diffuse disk.

The first morphological classification of celestial objects took place in the early 20th Century, which began with The Great Debate (also known as the Shapley - Curtis Debate), which was held on 26 April 1920 at the Smithsonian Museum of Natural History. The debate sought to substantiate the nature of the so-called spiral nebulae (such as Andromoeda) that had been readily observed in the previous cen-

Figure 3.2: Hubble Space Telescope image of the planetary nebula: Kohoutek 4-55 (or K 4-55). Image credit: NASA STScI.

tury. Harlow Shapley believed that the nebulae were relatively small and within the confines of the Milky Way, whilst Herber Curtis argued that they were in fact independent galaxies, implying that they were at exceedingly large distances. Distances which at the time, many astronomers refused to believe were conceivable. However, whilst these nebulae were originally grouped via their common diffuse shapes, it was revealed that they were comprised of stellar material and Curtis' strongest evidence for them being outside the Milky Way was that Andromeda alone was observed to have more novae explosions than the entirety of the Milky Way itself. The argument was settled a few years later when in 1927 Georges Lemaître and in 1929 Edwin Hubble independently theoretically (by the former) and observationally (by the latter), showed that the Universe is expanding and the spiral-nebulae were indeed entire galaxies at immense distances from the Milky Way (Lemaître, 1927; Hubble, 1929).

Edwin Hubble's observations of many galaxies led to the Hubble Sequence, which was taken as a morphological 'classification system' for galaxies, showing a sequence based on their shapes (Fig. 3.3). Hubble's scheme separates galaxies into three main classes: ellipticals, lenticulars and spirals. Whilst the Hubble sequence is still the most widely used naming system for morphological classes of galaxies, its common physical interpretation of an evolutionary sequence, from elliptical (known as early-

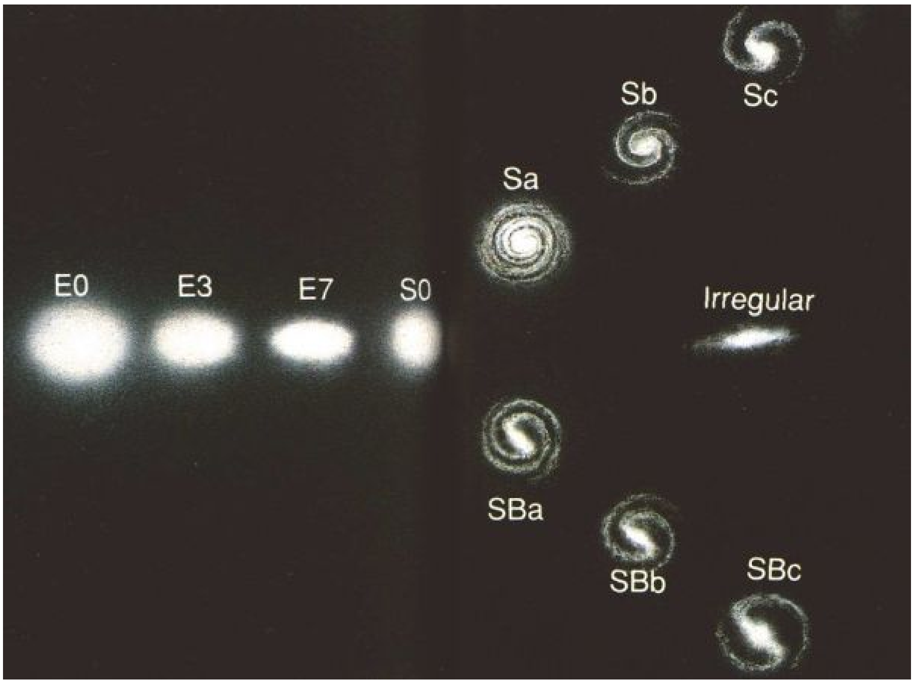
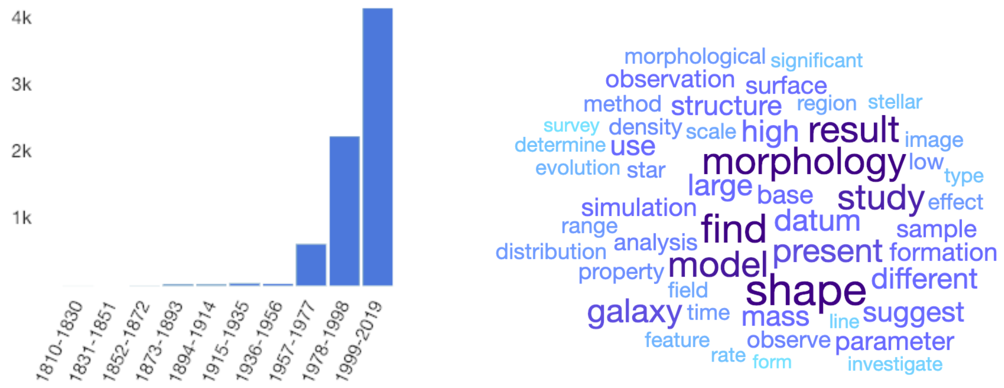


Figure 3.3: Edwin Hubble's observational tuning fork diagram of galaxy morphology classes (Hubble, 1982)

type galaxies) through to spiral (late-type galaxies), has since been shown to be incorrect. From Hubble's paper: 'The nomenclature, it is emphasised, refers to position in the sequence, and temporal connotations are made at one's peril. The entire classification is purely empirical and without prejudice to theories of evolution' (Hubble, 1927). Baldry (2008) wrote that: 'Hubble took these terms from spectral classification of stars to signify a sequence related to complexity of appearance, albeit based on images, not spectra'. A strong argument against galaxies evolving from elliptical through to spiral is the measured rotation speeds of each, with spirals observed to rotate much faster than ellipticals. Therefore, conservation of angular momentum (up to a factor of five higher in spirals than ellipticals (Shi et al., 2017)) prevents such systems evolving to the latter state. The Hubble sequence is hence a perfect example of the difference between unsupervised clustering and supervised classification, which will be revisited later in this chapter.

Looking at the current status of shape research in astronomy, searching The SAO/NASA Astrophysics Data System[1] (ADS) for the terms 'morphology', 'morphological' or 'shape' in the title of refereed astronomy articles returns over 7000 results (breakdown by years shown in Fig. 3.4a). The vast majority of these pertain to galactic morphology (see Fig. 3.4b where 'galaxy' is a large highlighted term from the content of those articles). In recent years, this is in part due to the great success

[1] https://ui.adsabs.harvard.edu/



(a) Number of referred journal articles published.

(b) Word cloud of terms found in the returned articles. Word size corresponds to usage frequency.

Figure 3.4: Overview of The SAO/NASA Astrophysics Data System refereed astronomy articles featuring the terms 'morphology', 'morphological' or 'shape' in the title

of Galaxy Zoo and the Zooniverse platform (as we saw in Sect. 2.4.2). Including the term 'Hii regions' to the previous search criteria returns 109 results on ADS. This is not limited to having Hii region in the title, nor is it restricted to Galactic Hii regions as is the topic of this thesis. It does suggest, however, that the community believes there is useful information to be gained by investigating the shapes of Hii regions in particular.

3.1.2 Mathematical Shape

Turning our attention back to Earth, the concept of shape from a mathematical perspective will now be introduced. **Shape** is defined by Kendall (1977) as 'all the geometric information that remains when location, scale and rotational effects are filtered out from an object'. Therefore, an object's shape is invariant under Euclidean transformations, providing an intrinsic property of the object in question. If one would like to maintain information of scale, an object's **size-and-shape** can be considered. Two objects are said to have the same size-and-shape if a rigid body transformation of one shape can match it exactly on to the other. The entirety of shape statistics follows from this general definition of shape and what affects it. In order to understand how to compare shapes, we need to look at how we use coordinates or 'landmarks' to quantify shapes.

'A **landmark** is a point of correspondence on each object that matches between and within populations' (Dryden and Mardia, 1998). Three basic types of landmarks exist: **anatomical, mathematical** and **pseudo-landmarks**. Anatomical landmarks, as the name suggests, are points of organism shapes assigned by ex-

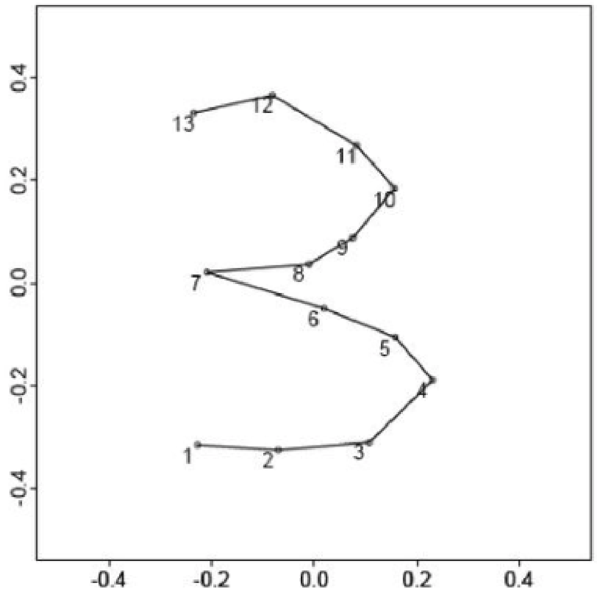


Figure 3.5: Examples of mathematical and pseudo landmarks, Figure 10 from Dryden and Mardia (1998). Digitised handwritten '3' with 13 labelled landmarks. Landmarks 1, 4, 7, 10 and 13 are mathematical landmarks, the remaining are pseudo landmarks at approximately equal distance between the mathematical landmarks.

perts in some biologically meaningful way, for example, at the corners of the eye and tip of the nose for a face. Such landmarks are hence widely used in facial recognition software. Mathematical landmarks are points located on an object according to some mathematical or geometrical property of the figure. This could be a region of high curvature or an extreme point on the shape. An example of the uses of these landmarks is in handwriting digitisation (Fig. 3.5). Pseudo landmarks are constructed points located around the outline of an object or in between mathematical landmarks. Figure 3.5 shows an example of landmark placement on a digitised handwritten number three. In this example, mathematical landmarks are assigned at points 1 (extreme bottom left), 4 (maximum curvature of bottom arc), 7 (extreme end of central protrusion), 10 (maximum curvature of top arc) and 13 (extreme top left). The remaining landmarks are all pseudo landmarks assigned at approximately equal distance between each of the mathematical landmarks. It is clear from this image that the pseudo landmarks are assigned at exact intersects between vertices, where the image has been digitised from the handwritten curve. In this thesis, equally spaced pseudo landmarks are considered along the boundary of each Hii region in a similar, quantised manner. This is further explained in Sect. 3.1.4.

When comparing shapes using shape statistics, it is common for landmarks to be treated with a one-to-one correspondence between shapes. This allows for computation of mean shape and superimposition using Generalised Procrustes Analysis (GPA) (Gower, 1975). In GPA, the landmarks of a set of shapes are aligned and the mean placement of all landmarks calculated. This is useful if one would like to determine how much a given shape differs from the sample mean shape. A recently developed computational tool to determine the optimum number of landmarks to characterise an object's shape is given in (Watanabe, 2018). However, it may not always be possible to have the same number of landmarks per object, especially when considering size-and-shape. In this instance, one can chose to keep a constant spatial resolution when attributing the landmark placements; hence the number of landmarks are determined from the data (Bookstein, 1986). In this situation, one considers the landmarks of an object represent a '**shape-space**' and the '**shape-distance**' between objects can be determined by using a suitable '**shape-descriptor**'. This represents how (dis)similar the objects are. Shape-descriptors, in the most simple cases, are measures of radius or eccentricity. In this work, a curvature distribution of each HII region was defined as the shape-descriptor, which is determined from the size-and-shape landmarks at a constant spatial resolution. Before this is described in detail, the next subsection explains how the landmarks are systematically assigned to the boundaries of the HII regions that are analysed in this work.

3.1.3 Extracting the Shape of HII Regions

This subsection explains how the morphologies of the HII regions, analysed in this work, were extracted using image contouring at a given surface brightness. Section 2.4.1 provides an overview of the MAGPIS radio data and survey, which the observational HII region data were taken from. Described here, is the general approach to how the image data were systematically quantified into landmarks across the observational sample. For the synthetic observations, details of how the shapes of those regions were extracted is explained fully in Chap. 5.

To generate contours from each region, the signal value was first determined by applying sigma clipping to the pixel values of the image tile. During sigma clipping, the median (m) and standard deviation σ of the sample is determined. Any values that are smaller or larger than $m \pm \alpha\sigma$ are then removed from the sample. For the MAGPIS tiles, $\alpha = 5$ was used. This essentially removes the signal from the pixel values, leaving us with the Gaussian noise profile for the tile (Fig. 3.6). Taking the mean and σ from the sigma clipped distribution then allowed

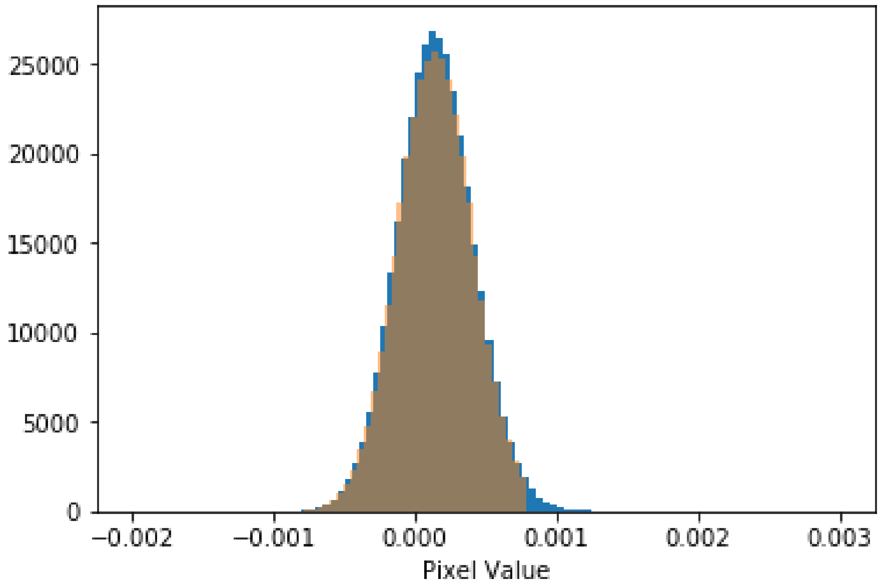


Figure 3.6: Histogram of pixel values from an example MAGPIS image tile containing an HII region. Blue shows the original pixel values, orange shows the remaining pixel values after sigma clipping is applied with an upper threshold of 5σ.

for the contours to be applied at a constant surface brightness across each individual tile. Contour levels were applied to each tile with values of 0.5, 1, 1.5 and 2σ above the clipped-mean, with a smoothing of 3 pixels, which accounted for the $5.8''$ beam size of MAGPIS, given the $2''$ pixel scale (see Fig. 3.7). This enabled a systematic boundary of each region to be identified, whilst accounting for the varying fluxes and noise. Visual inspection of the contouring procedure showed that the 1σ contour was most effectively capturing the 'edge' of the HII regions in the MAGPIS data. Both the 1.5 and 2σ levels frequently missed some of the extended emission features and were concentrated around the brightest pixels (this can be seen for the 2σ level in Fig. 3.7). The 0.5σ level extended out into the image noise for many of the regions. Hence, the 1σ contour levels were selected to describe the shape of each HII region. Further testing of how different sigma levels affect the methodology and results are detailed in Chap. 4.

The final sample of HII region shapes were obtained from the 1σ contoured MAGPIS image tiles. Since this was the first data set considered to test this shape analysis method, the selection criteria was more stringent. HII regions that were at the edge of image tiles were discarded, as were HII regions where the contouring procedure did not generate at least one closed boundary around them, usually due to noise issues. This resulted in a sample of $n = 76$ HII regions, comprising different shapes and sizes. Further details of the sample are given in Chap. 4. The contouring

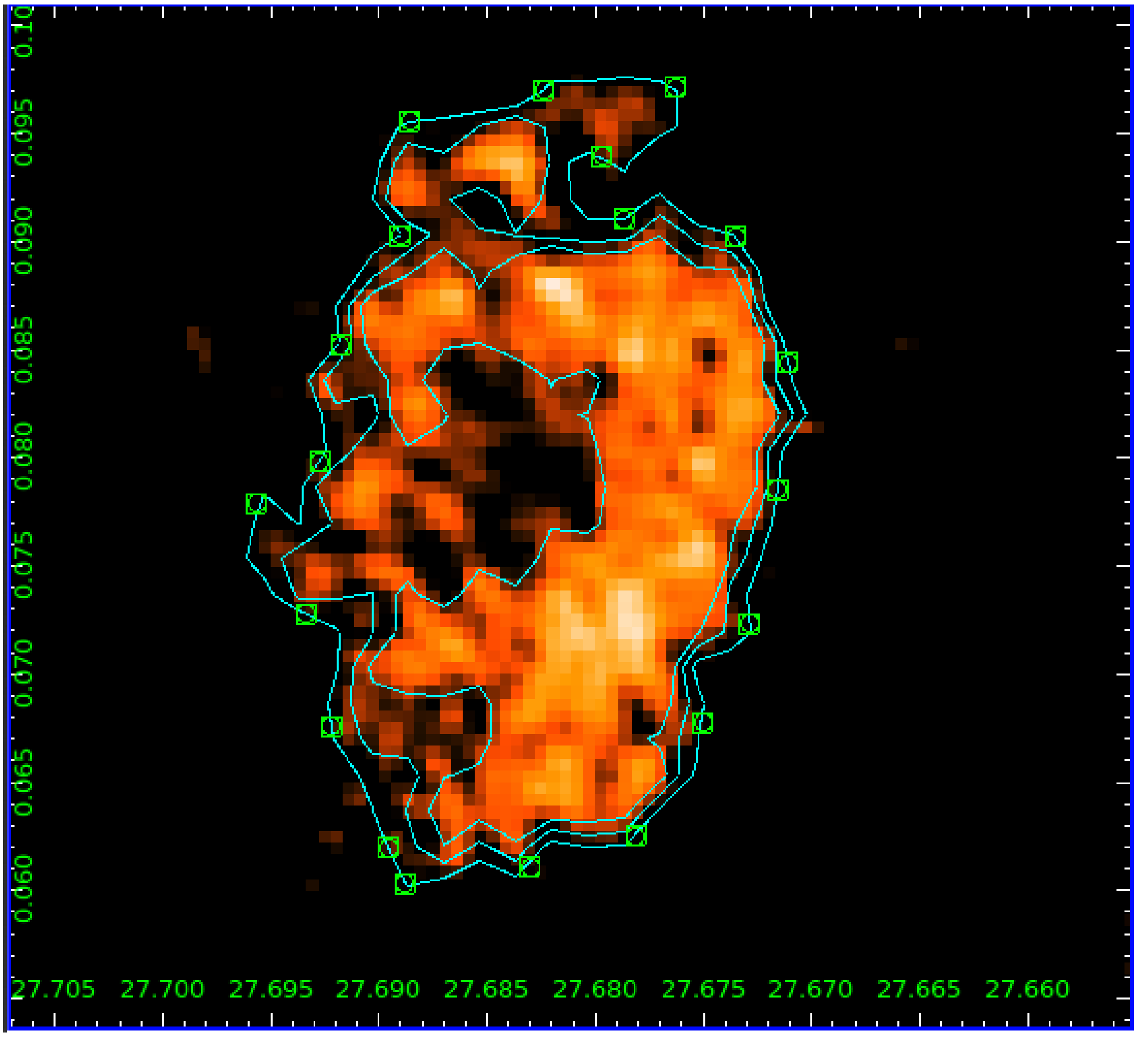


Figure 3.7: Example of the shape extraction from one of the MAGPIS HII regions. Contours are shown in cyan with values of 1, 1.5 and 2σ above the sigma clipped mean (outer to inner, respectively). Interpolation spline knots are shown as the green points along the 1σ contour.

procedure produced a set of l, b Galactic coordinates for the boundary of each HII region, with a coordinate pair at each pixel value. After translating the centre of each set to the origin of the coordinate axes, each HII region's heliocentric distance was then used to change from angular Galactic degrees to spatial parsec coordinates (see Sect. 2.4.3 for the details of these distances). Since all HII regions were within ±0.8° of the Galactic Plane, the correction factor to account for the spherical coordinate system was neglected. The next subsection explains how the pseudo landmarks were assigned along these spatial boundaries using cubic interpolation splines.

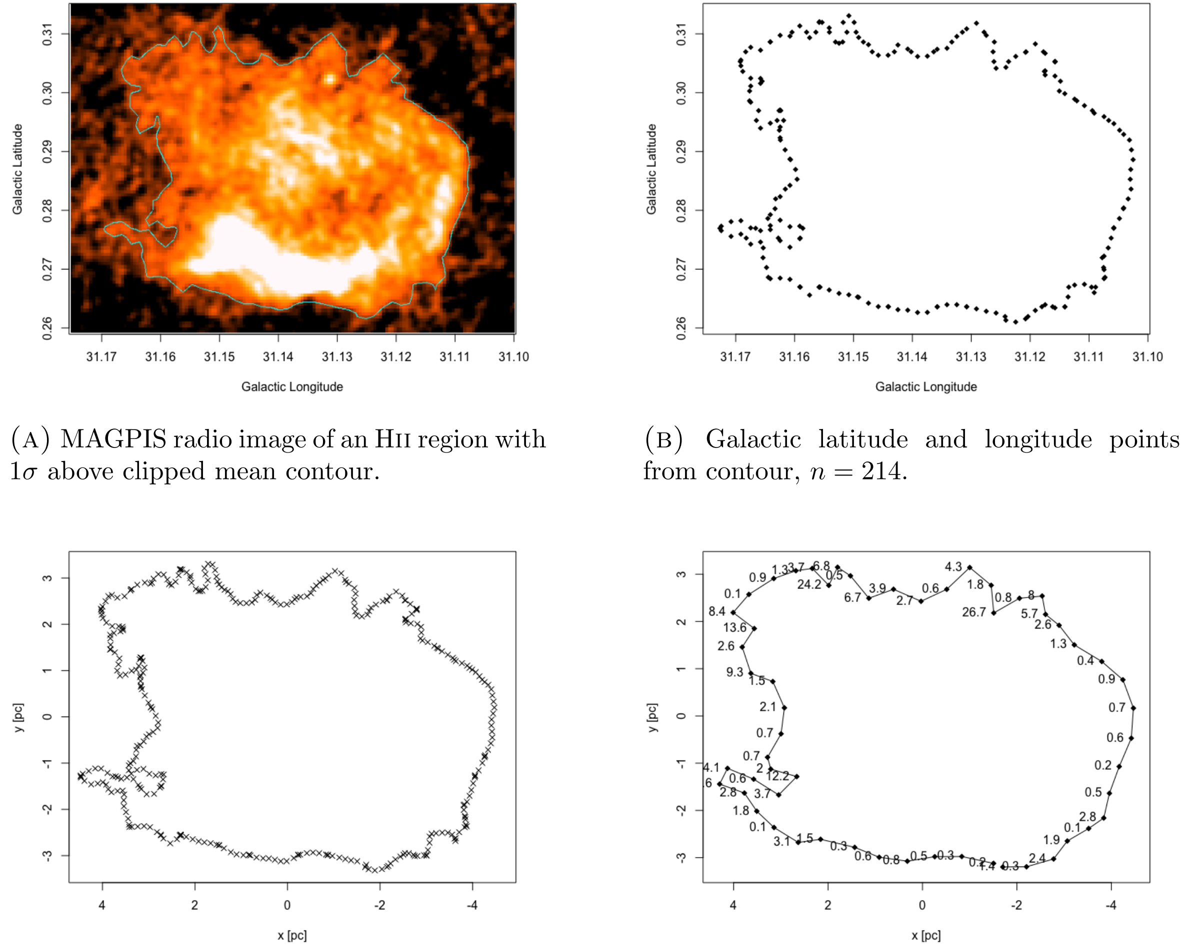
3.1.4 Curvature Shape Descriptor

In order to sample the contour coordinate points to assign landmarks for shape comparison, cubic interpolation splines were fitted through the coordinate data. This hence allowed for the shape-descriptor, used to compare the shapes of our sample, to be defined. The shape-descriptor used was the local curvature, which is defined as the reciprocal of the radius, hence this descriptor preserves the object's size-and-shape invariance. By taking the boundary of each region as $f(t) = x(t)y(t)$, where t is the spatial interval between each (x, y) Cartesian coordinate pair, the curvature, $k(t)$, was then calculated along the boundary, $f(t)$, by the following equation, where the primes denote the first and second derivatives with respect to t.

$$k(t) = \frac{|x'y'' - x''y'|}{(x'^2 + y'^2)^{\frac{3}{2}}} \tag{3.1}$$

These derivatives were obtained directly from the cubic interpolation splines that were fitted to the data, providing $f(t)$. Cubic splines determine the derivatives of each coordinate independently with respect to the spatial interval, t; they are also known as smoothing splines, since the condition for fitting is the first and second derivatives are continuous. Natural splines assume that the second derivatives are zero at the endpoints, which can result in a loss of information. Therefore, the method of Forsythe et al. (1977) was chosen for the spline fitting. This method determines the end point derivatives from the fist and last four points of the curve, thus maintaining a smooth fit throughout.

Figure 3.8 shows an example MAGPIS Hii region with the corresponding points ($n = 214$) extracted from the contouring procedure. We see from panel (a) that the sampling along the boundary is not at a constant spatial interval, with more points placed closer together at some regions of high curvature. In order for the spline points to be considered as the pseudo landmarks for the shape-space, the spline points needed to be placed as close to equidistant as possible. This was achieved via consecutive spline applications. Panel (c) shows the first spline run of the input data, with $n = 350$ resulting spline points (also known as 'knots'). The spatial sampling of the input data for this first spline pass uses 214 time-steps, which corresponds to the number of coordinate pairs in the input. This resultant oversampling of the input data has the same uneven spacing as the original data-points. From these knots, however, the spatial distance between each consecutive knot was calculated, and then used as the time-step interval for the second spline pass shown in panel (d). This resulted in the approximately equally spaced knots shown in this panel. The



(A) MAGPIS radio image of an HII region with 1σ above clipped mean contour.

(B) Galactic latitude and longitude points from contour, $n = 214$.

(C) Points from first interpolation spline fitting through contour, $n = 350$

(D) Points from second spline fitting with approximately 0.54 pc spaced spline points. Calculated curvature, $k\,[\mathrm{pc}^{-1}]$, values are also shown at each point.

Figure 3.8: Example of the spline fitting procedure to an HII region from the MAGPIS sample.

corresponding k values for each knot, calculated from Eq. 3.1, are also shown.

The points for interpolation (spline knots) could then be set at a given spatial resolution, whereby the derivatives, and thus the curvature, were determined at each knot. This allowed for each region's shape to be represented by its corresponding curvature distribution, $k(t)$. The mean knot interval for the observational data was set to 0.54 pc, which corresponds to the spatial beam size of the MAGPIS survey at a distance of 19.2 kpc - the lower error bound of the heliocentric distance to the farthest object in the sample. This limiting sampling resolution aimed to ensure that no bias was given to objects at a nearer distance, where the spatial resolution of the images is higher. For such objects where the initial number of data points in $f(t)$ was far greater than the number of spline knots, the boundaries were notably under-sampled, leading to the small amount of variation in the 0.54 pc knot intervals. The resulting knot intervals were normally distributed with a mean value of 0.54 pc and a standard deviation of 0.16 pc. Resulting effects from changing this knot interval depend upon whether or not the overall description of the object's shape changes, and are discussed in more detail in Sect. 4.3.1. Figure 3.9 shows an example of two different shape boundaries with centred, spatial coordinates (*left*), with the corresponding empirical cumulative distribution functions (EDF) for the curvature at each knot (*right*). The use of EDFs for summarising the curvature distributions allows for an unbiased, consistent estimator of the population, with no loss of information due to binning of data; this is explained further in the next section.

Having described in detail how the H_{II} region shapes are defined and extracted, we now must consider how they are systematically compared and grouped. The next sections outline the various measures that are employed in the following results chapters.

3.2 Statistical Distance Measures

Astronomers' knowledge of the Cosmos and its constituents is usually limited by having few observational parameters that give only a glimpse of the physical underlying conditions. Astrophysical understanding often involves the modelling and distribution of data with the goal of finding relationships between the observational parameters. Statistical measures give us the tools by which to sample enormous astronomical time-scales that should not be feasible given the human lifetime. Re-

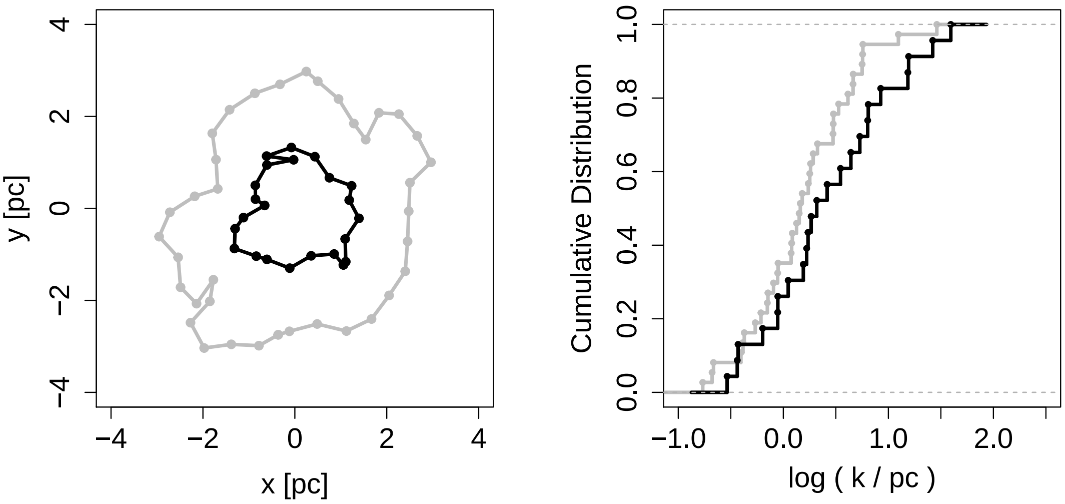


Figure 3.9: Left: Boundaries of HII regions G038.840+00.495 (grey) and G018.152+00.090 (black) with centred, spatial coordinates. Points indicate interpolation spline 'knots', where the curvature, k, was calculated. Right: Corresponding empirical cumulative distribution functions (EDFs) for the curvature values along the boundary of each HII region.

garding these inherent difficulties intrinsic to astronomy, particular statistical distributions and relationships between observed variables are sometimes assumed with little or no basis. For example, the logarithm of a variable is frequently taken to reduce ranges and remove units and normality is assumed. However, few astrophysical theories can actually predict whether the distribution is indeed Gaussian under such logarithmic, exponential or other transformations of the variables.

The use of **nonparametric statistical measures**, which make no assumptions of the underlying probability distribution, are hence of great importance for astronomers and our understanding of relationships between observables. A well known example of a nonparametric test within astronomy is the **Kolmogorov-Smirnov** (K-S) two-sample test (Kolmogorov, 1933; Smirnov, 1939; Dodge, 2008). Details of how this and the more robust **Anderson-Darling** test are given in this section. These statistical measures are employed in this work to systematically compare the shapes of the HII regions via their curvature distribution functions, with no assumptions made about the data nor its distribution. In this sense, the statistical (shape) 'distance' nomenclature, is the quantitative result from the utilised test statistic, and should not be confused with the spatial distance to the HII regions.

3.2.1 Kolmogorov-Smirnov Test

Before formally introducing the two-sample Kolmogorov-Smirnov (K-S) test for comparing distributions of data, let us first revisit the empirical distribution functions mentioned in the last section; which in this work are used to describe the curvature (and hence shape) distribution of each HII region. Explicitly, the empirical distribution function (EDF) is the simplest and most direct nonparametric estimator of the underlying cumulative distribution function (CDF) of the overall population. The dataset within the EDF $(k_1, k_2, ...k_n)$ is assumed to be drawn as independent and identically distributed samples from a common distribution function, the CDF. This makes sense for the curvature values within the EDF representing each HII region, since they were sampled at a given spatial resolution from a continuous curvature distribution along the boundaries.

$$X_n(k) = \frac{1}{n}\sum_{i=1}^{n} I[k_i \leqslant k] \tag{3.2}$$

Equation 3.2 shows the EDF X_n for the variable k and reads as: for a given value of k, the EDF is the number of elements in the sample $\leqslant k$, divided by the number of elements in the sample. The EDF ranges from 0.0 to 1.0 with step heights of $1/n$ located at each of the values k_i. This hence explains why the use of an EDF for representing distributions is far more robust than a histogram, since there are no arbitrary choices of bin widths or origins, which lead to loss of information in the bins. One can argue that since we are more accustomed to viewing histograms, it may be easier to get a 'feel' for the distribution of the data they represent (i.e. its mean and spread if it is Gaussian), however, when it comes to statistically comparing the data, the EDF is by far the better choice.

The K-S test is the most commonly used statistical distance measure (supremum) for comparing EDFs (in the two-sample case) or to compare a given EDF with a particular model specified in advance (the one-sample case). The two-sided K-S test is given by the following equation:

$$M_{KS} = \max_x |X_{(n)}(k) - Y_{(m)}(k)| \tag{3.3}$$

A large value of the supremum, M_{KS} would allow for rejection of the null hypothesis that the EDFs X and Y are from the same parent distribution. If the EDFs are identical, $M_{KS} = 0$.

The two-sided K-S test is distribution free (non-parametric), meaning that it does not assume normality or any particular distribution of the EDFs[2]. There is no restriction on the size of the two samples, nor must they be equally sized. Critical values of probabilities and significance levels are also widely available and easily computed from the data. It is therefore easy to see why the K-S test is used in hundreds of astrophysical papers every year, due to its many advantages. However, the K-S test does suffer from some noteworthy limitations. One being it is not actually *that* sensitive to differences between EDFs, since it only considers the maximum deviation. Furthermore, due to convergence at low and high values, it is also insensitive to these differences for two EDFs. The next statistical distance measure introduced, aims to address each of these issues, providing a more thorough distance estimator between two samples.

3.2.2 Anderson-Darling Test

The two-sided Anderson-Darling (A-D) test statistic (Anderson and Darling, 1952) is a weighted variant of the Cramér-von Mises (CvM) statistic (Cramér, 1928). The CvM is an extension of the K-S test, such that it compares two EDFs, but instead of just considering the supremum the two distributions, it measures the sum of the squared differences between each EDF. It is therefore sensitive to both local and global differences. As with the K-S test, however, it still suffers from the convergence of the EDF at low and high values to 0 and 1, respectively. The A-D test accounts for this shortcoming by appropriately weighting the test statistic to the tails of the distributions, whilst also considering the sum of squared differences. The two-sample A-D test, T_{AD} (Pettitt, 1976), generalises to the following formula:

$$T_{AD} = \frac{1}{nm} \sum_{i=1}^{n+m} \frac{(N_i Z_{(n+m-ni)})^2}{i Z_{(n+m-i)}} \tag{3.4}$$

where $Z_{(n+m)}$ are the combined and ordered EDFs $X_{(n)}$ and $Y_{(m)}$ (for respective sample sizes n and m), and N_i are the number of observations in $X_{(n)}$ that are less than or equal to the i^{th} observation in $Z_{(n+m)}$. In this equation, the numerator deals with finding the sum of the squared differences, whilst the denominator is the weighting factor to account for the convergence at the tails. The test statistic T_{AD} represents a dissimilarity measure between the two samples, whereby the null hypothesis (that the two samples originate from the same parent sample) is rejected

[2] There is a situation where the one-sided K-S test is not distribution-free, when the model parameters are estimated or derived from the same dataset being tested. For example, when using the K-S test to test for a normal distribution, the test is not distribution free if the mean and variance of the normal distribution are specified from the data in question. Instead, one should aim to use a similar dataset or astrophysical theory to provide the model against which to test.

Table 3.1: Dissimilarity matrix of A-D pairwise test statistic scores (T_{AD}) for subsample of HII region curvature distributions.

HII Region l	10.965	12.432	17.33	17.921	18.081	18.141	18.154	18.452
10.965	0.00	0.43	1.97	2.42	4.12	1.09	2.55	1.63
12.432	0.43	0.00	1.89	3.20	5.75	1.34	3.98	2.87
17.33	1.97	1.89	0.00	1.31	1.68	2.30	1.65	1.24
17.921	2.42	3.20	1.31	0.00	0.42	3.59	0.86	0.72
18.081	4.12	5.75	1.68	0.42	0.00	4.87	0.52	1.21
18.141	1.09	1.34	2.30	3.59	4.87	0.00	3.28	2.47
18.154	2.55	3.98	1.65	0.86	0.52	3.28	0.00	0.52
18.452	1.63	2.87	1.24	0.72	1.21	2.47	0.52	0.00

for large T_{AD}, and $T_{AD} = 0$ for identical distributions.

The A-D test has the same advantages as the K-S test, such that it is distribution free, however, the A-D test cannot be applied to small samples. Furthermore, the A-D test addresses both of the issues outlined for why the K-S test is insensitive to some EDF differences and has repeatedly been shown be more effective than both the K-S test and CvM test for determining statistical distances between samples. For further information of comparisons between the A-D test and the K-S test see e.g. Babu and Feigelson (2006), Hou et al. (2009) and Engmann and Cousineau (2011).

Following from these arguments and early performance testing for the data analysed in this work, the two-sided A-D test statistic was chosen to represent the shape-distances. It was hence computed pair-wise for each HII region using the curvature EDFs. This resulted in a symmetrical $N \times N$ matrix of T_{AD} dissimilarity measures. An example arbitrary subset of this matrix for eight HII regions in our sample (ordered by Galactic longitude) is shown in Tab. 3.1.

3.3 Statistical Clustering

Now that we have our quantitative morphological distances between pairs of HII regions, we can look at how we obtain groups from the data and look for similarities. This is the topic of multivariate **clustering** and **classification**. An important terminology point to mention here is that clustering refers to situations where the subpopulations are estimated from the data alone (known as unsupervised), whereas classification requires training datasets of known populations that are available independent to the dataset under study (hence termed supervised). This point is stressed because many times when we read or hear the term classification of data, what has actually taken place is unsupervised statistical clustering. The reason for this is because the nature of classification itself is the result of characterisation and understanding. When we say something is classified, it means we have interpreted enough information from that phenomenon, such that we can link it to common objects and have a reasonable amount of knowledge about it. Clustering is hence on many occasions the first step towards this goal and the terms are used synonymously[3].

Astronomical examples of where the term 'classification' was used synonymously with 'clustering' include: Edwin Hubble's 'classification' scheme of galaxy morphologies - mentioned earlier in this chapter; the many classes of variable stars which are often named after a bright prototype (e.g. RR Lyr (Shapley, 1916), FUors (Herbig, 1989)); types of active galactic nuclei with multiple overlapping classes (e.g. Fanaroff-Riley radio galaxies, Seyfert galaxies, BL Lac objects, quasars, to name a few from Urry and Padovani (1995)). In many cases, the unsupervised clustering of objects has led directly to a supervised classification scheme, whereby discriminatory rules are applied to the training set in order to group the objects based on physically or theoretically defined common classes. Examples of this include supernovae events, which are classified by their optical spectra into e.g. Type Ia, Ib, Ic, II, etc (Filippenko, 1997); and protostars that are classified by their spectral energy distributions into Class 0, I, II and II (Adams et al., 1987).

The need for automated clustering and classification tools in astronomy has become increasingly important in the era of big-data, with wide field surveys now ubiquitous and the amount of data literally and figuratively astronomical. In this section, the different methods for performing statistical clustering are explained and examples of each are provided. The hierarchical clustering method employed to identify groups of HII regions sharing common morphologies is then explained in full.

3.3.1 Partition Clustering

There are two basic types of statistical clustering: **partition** and **hierarchical**. These each, respectively, take a top-down or and bottom-up approach[4]. A partitioning method constructs a number of clusters (groups) from the data set, which

[3] I am guilty of this myself, by entitling research talks 'classification of HII regions', when actually meaning clustering, although I refrained from this for my thesis and journal publications.

[4] This is true when speaking of agglomerative hierarchical clustering, divisive hierarchical clustering also takes a top-down approach, but for reasons outlined in the next subsection, is generally not used when dealing with hierarchical clustering.

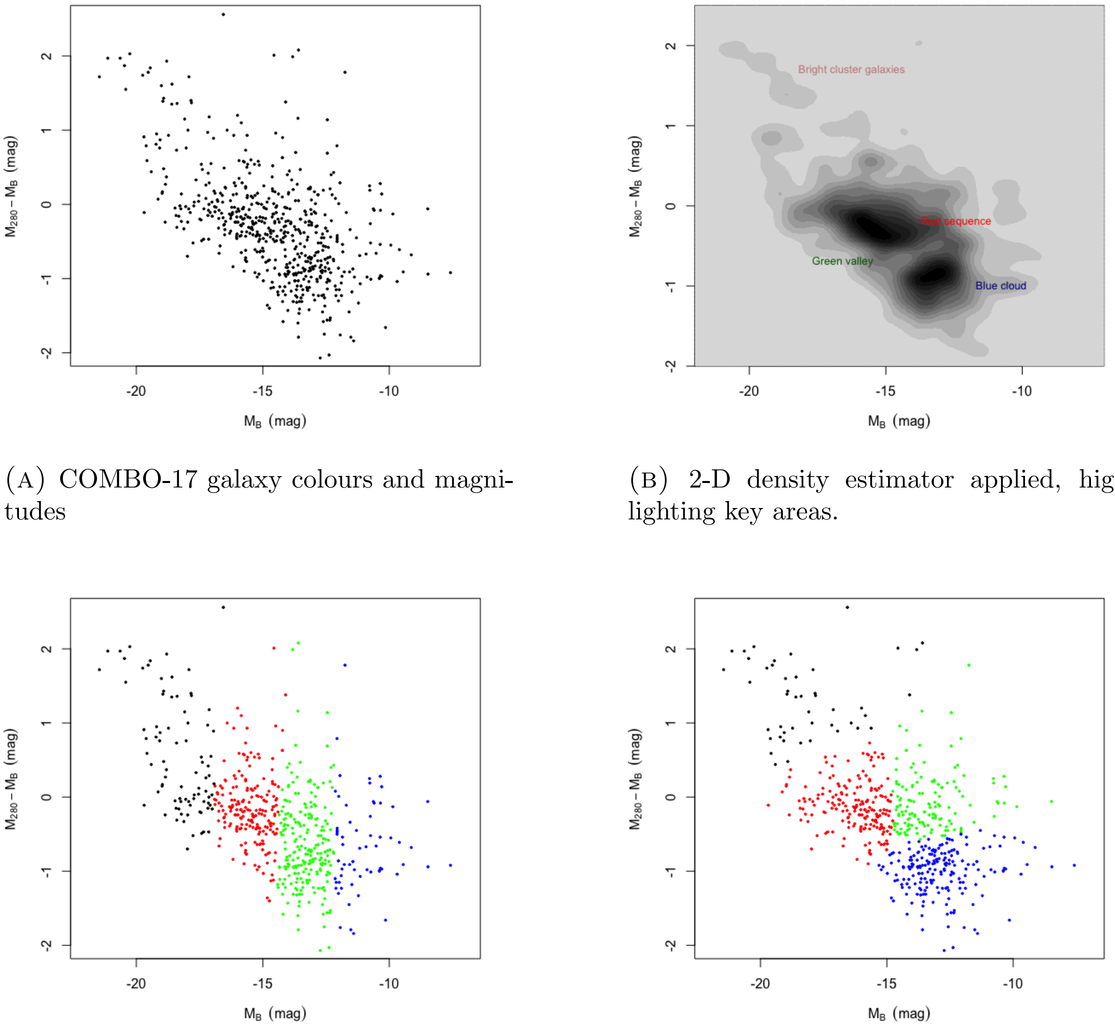
together must satisfy the conditions that each group contains at least one object and each object must belong to exactly one group (Kaufman and Rousseeuw, 2009). The number of groups must hence be specified before running the algorithm and is usually repeated with different set number of groups to investigate the resulting clusters and associations. The way in which the groups are constructed relates to the type of partition clustering used. The most common of which is known as k-means clustering (such that 'k' refers to the number of groups, however, this labelling will not be adopted throughout, as to avoid confusion with the curvature values).

To perform k-means clustering, one starts with the k seed locations which represent the group/cluster centroids. Each data point is then iteratively assigned to the nearest cluster to reduce the sum of within-cluster squared distances. The centroids of each cluster are thus recalculated after every iteration by the following equation:

$$\bar{x}_j' = \frac{n_j \bar{x}_j + x_i}{n_j + 1} \tag{3.5}$$

Where $\bar{x}_j'$ is the updated centroid position of the j-th cluster with the addition of a new data point x_i, and n_j is the current population of the cluster. Unlike for the hierarchical methods, that we will see next, the pairwise distance matrix for the data does not have to be calculated and stored beforehand when performing k-means clustering. Convergence of the cluster centroids is usually rapid, however, there is no guarantee that the optimum solution is achieved. To demonstrate this, the example used by Feigelson and Babu (2012) of the COMBO-17 (Classifying Objects by Medium-Band Observations in 17 Filters) photometric survey of galaxies (Wolf et al., 2003) has been reproduced here.

Figure 3.10a shows the photometric COMBO-17 data for 517 galaxies, with their magnitude in the blue band (M_B) plotted against the ultraviolet-to-blue colour index ($M_{280} - M_B$). In Fig. 3.10b, a two-dimensional density estimator has been applied to smooth the data, the resulting structure shows the unsupervised cluster groups termed 'blue cloud', 'green valley' and 'red sequence', along with the outlier bright cluster galaxies (for details of this clustering example, see Bell et al., 2004). Figure 3.10c shows the result of k-means clustering on this dataset, with $k = 4$ groups shown by different colours. We can see that the manner in which the groups are delineated does not match up well to the expected groupings from the smoothed figure. This is explained by the range of values on each axis, with the larger range belonging to the blue band magnitudes. Hence, the delineations in the groups is only along this axis to reduce the within-cluster distances. Figure 3.10d shows the results of k-means clustering on the centred and scaled COMBO-17 data. That is, the data on each axis is centred at zero and has its values scaled by the standard deviation



(A) COMBO-17 galaxy colours and magnitudes

(B) 2-D density estimator applied, highlighting key areas.

(C) Application of k-means clustering to data with k=4 clusters.

(D) Application of k-means clustering to scaled data with k=4 clusters.

Figure 3.10: Example of applying k-means partition clustering to the COMBO-17 galaxy magnitude data.

of the respective data. The resulting clusters are now delineated in both axes and are somewhat closer to representing the clusters outlined in the smoothed data. Whilst this is not the optimum arrangement as shown in the smoothed density plot in Fig. 3.10b, k-means clustering does allow for fast delineation and investigation of potentially correlated properties. It is therefore commonly used in astronomy as an automated tool for clustering large datasets and as a way of training machine learning algorithms (e.g. Sánchez Almeida and Allende Prieto, 2013; Garcia-Dias et al., 2018).

Limitations of k-means clustering include: as mentioned, having to predetermine the number of clusters, k. This is easily overcome by running the procedure with different values of k, and then selecting the value that gives the smallest sum of within-cluster variances. A second limitation is that the initial seed locations can influence the result. This is more profound in larger datasets and is therefore common to compare the results from a number of random initial seed locations. However, in cases with extremely large datasets, it may be unfeasible to try enough initial conditions to ensure that optimum within-cluster variance has been found. A possible solution to this is to use median values rather than mean values for determining the centroid positions. This is known as k-medoid partitioning, where the summed distances, rather than the squared differences, are minimised. A further advantage of this method is that it identifies a prototypical group member for each group (in k-means clustering, the centroid does not necessarily have to be located at the location of an object in the group). A recent study of globular star clusters used this method, whereby the representative star clusters were obtained for different areas of the Galaxy (Pasquato and Chung, 2019).

3.3.2 Hierarchical Clustering

Hierarchical algorithms for unsupervised clustering do not create a single partition with k clusters as with partition clustering. Instead, all possible values of k, including $k = n$, are obtained in one application. There are two types of hierarchical clustering: **agglomerative** and **divisive**, which construct their hierarchy in opposite ways. Agglomerative clustering starts with $k = n$ clusters, such that each individual data point is its own cluster. It then merges objects into groups and iteratively grows the groups, according the type of agglomeration method used. Divisive methods, on the other hand, start with all objects grouped together and in each iteration, splits the data into subgroups (hence differentiating it from partition clustering, which finds groups based upon a predetermined number of groups). One might wonder: why bother using partition clustering with a predetermined number of groups, when one could use divisive hierarchical clustering and explore all possible groups? This actually highlights the main difference between the two clustering methods, such that hierarchical methods form their clusters 'along the way', meaning that once an object is assigned to a cluster, that is its final assignment. Partition clustering, however, allows for redetermining of group membership, based upon the minimising cluster centroid methodology. It is therefore important when using hierarchical clustering to understand how groups are iteratively formed, which can be accomplished several different ways. Whilst hierarchical and partition clustering both aim to find groups in multivariate data, their specific goals are slightly different. Partition clustering aims to find the 'best' groups and hierarchical clustering allows for investigation of connected structure within the groups, and is hence often used by biologists and taxonomists for finding evolutionary structure.

A further reason for partition clustering still being used in favour of *divisive* hierarchical clustering is that divisive procedures have historically not been utilised as much, due to high computational expense. Since the first step of the algorithm would be to consider all possible divisions of the data into two subsets, this becomes infeasible with a large number of combinations. In fact, there is only one refereed astronomical journal article on ADS that mentions 'divisive hierarchical' in the title or abstract (Dumitrescu et al., 1997). However, ∼2000 articles include the terms 'hierarchical clustering' and, in nearly all cases, this would be referring to an agglomerative method. Therefore, for the remainder of this section, only *agglomerative* hierarchical clustering is further considered.

The agglomerative algorithm most familiar to astronomers is referred to as the 'friends-of-friends' or 'nearest-neighbour' algorithm, popularised in the 1970s by the search for galaxy groups in redshift surveys. This method is formally known as the **single-linkage** method and is the oldest of the hierarchical clustering methods (Florek et al., 1951). The agglomerative method determines how merges take place as objects are first paired and then groups grow, until the final two groups are merged. The result is commonly plotted on a **dendrogram** (or tree diagram). In order to perform agglomerative clustering, a Euclidean distance matrix is formed (Eq. 4.1) from the variables, which in this work, is the dissimilarity matrix of T_{AD} test scores:

$$D_{xy} = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2} \tag{3.6}$$

where x and y are the vectors of variables or dissimilarities between the respective objects. For example, when applying this to the HII region dissimilarity scores, D_{14}

Table 3.2: Euclidean distance matrix for the T_{AD} dissimilarity scores in Table 3.1

HII Region l	10.965	12.432	17.33	17.921	18.081	18.141	18.154	18.452
10.965	0.00	2.69	4.42	6.59	9.22	2.55	6.87	5.41
12.432	2.69	0.00	6.22	8.52	10.71	2.40	8.85	7.64
17.33	4.42	6.22	0.00	3.08	5.82	5.59	3.64	2.46
17.921	6.59	8.52	3.08	0.00	3.45	7.79	1.55	1.94
18.081	9.22	10.71	5.82	3.45	0.00	9.78	3.05	4.84
18.141	2.55	2.40	5.59	7.79	9.78	0.00	7.83	6.73
18.154	6.87	8.85	3.64	1.55	3.05	7.83	0.00	1.99
18.452	5.41	7.64	2.46	1.94	4.84	6.73	1.99	0.00

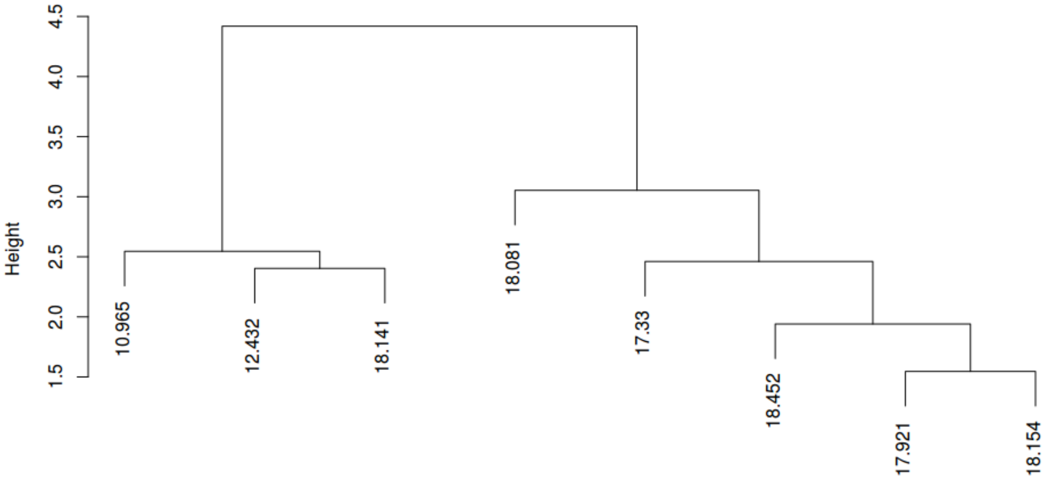


Figure 3.11: Dendrogram from applying single-linkage hierarchical clustering to the distance matrix in Table 3.2. Vertical axis shows height obtained from distance matrix for pairs; and from single-linkage method for subsequent merges.

is the Euclidean distance between regions 1 and 4, which accounts for the dissimilarities of each region with respect to all other regions. This allows for groups to be more accurately determined as outlier distances are highlighted in Euclidean space. The resulting distance matrix is again symmetric, with $D_{x=y} = 0$. The Euclidean distance matrix for the subsample of eight HII regions, formed from their A-D test scores is shown in Tab. 3.2. The cluster hierarchy resulting from the agglomeration method uses these distances, in the first instance, merging the pair of objects with the lowest distance score. New group distances are then calculated at each iteration by the method specified. In **single-linkage** clustering:

$$d_{(C,x)} = \min_{d}(d_{1,x}, d_{2,x}, ..., d_{i,x}) \tag{3.7}$$

here, $d_{(C,x)}$ is the distance between cluster C (of size i) and a new object x. Hence, the distance to the closest member of the cluster is used.

Figure 3.11 shows the resulting dendrogram from applying single-linkage hierarchical clustering to the eight HII regions with Euclidean distances listed in Tab. 3.2. In this figure, height corresponds to the distance at which a merge took place. We can see that the two objects that are joined at the lowest height are objects '17.921' and '18.154', which have a distance score of 1.55. The next merge joins object '18.452' to the first pair of objects since that has the next lowest score to one of the objects in that newly formed group. We can already see from this simple example that by only considering the minimum distances at each iteration, this method can result in elongated, chained groups. Outlined next is the procedure for three further agglomerative methods. The COMBO-17 galaxy data will then be revisited to show a comparison of the four methods.

Complete-linkage clustering aims to find groups that are more symmetrical and compact by using the group-point distance measure:

$$d_{(C,x)} = \max_{d}(d_{1,x}, d_{2,x}, ..., d_{i,x}) \tag{3.8}$$

such that the furthest, rather than the closest member of the cluster is used as the new distance.

An intermediate to these two methods is the **average-linkage** clustering method:

$$d_{(C,x)} = \frac{1}{i}\sum_{n=1}^{i} d_{n,x} \tag{3.9}$$

where the mean average distance between the cluster C and the new object x is considered at each iteration.

The final agglomerative method we consider is the **Ward's minimum variance** method: 'ward.D2'[5], which implements the Ward and Joe (1963) clustering criterion (Murtagh and Legendre, 2014). To iteratively form the groups using Ward's method, each value in the input distance matrix, D, $(i, j, k,\ etc. \subset D)$ starts as an $n = 1$ group $(C_i, C_j, C_k, etc.)$. As with the previous methods, the closest pair of objects are first agglomerated into a new group. After each iteration, the new inter-group distances are redefined from the new group centres (of group sizes n), following the

[5] The 'ward.D' method yields the same clustering results when using D^2_{xy} as the input

Lange and Williams (1967) dissimilarity update formula:

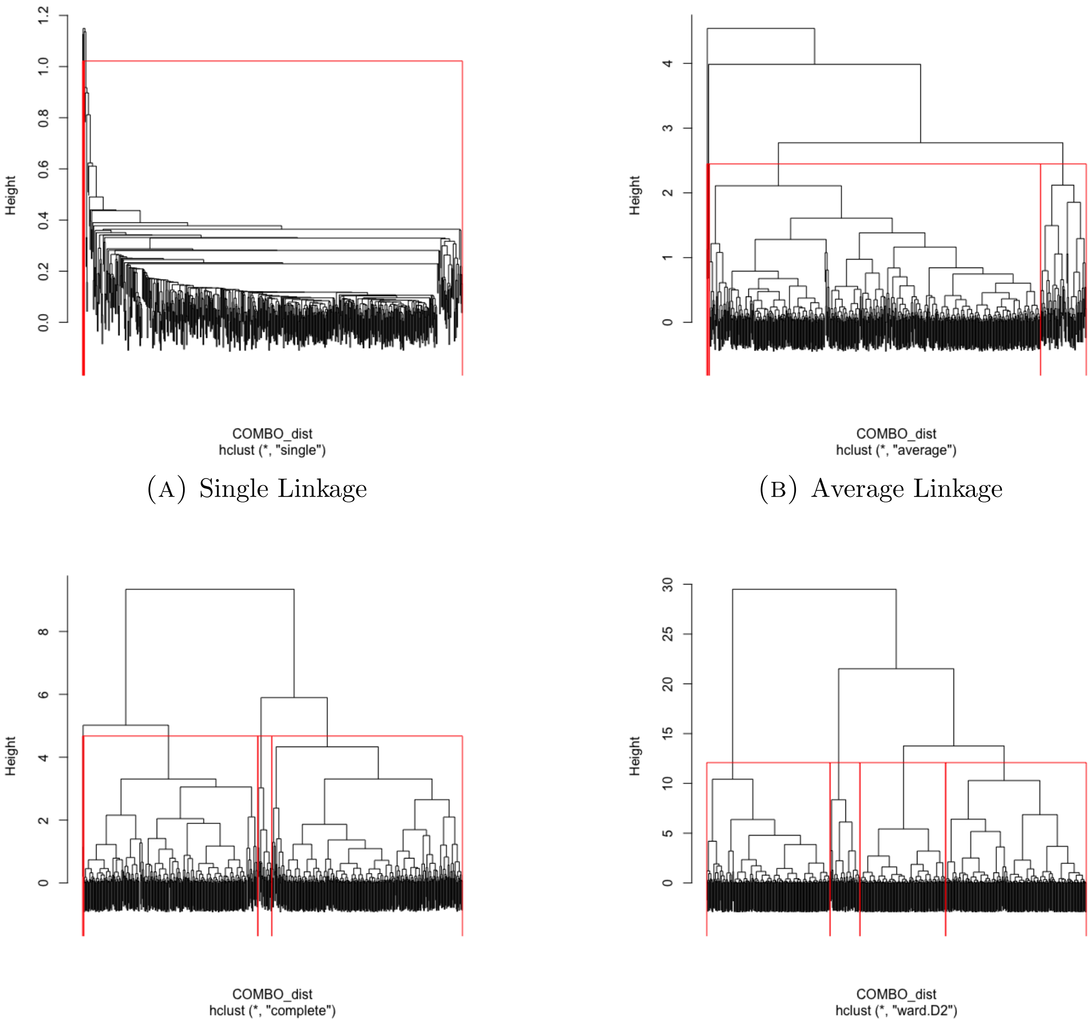
$$
\begin{aligned}
d^2(C_i \cup C_j, C_k) = \frac{n_i + n_k}{n_i + n_j + n_k} d^2(C_i, C_k) + \\
\frac{n_j + n_k}{n_i + n_j + n_k} d^2(C_j, C_k) - \frac{n_k}{n_i + n_j + n_k} d^2(C_i, C_j)
\end{aligned}
\tag{3.10}
$$

Ward's agglomerative method is based on a sum-of-squared differences criterion for producing groups. It ensures that at each iteration, the distances of the newly created group to the rest of the data is selected such that the within-group-dispersion is minimised. This dispersion is proportional to the squared Euclidean distance between group centres.

For each of the four methods outlined above, at each iteration, for k groups (where k originally equals the number of objects in the input distance matrix, and then decreases after each iteration), there are $k - 1$ agglomerations. The process is repeated until the final two groups are joined.

To investigate how each agglomerative method performs, we will return to the COMBO-17 galaxy data. A Euclidean transformation was applied to the scaled and centred blue magnitudes and ultraviolet-to-blue colour index values. Agglomerative hierarchical clustering was then applied to the distance matrix using each of the four methods outlined here. The resulting dendrograms are shown in Fig. 3.12 and the colour-magnitude diagrams with the points coloured by four groups identified by cutting the dendrograms is shown in Fig. 3.13. To cut the dendrogram into groups, one considers from the top-down the number of divisions in the tree. This is easily seen from the bottom two panels in Fig. 3.12. The top two panels, displaying single- and average-linkage show one and two (respectively) large groups, with small outlier groups on the left hand side. Three groups are identified by the complete-linkage method with one small outlier group on the left. Ward's method identifies four clear groups.

The corresponding colour-magnitude diagrams in Fig. 3.13 show that single-linkage has performed poorly in identifying cluster groups, with one large cluster and three individual outlier points located at the top and bottom of the graph. Average-linkage is able to split some of the redder galaxies from the main group (coloured green on this graph), with four outlier galaxies in two groups at the top and bottom. Complete-linkage shows more of a distinction in this respect, whilst also separating out some of the outlier bright cluster galaxies (shown in red). Ward's method is able to identify four distinct groups, in a similar fashion to the results of the k-means clustering shown earlier. The groups identified from Ward's method



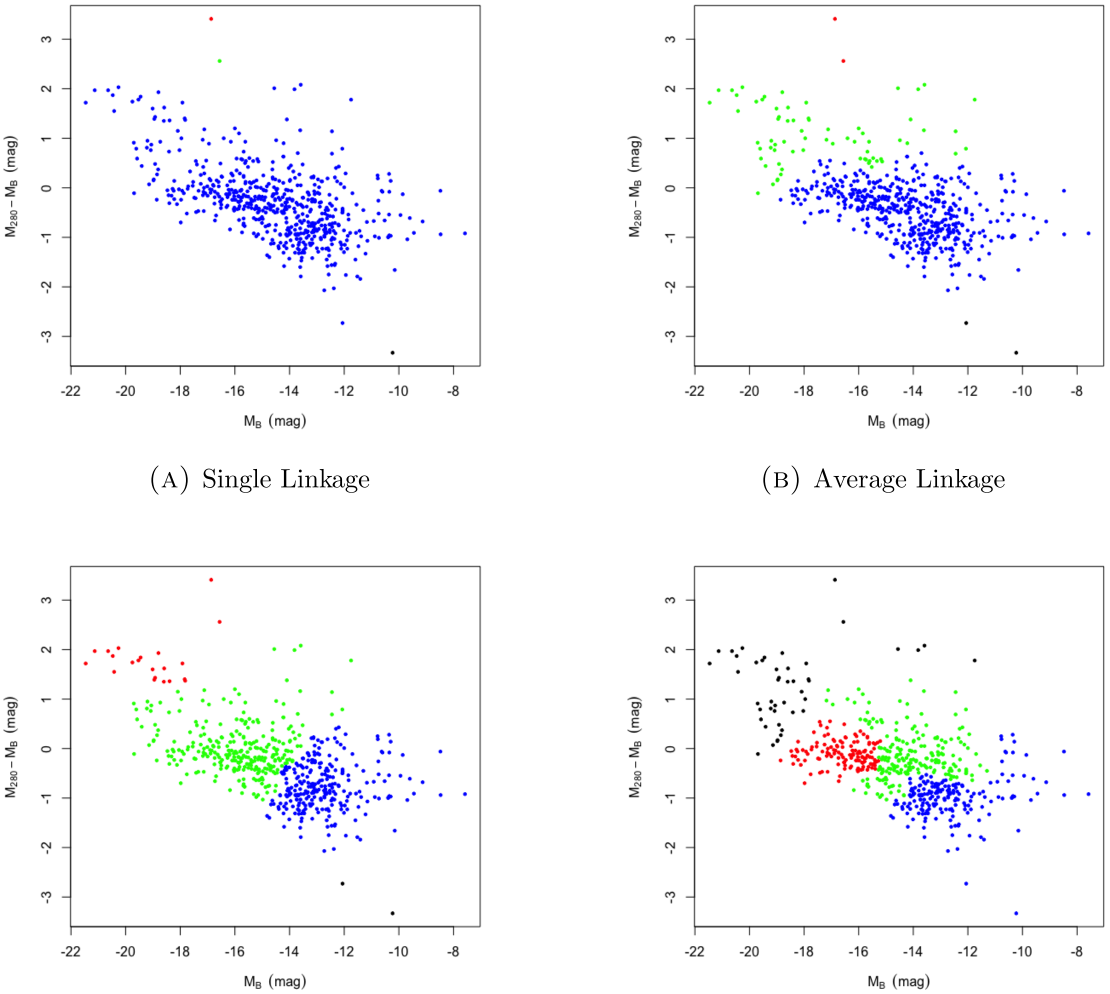
(A) Single Linkage

(B) Average Linkage

(C) Complete Linkage

(D) Ward's Minimum Variance

Figure 3.12: Example dendrograms for four different hierarchical clustering algorithms applied to the COMBO-17 colour magnitude data. The red boxes delineate the first four obtained clusters.



(A) Single Linkage

(B) Average Linkage

(C) Complete Linkage

(D) Ward's Minimum Variance

Figure 3.13: Corresponding colour-magnitude diagrams for the COMBO-17 galaxy data; with the obtained clusters from the respective methods in Figure 3.12 identified by the coloured points.

show less of a distinct separation along each axes when compared to those obtained by the k-means clustering.

The advantage of hierarchical clustering over partition clustering is the ability to determine structure from the resulting dendrogram. One can see how and when certain groups merge into larger groups with respect to others. As we can see from the dendrograms of the COMBO-17 data, the choice of agglomerative method can have a substantial influence on the determined groups. The performance of the different agglomerative methods have been extensively evaluated from simulated data and in all cases, it was found that the single-linkage method is the least successful (Everitt et al., 2001; Ferreira and Hitchcock, 2009). Ferreira and Hitchcock (2009) found that Ward's method generally outperforms other methods, especially when there are no large differences in group sizes. Following these arguments, and the examples shown in this section, Ward's minimum variance method was selected as the statistical clustering method to use in this investigation. The resulting groups identified from its application to the HII region shape data are fully explored in Chapters 4 & 5.

The following chapter details the application of the shape extraction, statistical distance measure and unsupervised clustering methods outlined here to the MAGPIS observational HII region data.

Chapter 4

Radio Continuum Observations

This chapter details the application of the shape analysis and statistical clustering for a selection of Galactic HII regions from the MAGPIS survey. Many of the results and discussion herein are published by the thesis author in Campbell-White et al. (2018).

4.1 Introduction

This chapter covers the analysis and discussion of the high resolution MAGPIS image data of HII regions (detailed in Chap. 2) using statistical shape analysis (detailed in Chap. 3). In the literature thus far, radio continuum images of HII regions have been observed to share a common morphology of a closed disk (Fig. 4.1, left), which is inferred as a projection of the extended Strömgren sphere around an ionising source(s) (e.g Mezger and Henderson 1967; Beltrametti et al. 1982; Franco et al. 2000; Quireza et al. 2006). Similarly, mid-infrared (MIR) bubbles, the larger of which are all confirmed to be the result of HII regions (Anderson et al., 2011; Deharveng et al., 2010), have been characterised by their common morphologies of ring like structures (Fig. 4.1, right), and studies have catalogued their radii, eccentricities and shell thickness's (e.g. Churchwell et al. 2006, 2007; Beaumont and Williams 2010; Arce et al. 2011; Simpson et al. 2012). In each of these cases, inhomogeneities

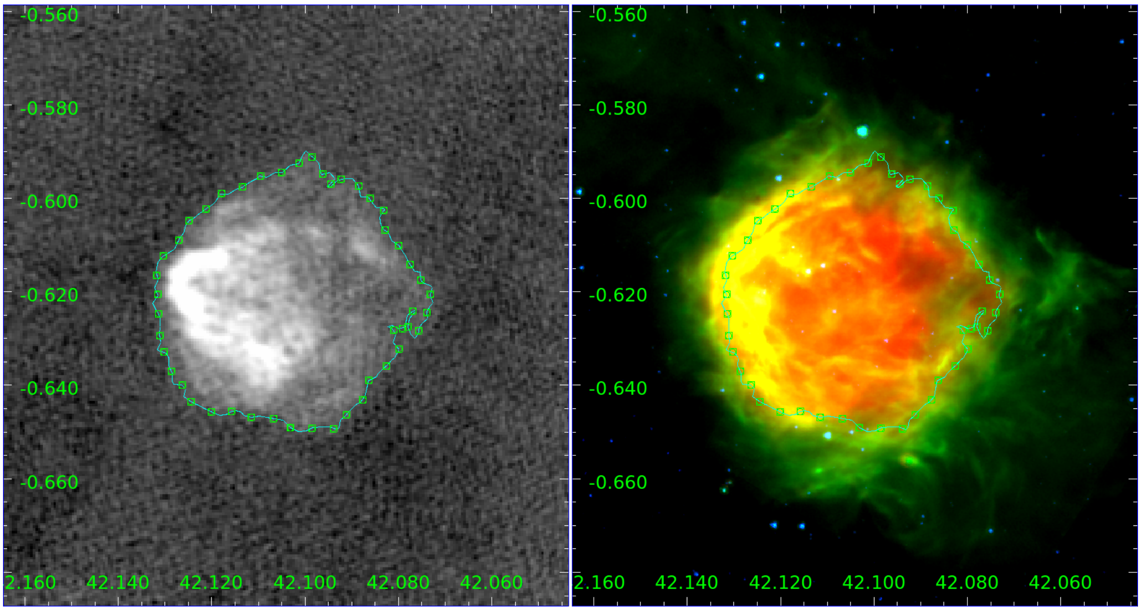


Figure 4.1: Hii region G042.103-00.623. Left: MAGPIS 20 cm radio continuum image with contoured boundary. Right: False colour MIR composite of *Spitzer* GLIMPSE 4.5 μm (blue), 8 μm (green) and MIPSGAL 24 μm (red). The same contour from the extracted Hii region radio shape is over-plotted.

in the natal molecular clouds can lead to perturbations from these ideal morphologies. The example Hii region shown in Fig. 4.1 is indeed a spherically symmetric, unperturbed example, however, this is not the norm for these regions. The full sample in Appendix B shows many of the inhomogeneities, elongations and asymmetries. If a link can be shown between the shape of the Hii regions and the physical conditions, this can be used to better understand the affinity between massive stars, their formation and the surrounding ISM.

In terms of statistically clustering and classifying Hii regions and bubbles via their observed shape, a broad morphological scheme was introduced by Churchwell et al. (2006, 2007, hereafter C06), who first studied stellar bubble images from the GLIMPSE data (Sect. 2.4.2). They considered almost 600 MIR bubbles along the Galactic plane by the shape of their shells, e.g. closed, broken, group. This approach was also used by Anderson et al. (2011) and Bania et al. (2012) for the Green Bank and Arecibo Hii Region Discovery Survey catalogues, respectively. Despite the fact that only half of all Hii regions are observed as MIR bubbles (Anderson et al., 2011), it is easier to distinguish features by eye in the MIR shells than the radio continuum disks. However, the radio continuum data leads itself to automised shape extraction more readily, not suffering from point source contamination or filamentary structure. More recently, Anderson et al. (2014) categorised over 8000 regions in the WISE Catalog of Galactic HII Regions (Sect. 2.4.3) based on the presence or absence of radio emission, and do not include any morphological flags (MIR or otherwise).

H II regions are known to trace the locations of active star formation (SF) (e.g. Thompson et al. 2012; Kendrew et al. 2012). Due to the clustered nature of massive SF, insight into the physical properties of H II regions allow for inferences to be made as to the initial conditions of massive star cluster formation. In this work, the homogeneity of the high resolution radio images available of H II regions is utilised, to systematically compare their shapes. As explained in Chap. 3, mathematical shape provides an unbiased, intrinsic characteristic of an object, which can be readily compared through statistical means. The aim of this work is to find associations between the region shapes and physical parameters/initial conditions of star cluster formation to ultimately produce a training set for supervised classification of H II regions.

Before beginning the analysis of the radio data, we will first have a brief summary of the H II region data, the corresponding shape extraction procedure and the statistical measures applied:

The H II region data were taken from the 20 cm (1.4 GHz) Multi-Array Galactic Plane Imaging Survey (MAGPIS Helfand et al., 2006). MAGPIS combined VLA images with those from the 100m Effelsberg telescope to correct for missing fluxes in the extended emission regions (full details in Sect. 2.4.1). The resulting MAGPIS images have an average angular resolution of $5.8''$, with a pixel scale of $2''$ and 1σ sensitivity of < 0.15 mJy. MAGPIS covers $|b| < 0.8^\circ$ and $48.5^\circ > l > 5^\circ$. Within this range, there are 710 'Known' WISE H II regions, 405 of which have a determined distance in the WISE catalogue and 243 are positionally coincident with at least one MWP large bubble[1].

The morphologies of this sample of H II regions at a given surface brightness were then extracted using an image contouring procedure (full details in Sect. 3.1.3). To generate contours from each region, the signal value was first determined by applying sigma clipping to each image tile. Different contour levels were then applied, with visual inspection of the procedure showing that the 1σ above the clipped mean level most effectively captured the 'edge' of the H II regions. Further testing of how different sigma levels affect the methodology and results are detailed later in this chapter. Regions that were at the edge of image tiles were discarded, as were ones where the contouring procedure did not generate at least one closed boundary around the region. A sample of $n = 76$ H II regions was thus established, comprising different shapes and sizes. Distances to these regions were taken from the WISE

[1] Matching the H II regions with the large MIR bubbles ensured that the H II regions would be of a sufficient size to carry out the shape analysis. It would also allow for the radio shapes of this sample to later be compared to the MIR shapes, in future work

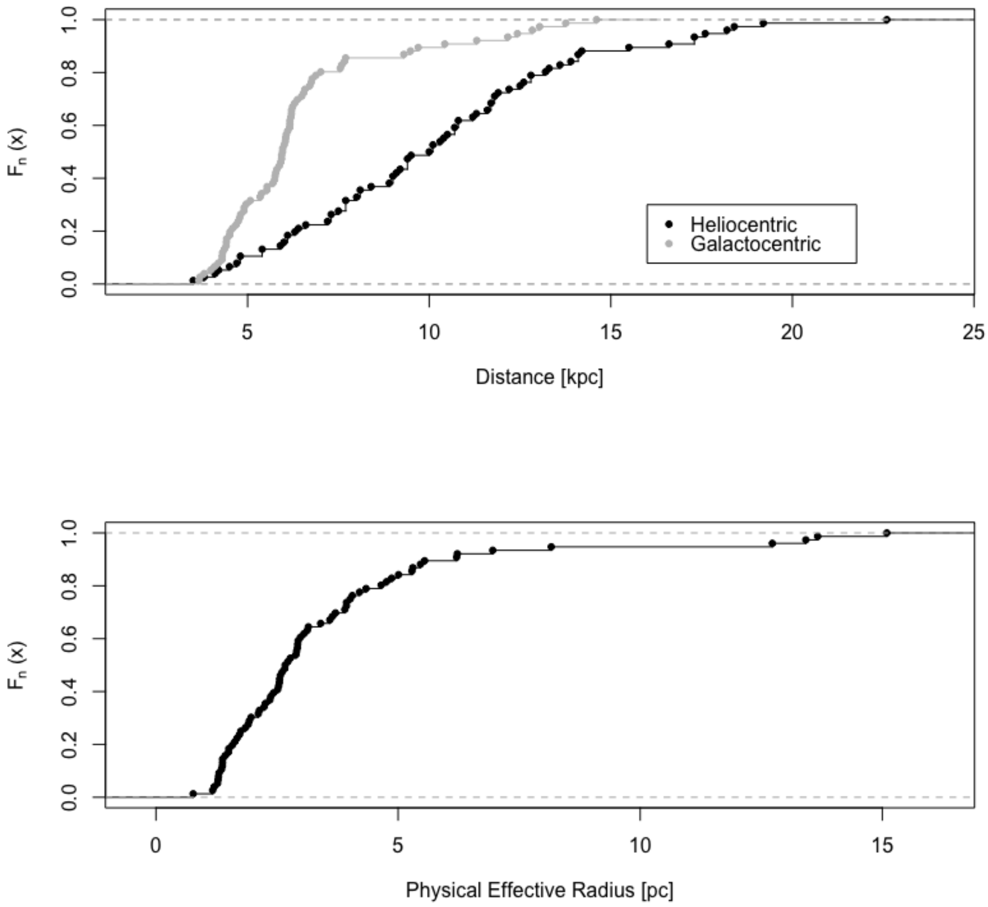


Figure 4.2: Distributions of distances of our sample of 76 HII region distances (top), and physical effective radii (bottom). Distances were taken from Anderson et al. (2014), and used to determine the effective radius. The effective radius was taken as the mean of the semi-major and -minor axes of the contoured regions.

catalogue (full details in Sect. 2.4.3, Anderson et al., 2014). The distribution of Heliocentric distances of our sample of HII regions range from 3.5 to 22.6 kpc with a mean of ∼ 10 kpc (Fig. 4.2, top). Galactocentric distances are also shown, with a mean ∼ 6.5 kpc. The mean physical effective radius of each region was determined from the contoured boundary and most are less than 5 pc (Fig. 4.2, bottom).

The contouring procedure produced a set of l, b coordinates for the boundary of each region. After translating the centre of each set to the origin of the coordinate axes, each region's distance was then used to change from angular Galactic degrees to spatial parsec coordinates. These coordinates were then taken as the pseudo-landmarks by which the shape was quantified. Interpolation splines were fitted through the landmarks, allowing for the calculation of the curvature at each spline knot (full details in Sect. 3.1.4). The curvature values for each knot then comprised the empirical distribution function (EDF) for the curvature shape description of each HII region. Each EDF was then compared pair-wise using the Anderson-Darling (A-D) two sample test statistic (full details in Sect. 3.2.2). This resulted in a symmetrical $n \times n$ dissimilarity matrix of statistical distance scores, to which a Euclidean transformation was applied. Hierarchical clustering using Ward's agglomerative method (full details in Sect. 3.3.2) was then applied to this distance matrix, in order to identify and investigate the resultant groups of HII regions, which should share a common quantitative morphology.

The resulting hierarchical structure is plotted on a dendrogram (Fig. 4.3), which shows each individual object, termed a 'leaf' and the 'branches' that join them. The height is the distance value at which the leafs/branches merge. For pairs of leaves, this is the distance value from the input Euclidean distance matrix. For all merges higher than this, the height value is that obtained from Ward's method. In the next section, the groups identified from the clustering procedure are discussed and how significant they are for investigating the physical properties of HII regions is investigated.

4.2 Groups from Clustering

Figure 4.3 shows the resulting hierarchical structure of the HII region shape data. Each HII region shown as a 'leaf' on the diagram is indicated by its Galactic longitude (to three decimal places). Cutting the dendrogram at a constant height delineates seven groups, identified by corresponding coloured integers below the leaves. The numbers in red above merges are the bootstrapped p-values, which will be discussed in detail in Sect. 4.3.1.

Although, with hierarchical clustering, one usually considers each leaf and how they form branches (groups), the advantage of having the diagrammatic representation of the dendrogram is that we can consider how the groups merge and the hierarchy ensues. On the other hand, one can instead consider the first split in the dendrogram to be the two largest distinct groups, with more groups forming at each subsequent split as one looks further down the dendrogram. The numbering of the selected groups is arbitrary. It is assigned by the algorithm following the sequence in which it considers each object. The position of groups in the dendrogram, however, is not arbitrary, with tighter groups (with a lower height score) positioned starting from the left at merges. By looking top-down, we see that Groups 7 & 6 are sepa-

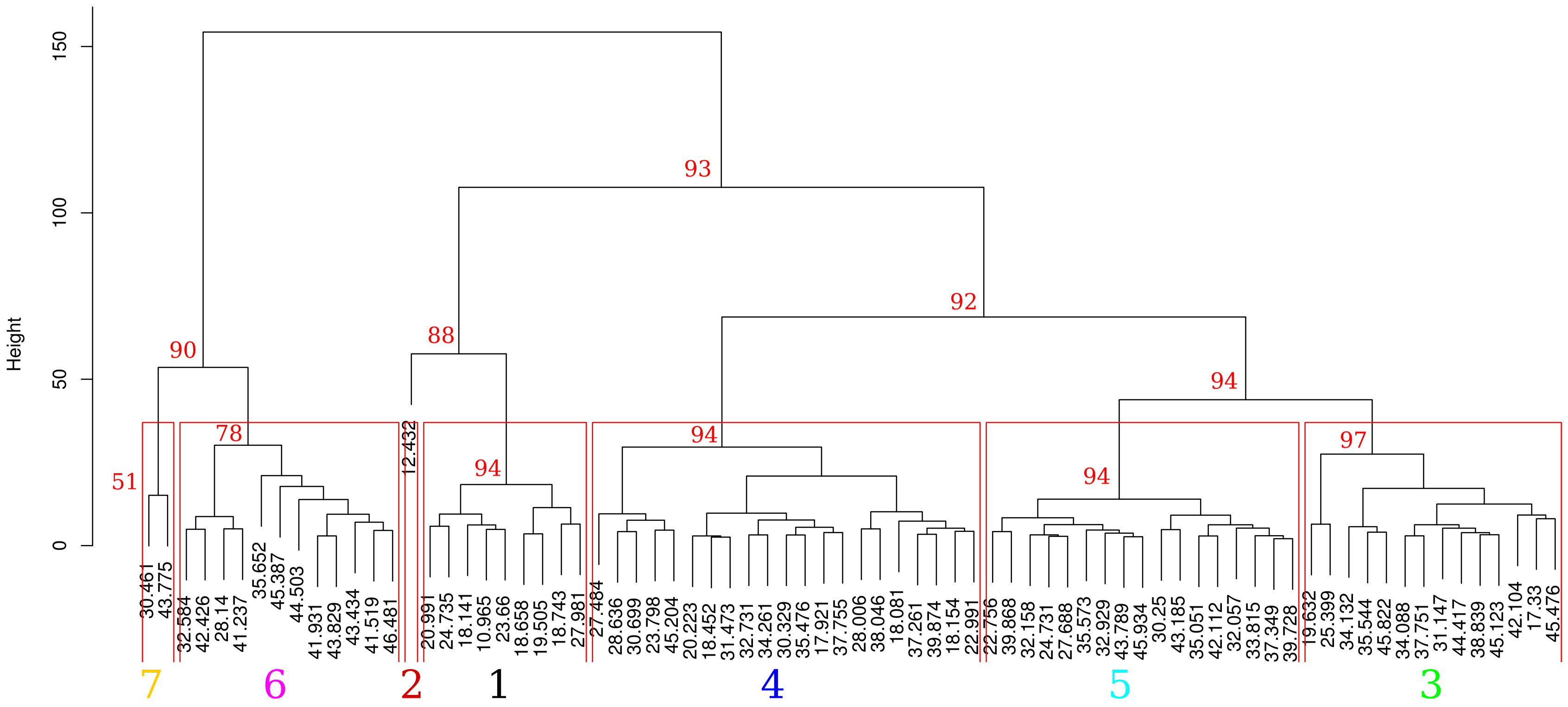
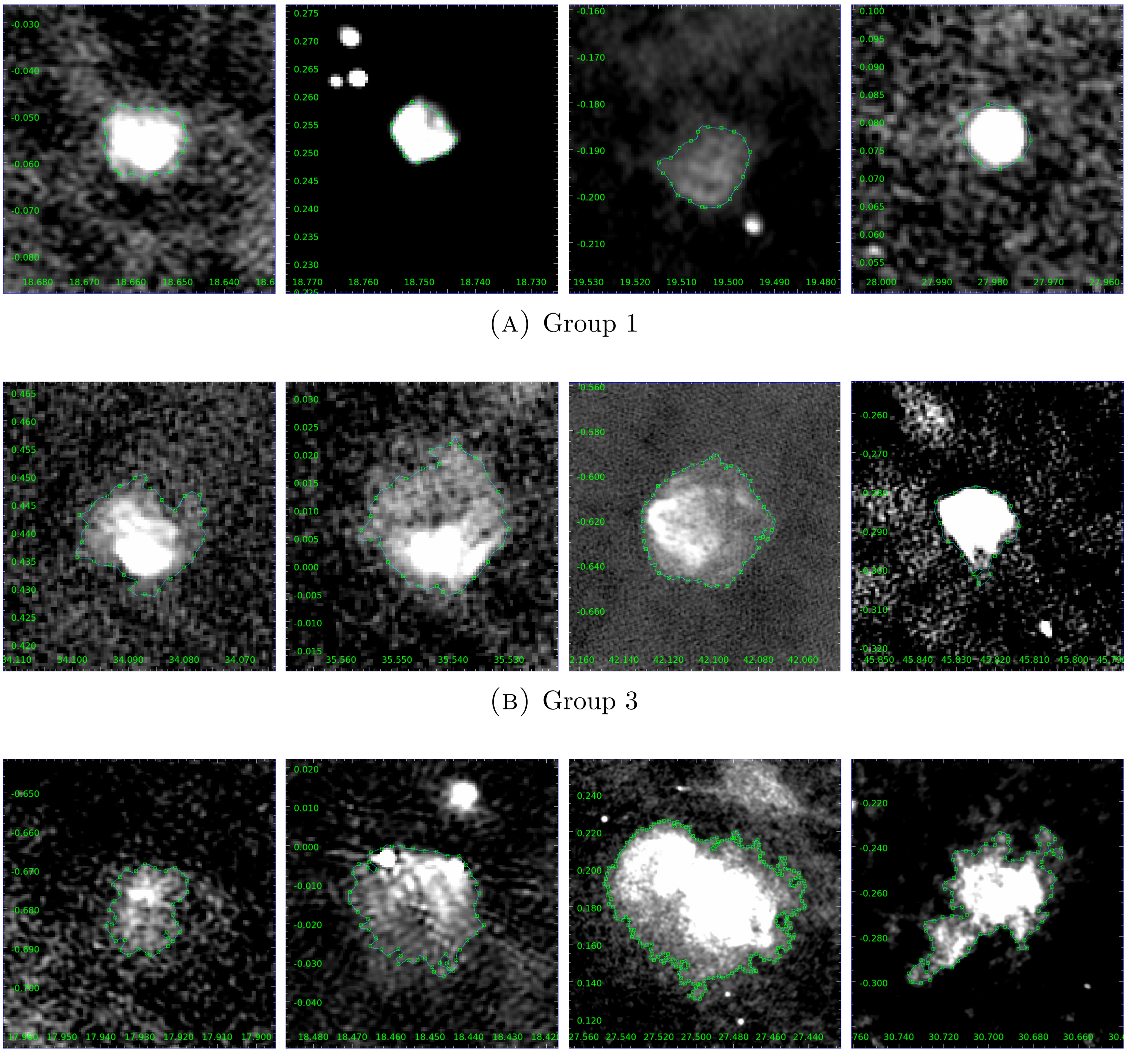


Figure 4.3: Dendrogram showing the results of our hierarchical cluster analysis of the shapes of our sample of Hii regions. Each Hii region is placed at the bottom of the dendrogram, with groups shown as merges on the dendrogram. The height score of a group is taken from Ward's agglomerative method. Red boxes delineate groups obtained by cutting the dendrogram at a constant height, which are indexed 1 – 7 at the bottom of the figure. Numbers above merged groups represent the p-values (in %) obtained from multi-scale bootstrap re-sampling of the data.



(a) Group 1

(b) Group 3

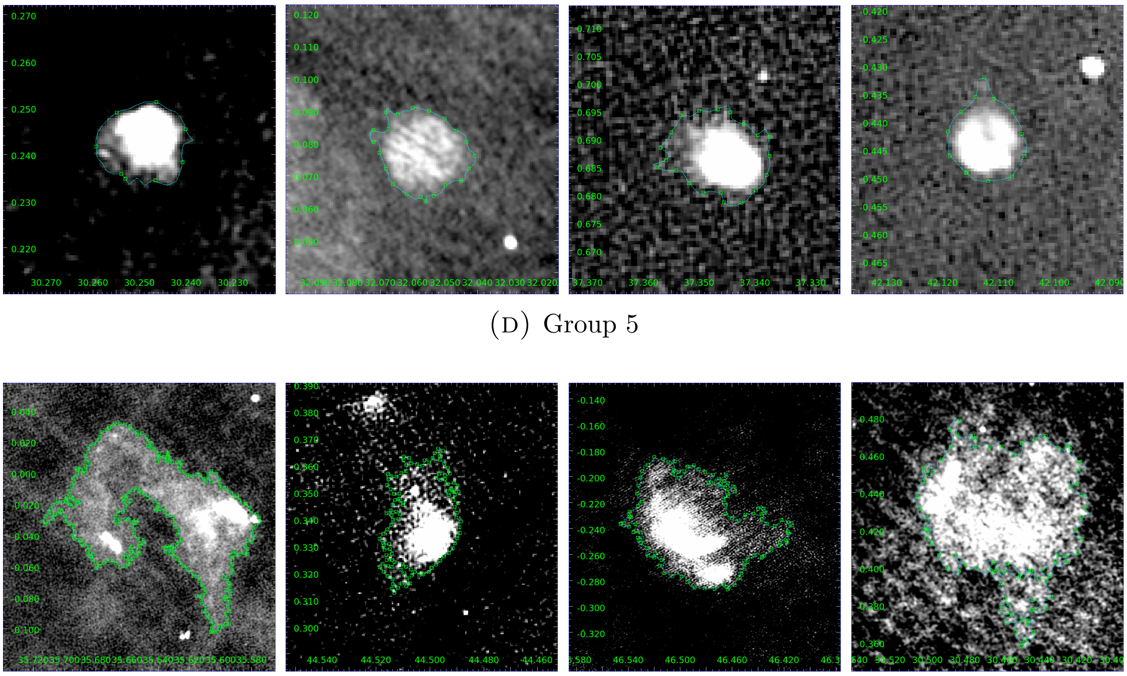
(c) Group 4

Figure 4.4: Example HII regions per group identified in the dendrogram in Fig. 4.3. In each image the extracted shape contour and spline points are shown.

rated from the rest of the groups at the highest level. We should hence expect to find most difference between these sets of regions. Considering the right side of this first split, we note that Group 1 and the outlier object in Group 2 are separate from Groups 4, 5 & 3.

4.2.1 Visual inspection

In order to determine whether the statistical clustering of the HII regions via their morphologies had produced suitable groups, visual inspection of the image and contour data was performed. An overview summary of four example HII regions assigned to Groups 1, 3, 4, 5 and 6 in the dendrogram in Fig. 4.3 is shown in Fig. 4.4. Images of each of the 76 HII regions by group can be seen in Appendix B. We can see from



(d) Group 5

(e) Group 6. Note that the region on the far right was one of the two listed in Group 7 in Fig. 4.3.

Figure 4.4: Continued from previous page

the grouped images that the amount of local curvature variance along the region boundaries changes between each group. These differences in shape is more apparent in some groups than others. In Appendix B, the MIR images are also shown for reference, with the 'edge' obtained from the contouring procedure shown over-plotted on all images. In most cases, the MIR emission is well spatially coincident with the radio continuum emission. There are some exceptions, where the radio emission extends further than the confines of the bubble: G010.964+00.006 in Group 1; G012.42900.049 in Group 2; G028.146+00.146 and G030.468+00.394 in Group 7. In many cases, the automatically extracted radio continuum 'edge' matches up extremely well with the 8μm shell (the green channel in the MIR colour-composites), suggesting that the interface between the ionised region and extended PDR is well traced by this edge.

Those grouped on the left hand side of the dendrogram (labelled Groups 6 & 7 in Fig. 4.3) possess the highest amount of local curvature variation along their boundaries. There are also visual similarities between groups originating from common parent branches - e.g. Groups 5 & 3. Group 1 appears to host the most unperturbed regions, with low amounts of local curvature variance, hence suggesting why it is a distinct branch on the dendrogram. The regions in Group 4 display a range of morphological features, including boundaries that are circular, elongated, and largely perturbed. This group may hence represent the 'average' regions, with no explicitly defining features which would place them in to one of the other groups. It is not immediately apparent why the outlier region in Group 2 is placed into a single group, nor why it is joined to Group 1 further up the dendrogram. A possible explanation is that the amount of interpolation points used to determine this region's curvature distribution is far greater than most regions, since it is located at the furthest distance in the sample. This region will be discussed in more detail in Section 4.3.5.

4.2.2 Curvature EDFs

For some of the regions, it is hard to tell visually why they are placed into a given group. In order to more quantitatively determine how the clustering procedure is grouping the regions, the Empirical Distribution Functions (EDFs) of all k values for all regions within each of the seven groups identified in the dendrogram were compared. The resulting plot is shown in Fig. 4.5. It has been truncated to highlight the global differences between each EDF. The pair of objects in Group 7 have the largest fraction of high k values, followed by the objects in Group 6. These two groups merge with the rest of the groups at the highest point on the dendrogram, suggesting this is the cause for this first split. However, Group 4 which displays the next largest amount of high k values, is in the middle of the groups on the right side of the first split. This suggests that it is not exclusively the amount of high k values that is delineating the groups. This result was not unexpected, since the A-D test was used, which accounts for numerous discrepancies between EDFs. The EDFs of Groups 3 and 5 cross multiple times and if the cut on the dendrogram was made higher, these two groups would be the first to merge. Group 1 appears to have the lowest k values (which supports the observation that this group hosts the least-perturbed regions), with the outlier region in Group 2 having a distribution that lies between Group 1 and Groups 3 & 5, suggesting why it joins with Group 1 in the next merge on the dendrogram.

These results confirm that the pairwise A-D test of the k values and the hierarchical clustering are delineating groups as expected, with different groups possessing different, distinct overall EDFs. It remains to be seen from this plot alone, however, what other factors are causing the exact separation of the neighbouring groups. We can further visualise how the hierarchical clustering has generated groups by utilising an ordination technique for the input distance matrix.

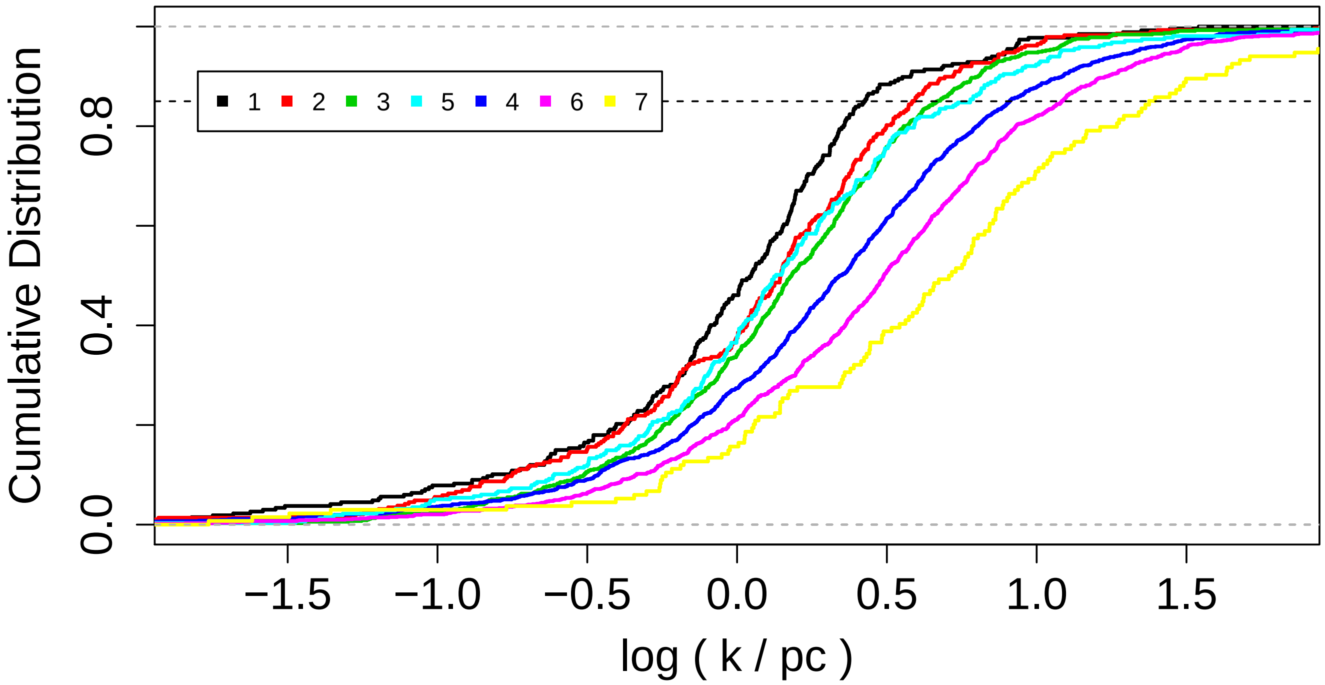


Figure 4.5: Empirical cumulative distribution functions for the curvature values of all HII regions belonging to the seven groups identified in Fig. 4.3. Note that this figure has been truncated at log $(k/\mathrm{pc}) = \pm 1.8$ to highlight the differences between EDFs at the centre of the distributions. The maximum curvature values for Group 7 are significantly higher than the other groups.

4.2.3 Multi-Dimensional Scaling

The input distance matrix allows Ward's method to find groups in multivariate Euclidean space, the same used by Multi-Dimensional Scaling (MDS, also known as Principal Coordinate Analysis, Torgerson, 1952; Kruskal, 1964). MDS reduces the dimensionality of a distance matrix to represent variances and similarities within data on an ordinance diagram. A set of uncorrelated, orthogonal axes are produced, where each axis created by the MDS procedure has an associated eigenvalue, whose magnitude indicates the amount of variation captured by that axis. A common, two-dimensional example of a use for MDS is to recover the coordinates of geographical locations, simply via their corresponding distances to one another. MDS is mathematically similar to principal component analysis (PCA), which is used often in astronomy. PCA seeks to reduce dimensionality between different variables, also producing sets of orthogonal axes refereed to as components. For MDS, the number of axes produced equals the number of objects in the distance matrix (which, in this instance, is the number of HII regions in the sample; in PCA it would be the number of variables considered). Each axis' relative eigenvalue gives the importance of that axis for summarising the variances in the distance matrix.

In order to determine the ordination coordinates and associated eigenvalues using MDS, one begins with a distance matrix **D**, which is computed using the standard Euclidean transformation (for classical MDS[2]):

$$\mathbf{D} = d_{xy} = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2} \tag{4.1}$$

we then apply a double-centring transformation to the distance matrix **D** using the following equation:

$$\mathbf{B} = b_{xy} = -1/2(d_{xy}^2 - d_{.y}^2 - d_{x.}^2 + d_{..}^2) \tag{4.2}$$

where $d_{.y}$ is the average of all d_{xy} taken across all of x, $d_{x.}$ is the average of d_{xy} for all y and $d_{..}$ is the average for all x and y values. We can then compute the p largest eigenvalues of **B**, $\lambda_1 > \lambda_2 > ... > \lambda_p$, whose magnitudes give the amount of variance signified by a given dimension. These eigenvalues then have corresponding eigenvectors $L = (L_{(1)}, L_{(2)}, \dots, L_{(p)})$, which are normalised such that $L'_{(i)} L_{(i)} = \lambda_i$. The ordination coordinates for each principal coordinate axis are hence the rows of

[2] Metric and non-metric MDS can be applied when the distance matrix is non-Euclidean, but will not be considered here since we already utilise the Euclidean distances between dissimilarities of HII regions

Table 4.1: MDS eigenvalue scores of the HII region T_{AD} shape distance matrix. The first 11 principal coordinate axes are shown.

MDS Axis	Eigenvalue λ	Relative Eigenvalue [%]	Cumulative Eigenvalue [%]
1	16658.768	0.601	0.601
2	8400.892	0.303	0.904
3	1913.342	0.069	0.973
4	242.756	0.009	0.982
5	208.151	0.008	0.989
6	140.503	0.005	0.994
7	31.457	0.001	0.995
8	26.012	0.001	0.996
9	23.847	0.001	0.997
10	15.977	0.001	0.998
11	8.239	0.000	0.998

L. As with Ward's agglomeration method, MDS is a least-squares optimisation of the data, which in this case, shows the optimum, centred configuration of the data.

For the same input distance matrix **D**, used for the hierarchical clustering of the HII regions; MDS reveals that the first two principal coordinate axes account for 60 % and 30 % of the total data variance, respectively. A summary of the first 11 MDS axes is shown in Tab. 4.1. A further 7 % is represented by axes 3, hence the first three axes account for 97 % of the data variation. The subsequent axes each capture less than 1% variation individually (with the remaining axes not shown in Tab. 4.1 each accounting for less than 0.1%); collectively making up the remaining 3% variation. In this instance, MDS successfully reduces the dimensionality of the data to essentially three dimensions.

The first two MDS axes are shown in Fig. 4.6 - top, and the second and third axes are shown on the bottom. If the original T_{AD} dissimilarity matrix was used as the input to PCA, the resulting PCA ordinance and eigenvalue scores would be equivalent to what we obtain using MDS in Euclidean space. Whilst PCA has a natural parametrisation of eigenvalues, corresponding to some combinations of the input variables, there is not such a correspondence in MDS. The input to MDS is a distance matrix rather than a set of variables, which in our case constitutes *shape*-distances between the HII regions. Hence, the MDS ordinance is used here to see if the results from hierarchical clustering delineate groups that visually correspond to groupings in the MDS Euclidean space. Associations between observables and the ordinations along the MDS axes can then be sought.

Figure. 4.6 shows that there is a clear correspondence between MDS axis 1 and the groups from hierarchical clustering. Group 2 (red triangles) which represents

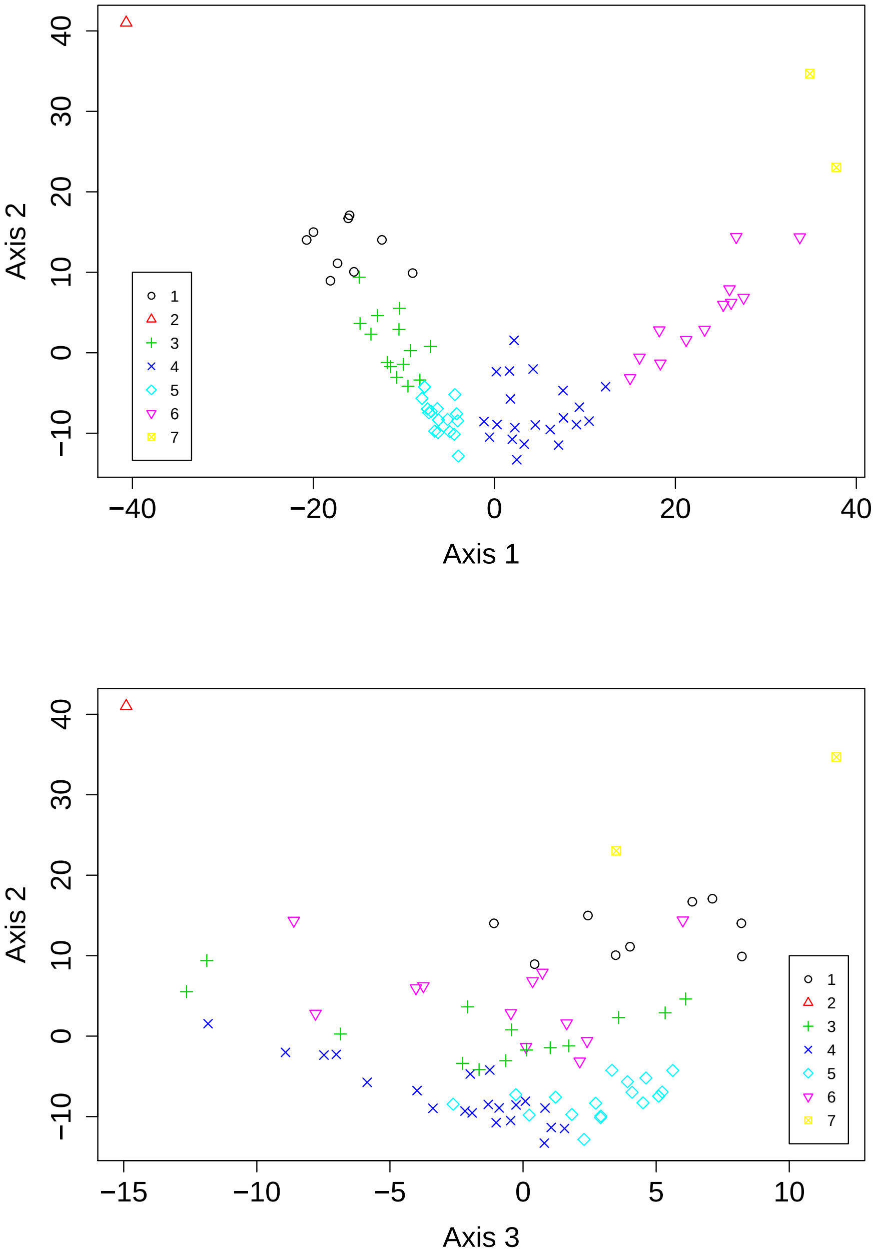


Figure 4.6: Ordination graphs showing the results of Multi-Dimensional Scaling of the distance matrix used for the hierarchical clustering. Each point represents one HII region, with colours and symbols matching the groups identified on the dendrogram in Fig. 4.3. The first and second principal coordinate axes are shown on the top and the third and second on the bottom. Together, these three axes represent 97% of the variability in the distance matrix.

the outlier region is seen to be ordinated away from the rest of the data in each of the first three axes. Excluding Group 2, the ordering of the groups along axis 1 matches the ordering in the EDF plot shown in Fig. 4.5. This suggests that the most variance in the data can be explained by the amount of high curvature points within each group. Groups 6 (pink inverted triangles) and 7 (yellow crossed circles) have the highest values along axis 1. Group 4 (blue crosses) are close to the origins of axes 1 and 2, but show a high variance along all three axes. This supports the notion observed perviously - that this group is representing 'average' regions, since this group's EDF is also in the middle of the set. Groups 5, 3 & 1 (cyan diamonds, green plus signs and black circles, respectively) all have some cross over along axis 1. However, with regards to Groups 5 (cyan diamonds) and 3 (green plus signs), which are the two that have similar EDFs and originate from the same parent branch on the dendrogram, MDS axis 2 reveals that there is a distinct separation of the groups along this axis. The variance along axis 2 is likely due to another feature of the curvature distributions, since regions with high scores on axis 2 are at the extremes of axis 1, but do not necessarily have a high fraction of high curvature values (Group 1 has the lowest). Groups 5 (cyan diamonds) and 4 (blue crosses) have similar scores on axis two, but they appear separated along axis 3 with some cross over around the origin of this axis. There is no further correspondence with group and axis 3, that is not already apparent from axes 1 and 2.

In addition to confirming that the hierarchical clustering procedure is identifying suitable groups of Hii regions based on their quantitative shapes, MDS also gives us a means of graphically comparing the results from hierarchical clustering to that of k-means partition clustering (detailed in Sect. 3.3.1). Figure 4.7 shows two results from applying k-means clustering to the distance matrix of Hii region shapes. The first and second MDS axes are shown and the symbols of points match those of Fig. 4.6, showing the seven groups identified from hierarchical clustering. In Fig. 4.7, however, the colours correspond to the groups obtained from applying the k-means partition clustering to the data. Seven groups are shown in Figure 4.7 upper panel, and six are shown in the bottom panel. The top panel shows somewhat of an agreement between the groups identified from both methods. Notable discrepancies are the outlier regions at high axis 2 scores are represented as one group by k-means clustering (as opposed to two by hierarchical clustering). There is also overlap between each of the remaining groups as seen by the differences in colours/symbols.

The most discerning difference between the two clustering methods, however, is seen when considering the lower panel, comprising six groups identified by k-means clustering. In hierarchical clustering, when groups are joined together at a higher

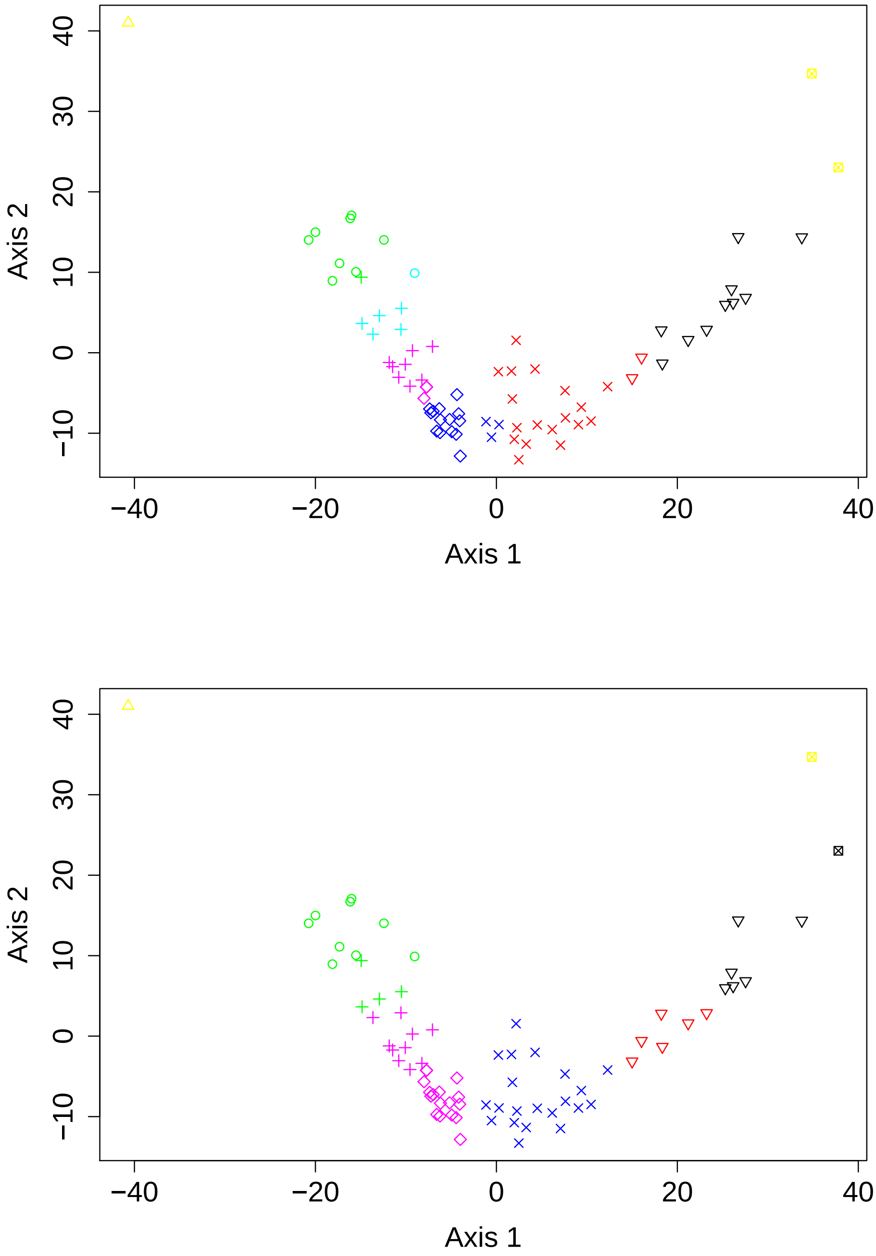


Figure 4.7: Ordination MDS graphs showing the results of k-means partition clustering of the distance matrix of Hii region shapes. Each point represents one Hii region, with colours identifying the groups from k-means clustering and symbols matching those in Fig. 4.6, which were the groups identified from hierarchical clustering. The top panel shows the results from $k = 7$ clusters and the bottom from $k = 6$.

cut, former group membership persists and a concatenation of groups occurs. The dendrogram in Fig. 4.3 shows that this would occur first for Groups 3 & 5. Figure 4.7, however, shows that by reducing the number of required groups by one, many changes of group memberships result. Firstly, the third yellow outlier region with the lowest axis 2 score is now joined to the black group below. Half of the cyan points from the top panel now belong to the green group in the lower, with the remainder joining the pink group. However, the pink group has now extended into what was the blue group; the blue group has also captured members from the former red group; and the new red group has half of the former black group. It should also be reported that whilst applying the *k*-means algorithm - group membership of individual data-points ordinated at the boundary of groups was changing between applications. This is due to the placement of group centriods and was one of the main limitations of *k*-means clustering noted in Sect. 3.3.1. Whilst one would have been able to see these noted results from the data and group assignments of the *k*-means procedure, the MDS ordination has made clear these group discrepancies within the *k*-means procedure. For the remainder of this analysis, *k*-means partition clustering for the purpose of identifying groups within the data will not be employed further.

Thus far in the analysis, visual inspection of the groups, the group EDF plot and the MDS ordinances show that the hierarchical clustering procedure is identifying groups that share similar morphological features. Both the group EDF plot and MDS ordination highlight quantitatively the similarities between Groups 3 and 5, yet visual inspection of Group 4 reveals a range of shapes are included within this group. Whilst it was just outlined that the group memberships are fixed in hierarchical clustering, we must still consider the reliability of these seven groups, which thus far have been identified by an arbitrary cut at a given height on the dendrogram.

4.3 Discussion

4.3.1 Group & Method Validity

As mentioned many times in Chap. 3, the extraction of shape from the HII region sample must be systematic and repeatable. Additionally, a future use of this unsupervised clustering method is to potentially identify a training set for use in machine learning supervised classification of HII regions. A number of robustness checks on both the shape data and performance of the analysis method were hence carried out; to determine how well the method copes with different selection choices.

Firstly, in order to test how reliable the identified groups are, a multi-scale bootstrap re-sampling (Shimodaira, 2004) method was applied to the data. This provides a p-value for each merge on the dendrogram, which represents the probability that each respective merge is intrinsic to the data. In standard bootstrap re-sampling, the original data of size n is considered, and bootstrapped data-sets (also of size n) are generated by sampling data from the original set with replacement. By applying the same hierarchical clustering procedure to the bootstrapped sets, one can compare how many times a group from the original data appears in a bootstrap replicated dendrogram. However, by always considering a sample size of n, biases can be introduced to the bootstrap probabilities (Shimodaira, 2004). Multi-scale bootstrap re-sampling generates a range of bootstrapped data-sets, which can be smaller or larger than n. For each sample size, hierarchical clustering is applied, and the number of times a given group from the original data appears in a replicated dendrogram is counted. This is repeated a number of times at each sample size, resulting in an approximately unbiased p-value for each group in the original data. For the distance matrix $\mathbf{D}$, the largest group obtained from our sample size of $n = 76$ contained 28% of the regions, the two smallest groups contained 2 and 1 region, and the next smallest group had 12% of the regions. The lower bound of the multi-scaling was hence set to $0.5 \times n$, allowing for groups of all sizes found in the final result to be recovered by the bootstrapping. This scaling was then increased in steps of $0.1 \times n$ to an upper-bound of $1.4 \times n$, at each step using 10^4 bootstrap iterations. This set up yielded reproducible p-values, where all p-values above 90% had a standard error less than 0.6%. By standard normal theory this equates to a confidence interval on the p-values of ±1.2%. The standard error of the p-values had plateaued by 10^4 iterations, reducing by a few orders of magnitude from testing with 100 iterations.

Figure 4.8 shows a reproduction of the dendrogram in Fig. 4.3 but with all approximately unbiased p-values listed below each merge. In the top right of the figure, the standard error for each p-value is shown, confirming the high confidence interval for each value. We can see that in this dendrogram containing all p-values that, with the exception of a few, most pairs of HII regions have a score of over 90%. For those merges that have lower p-values, such as the pair on the left-hand-side of the dendrogram in Group 7, which have a score of 51%; one would next consider the score of the parent merge, which in this case is 90%. This means that whilst those two objects are only paired together in half of the bootstrapped resamples - 90% of

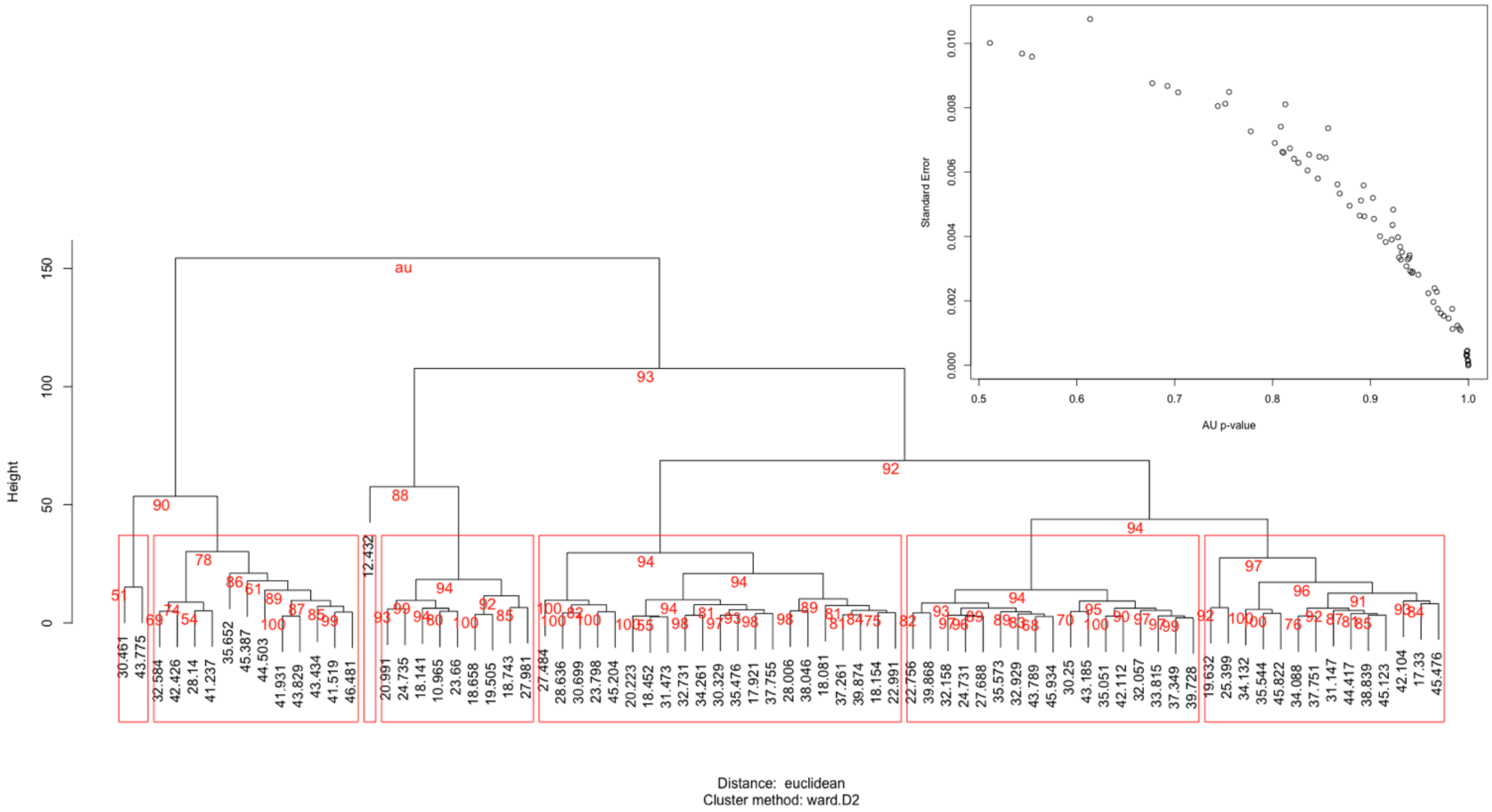


Figure 4.8: As Fig. 4.3, but with each of the p-values obtained from multi-scale bootstrap resampling of the data shown. The top right panel shows the corresponding standard error scores for the listed p-values.

the time they are both going to be grouped in some respect with the rest of the data comprising the entire left-split of the dendrogram. Keep in mind, this is the probability for those exact objects to constitute one group with no perturbations. We will see in the subsequent sections that differences in the input data can change the resulting structure of the dendrogram. Hence, the p-values reported here should be taken to simply confirm that the seven groups identified are not likely to be chance occurrences from the dataset, as they would be if they all possessed low p-values.

For the seven selected groups, the corresponding p-values for each merge are included in Fig. 4.3 in red above each group outlined by the constant cut. Subsequent merges higher up the diagram are also shown in this figure. For these seven groups, from left to right, the associated p-values are: Group 7: 51%, Group 6: 78%, Group 2 is the outlier object, which has a p=88% to be joined to Group 1, Group 1: 94%, Group 4: 94%, Group 5: 94%, Group 3: 97%. If Groups 7 and 6 are joined, the respective p-value for all of these objects to be within one group is then 90%. Hence, for the subsequent discussion of group properties and parameters, Groups 1, 4, 5, and 3 will be considered, and two regions from Group 7 will be merged into Group 6. Group 2 will continue to be referred to as the main outlier (the HII region that none of the other HII regions want to be friends with).

Before continuing with these six groups of HII regions, some further checks were carried out to determine how well the group structure of the dendrogram remains when varying certain selection choices pertaining to the shape data. Whilst the resulting hierarchical structure is fixed, it is fixed via the input distance matrix, which in turn is directly obtained from the shape descriptor of the HII regions: the curvature distributions. Certain selection choices were made throughout the shape extraction procedure, in order to automate and standardise the process. The first of these was the initial sigma value used to extract the boundaries of the HII regions. Our reasoning for using the 1σ contour level was explained in Sect. 3.1.3, with lower σ levels including too much noise and higher levels not capturing enough detail. Whether or not smaller changes to the sigma level affect the results is investigated here. The analysis was re-run with 0.8 and 1.2σ contours and in both cases, some of the group associations obtained from the 1σ data changed. This is explained by the fact that changing the threshold level inherently changes the shape considered. The extent to which the shape changes differs on an individual basis, with some regions much more susceptible than others. In turn, the computed pairwise distance matrix representing the region shapes also changes, which can have a knock-on effect for the hierarchical procedure. This is due to how the inter-group distances are computed by the agglomeration method, which considers each object's pairwise distance within a given group. In this application, the Ward method is less amenable to small within-group distance changes than, for example, the single-linkage method (Ferreira and Hitchcock, 2009). Nevertheless, larger differences in the distance matrix from those regions whose shape changes substantially *will* be captured by the agglomeration process, and the extent to which this affects the results would be apparent in the final groupings.

To determine whether lowering or increasing the σ level has more affect on the shape and resulting groupings, the bootstrapped p-values for the dendrograms of the entire $\pm 0.2\sigma$ data were first computed. In each case, they were notably lower than for the 1σ data with mean values of 71% for 1.2σ and 84% for 0.8σ, compared to 93% for the 1σ data. This suggests that the groups outlined at these levels are less credible than the 1σ level. In principal, one could test various sigma levels for resulting groups with the optimum p-values. Next, the shape data of randomly selected individual regions was changed to the 1.2 or 0.8σ level, and the analysis was repeated. On average, 18% of regions changed groups at 1.2σ and 29% of regions changed group at 0.8σ. Although the 1.2σ data was less likely to result in a individual region changing group, the lower overall p-values for this level may be explained by these shapes being more smoothed with fewer features, so the pairwise distances are generally lower and it is more likely for them to be grouped with different similar regions. For the 0.8σ regions, the larger fraction of individual regions changing group is expected since lowering the threshold allows for more detail and extended features with high curvature values to be captured. In all cases where a region changed group, the rest of the group structure from Fig. 4.3 remained. In addition, when a change was observed, it was to neighbouring branches, and no region moved into Group 6 from the right hand side of the dendrogram. This confirms that the hierarchical groupings are indeed not sensitive to small changes in the distance matrix, and how and when a region moves to/from a given group may be quantifiable. This will be discussed further using the synthetic observation data in Chap. 5. Although the arguments here support the choice of the 1σ level for defining the 'edge' of the HII regions, this may not be the case for other surveys with different noise profiles.

A similar test was carried out for the spline knot spacing. The spacing is in essence the resolution of the shape data, hence changing this would also in some cases change the description of the region shape, due to finer details either being included or smoothed out. Initially, a spacing of 0.54 pc was used, corresponding to the beam size of the survey at a far distance of 19.2 kpc. Hence, changes to a region's distance subsequently changes the knot spacing along its boundary. Smaller distances have larger spacings (lower resolution) and vice versa. The distances of regions in the sample were changed by a random amount between ±20%, which is 5% larger than the average distance error given by Anderson et al. (2014) for the kinematic distance estimates of the sample. This sought to address both issues of the HII region distance errors and how different resolutions affects the results. When decreasing the distance/resolution of an individual region, 18% changed groups. When increasing the distance/resolution, 53% of regions changed groups. Capturing more detail of the shape thus has a much more profound affect than smoothing the data. This agrees with the results from the MDS ordination that shows a correlation between assigned group and number of high-curvature points along the region boundary. The fact that increasing the spline resolution has the most notable affect on the groups supports the choice of using the far distance resolution to reduce distance biases, which is further discussed in the following Sect. 4.3.2. Whilst this is an important limitation to identify, it should be noted that changing the resolution of individual regions in this manner affects the systematic nature of the shape analysis. It is hence expected that changing the shape descriptor of a region by a substantial amount will cause it to move to a different group associated to another morphology class. Furthermore, certain regions of a given group each moved to the same group after changing the resolution. For example, each region that moved from Group 3 was relocated to Group 1 after increasing the resolution. This suggests that the morphological groupings could be further quantified with a more controlled sample.

In terms of the sample of HII regions considered here, the region images in Appendix B show that a varied selection was obtained, in terms of morphological features. To test whether the group results are sensitive to different samples, a number of regions from the data were removed and the analysis re-run . 18, 25, 32 and 38% of regions were removed at random and in each case, the resulting group associations matched that of Fig. 4.3 for the remaining regions. Similarly, whether the group structure remains when removing all regions belonging to a given group was tested. Again, all of the group structure is kept intact in the absence of any given group. These results suggest that the six groups identified do in fact represent distinct categories of morphologies, and is promising for the prospects of producing a training set in future work.

For the rest of this section, the physical properties of the HII regions within each group are considered. Any associations between these parameters and assigned group is investigated and whether it is some physical property responsible for the

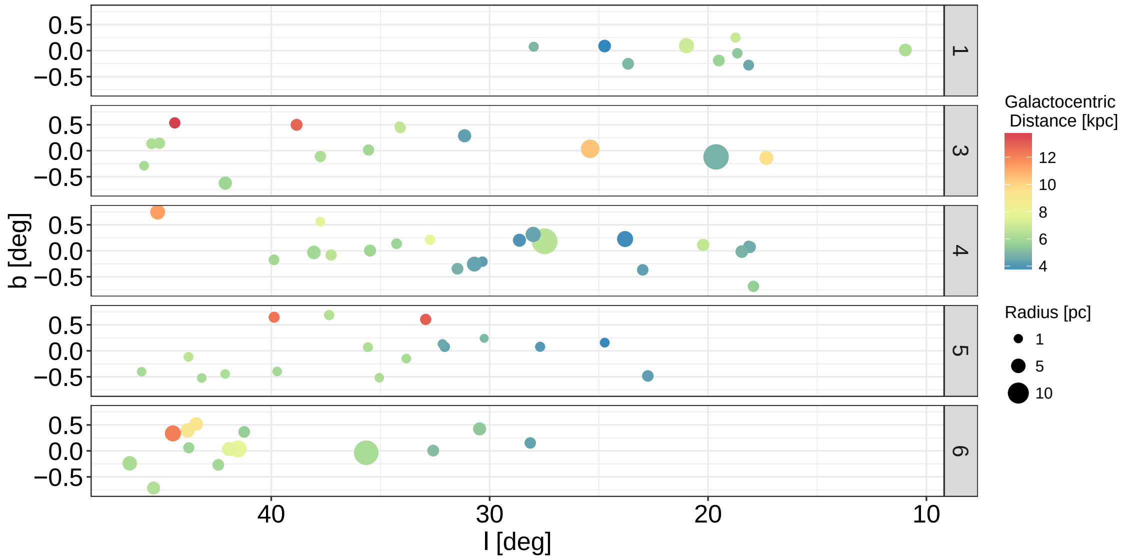


Figure 4.9: Galactic position of our sample of HII regions, split by group. Points are coloured by their Galactocentric distance. Note that the two regions in Group 7 from Fig. 4.3 have been merged into Group 6 and the outlier region in Group 2 is not shown.

30% variation along axis 2 of the MDS ordination, having already shown the 60% variation represented by axis 1 is likely linked to the amount of high-curvature values within each region. Hence, any physical association with the amount of high curvature values was sought. Also discussed in this section are the outliers identified by hierarchical clustering and the assumptions made about the assigned distances and ambient densities are investigated. Also outlined are potential future applications of this statistical methodology to larger samples of HII regions, and other diffuse astronomical objects.

4.3.2 Position, Radius and Distance

Figure 4.9 shows the positions of the HII regions – Galactic longitude and latitude, with points coloured by Galactocentric distance and scaled by physical effective radius – split by assigned group. Group 1 only contains regions with $l < 30°$ and Group 6 only has regions with $l > 28°$, however this is not exclusive, with Groups 3, 4 & 5 containing regions with a range of l values. There are no group preferences for regions with positive or negative latitude values, nor large or small Galactocentric distances. This suggests that the various shapes of HII regions located in the first half of this quadrant of the Galaxy are homogeneously distributed in Galactic latitude, with a small preference for those at a given Galactic longitude to share a common morphology.

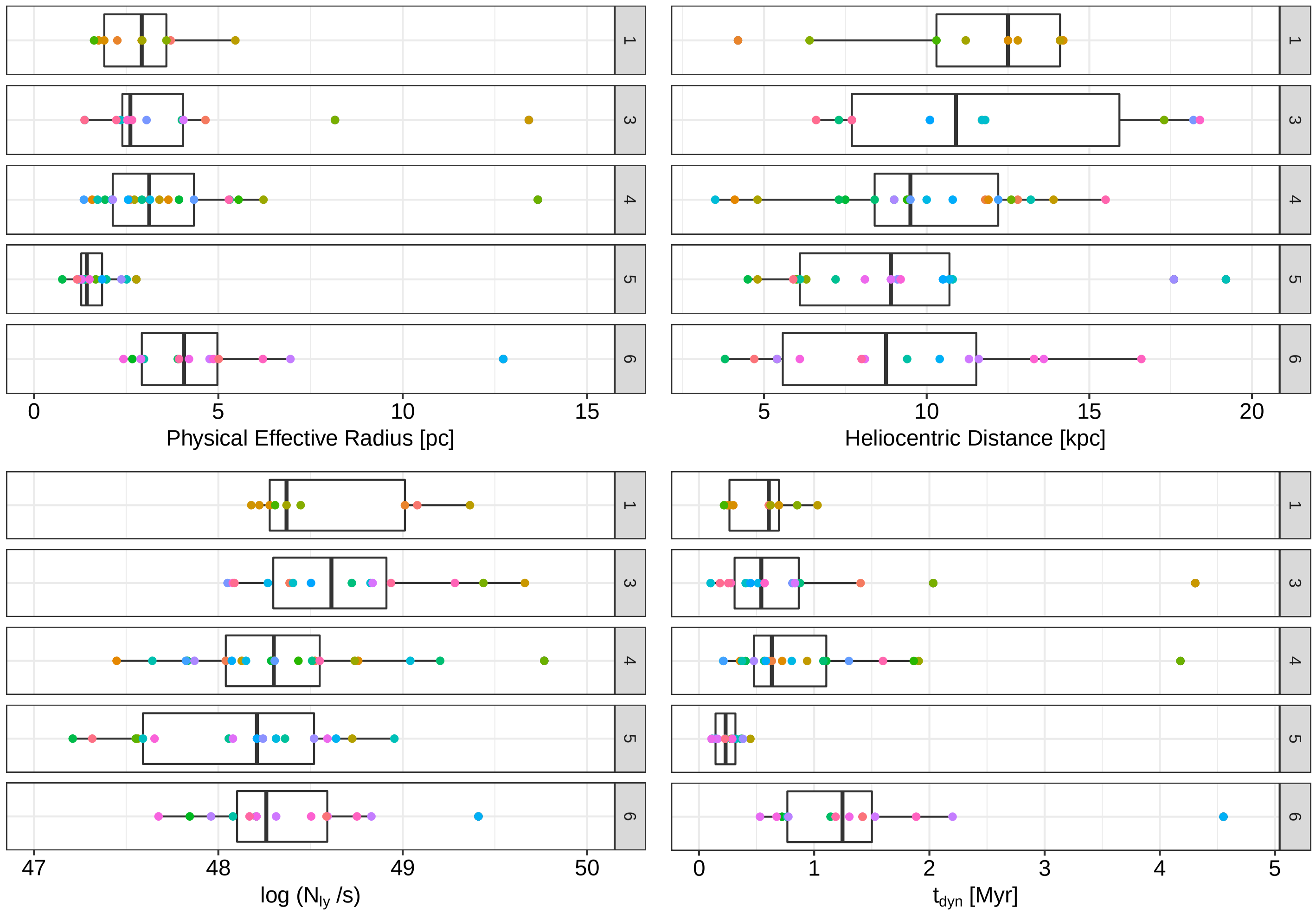


Figure 4.10: Summary of HII region properties by group. Top left shows physical effective radius, top right shows heliocentric distance, bottom left shows the number of ionising photons within the boundary of each region and bottom right shows the dynamical age. Points are consistently coloured by the region's Galactic longitude, allowing for identification of corresponding regions across plots. As with Fig. 4.9, the outlier region in Group 2 is not shown and the two regions from Group 7 are merged into Group 6.

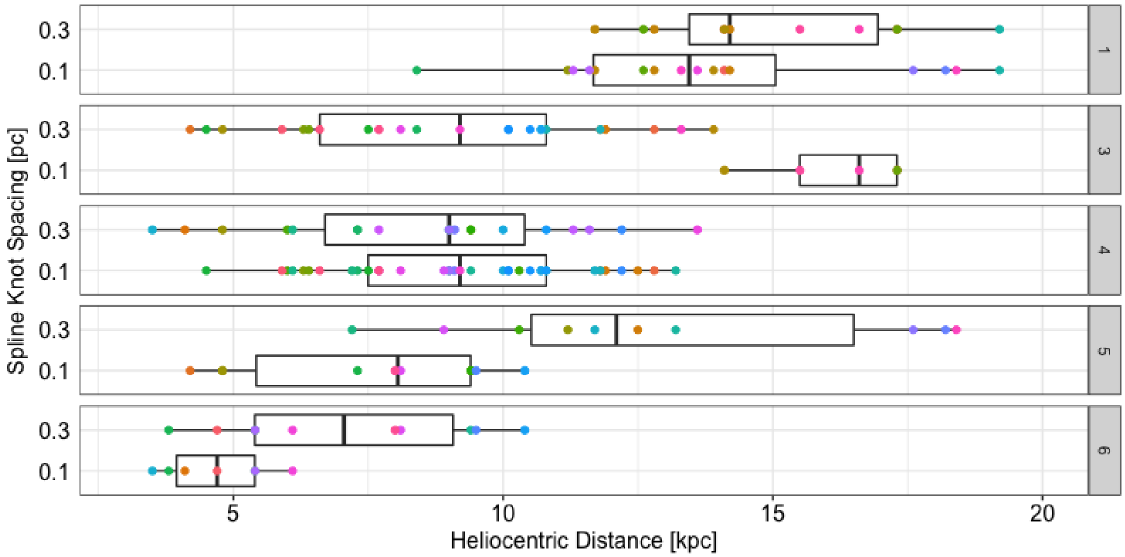


Figure 4.11: Comparison box-plot of the heliocentric distances of the MAGPIS Hii regions split by group, when using smaller spline knot spacings. The top box in each facet group represents the distance distributions for a spacing of 0.3 pc, the bottom distributions are for a spacing of 0.1 pc. As with Fig. 4.10, points are consistently coloured by the region's Galactic longitude.

Figure 4.10, top-right, shows the distribution of Hii region heliocentric distance by group. As previously outlined, we sought to ensure that no bias was given to regions at a near distance. This intent is confirmed by Fig. 4.10, with no preference for regions at any given distance appearing in a particular group. Further to our tests using different spline intervals (Sect. 4.3.1), it was tested whether considerably smaller intervals (equal to 0.1 pc and 0.3 pc), that captured much more curvature detail for the closer regions, would show a bias in the clustering results and found that, in these cases, there was a preference for objects at both large and small distances to be sorted into the same groups (Fig. 4.11). This preference was most profound at the smaller interval of 0.1 pc, where 72% of regions at a distance greater than 12.5 kpc were placed in one group. With this spacing, Group 3 hosts regions exclusively at a far distance, whilst Group 6 hosts regions that are exclusively nearby. For the interval of 0.3 pc, 39% of the far regions were then placed into Group 1 and there are no far regions placed into Group 6. Since there is no preference for the far regions, nor those at the nearest distances, to be placed in a given group at the 0.54 pc spline interval used, we can infer that any associations between region parameters and groups are not associated to any distance/angular resolution bias.

There are associations between the region radii and assigned group (Fig. 4.10, top-left). Since the curvature measurements of each region concern the size-and-shape, this is not an unexpected result. However, effective radius as a measure was not a direct input to the hierarchical clustering, since the curvature distributions or average curvature values per region do not always correlate with the effective radius. Therefore, the fact that these results have the following associations means that this size information is in some cases retained by the curvature distributions: 59% of regions with a radius less than 1.6 pc are assigned to Group 5, and 53% of regions with a radius greater than 4 pc are assigned to Group 4. Because of this association between large regions and Group 4, size cannot be attributed as the primary reason for the variance along axis 2 of the MDS ordinance, since Group 4 has low scores along this axis. The HII regions physical size can be linked to both ionising mass and age, for an assumed constant ISM density. The amount of ionising flux from each region was therefore measured, to provide estimates for the stellar masses, which when compared to the size of the regions, can also give estimates for the region dynamical ages.

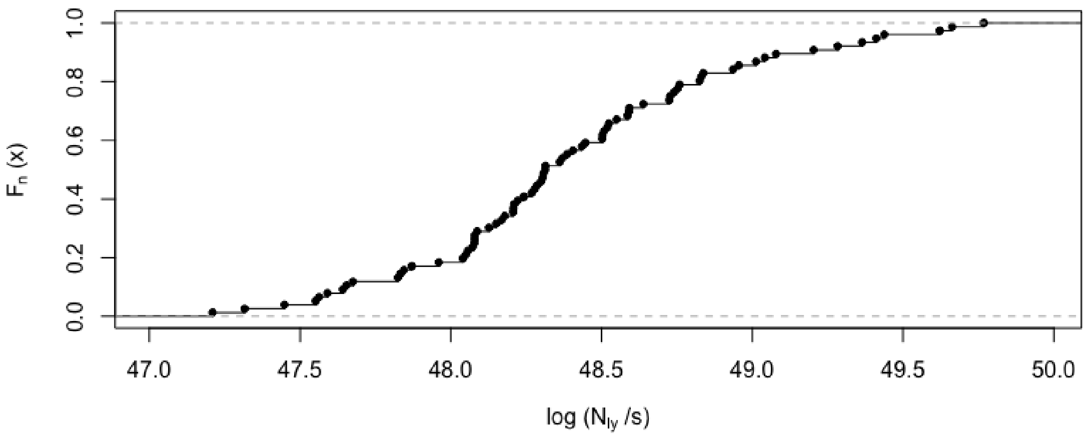
4.3.3 Lyman Continuum Flux

Using the integrated radio continuum flux density from within the boundary of each HII region, the number of Lyman continuum photons was estimated from the following equation (Matsakis et al., 1976), derived from the model developed by Mezger and Henderson (1967) and assumptions outlined by Rubin (1968) (see Sect. 2.3.2 for details):

$$N_{ly} = 7.54 \times 10^{46} \left(\frac{S_\nu}{\mathrm{Jy}}\right) \left(\frac{D}{\mathrm{kpc}}\right)^2 \left(\frac{T_e}{10^4\mathrm{K}}\right)^{-0.45} \left(\frac{\nu}{\mathrm{GHz}}\right)^{0.1} \tag{4.3}$$

where S_ν is the measured total flux density, D is the distance to the Sun, T_e is the electron temperature and ν the frequency of the radio emission. All regions are assumed to be optically thin and have an electron temperature of 8,500 K. For S_ν, the intensity values from the MAGPIS images are given as Jy/beam, the total emission was therefore extracted by first determining the beam integral ($1.133 \times \mathit{beam\ size}^2 \simeq 38.1''^2$) and dividing by the pixel scale, which gives the number of pixels in the beam ($\simeq 9.5$ pixels). Then, dividing the sum of the pixel values within the boundary of each region by the number of pixels in the beam yields the flux density in Jy (full details in Sect. 2.4.1).

Derived S_ν and estimated N_{ly} values for each of the 76 HII regions are listed in Table A.1 in Appendix A. These ionising flux estimates are lower limits due the assumption of ionisation bound regions. A small amount of ionising flux can escape into the outer PDR, and some flux may be attenuated by dust within the region. After identification of the probable ionising sources for a sample of HII regions,



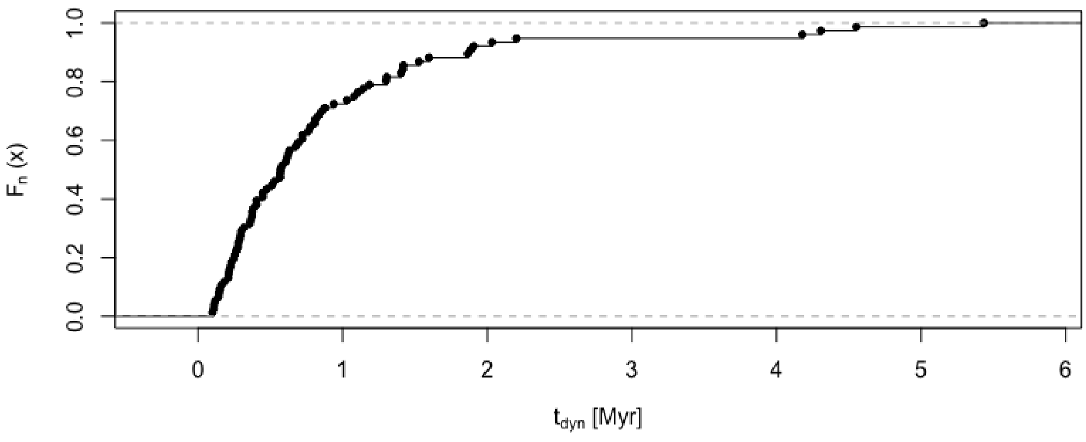


Figure 4.12: Distribution of the number of ionising photons, N_{ly}, within the boundary of each Hii region (top) and the dynamical age, t_{dyn}, of each region (bottom).

Watson et al. (2008, 2009) estimate that the N_{ly} values calculated from the MAGPIS images are a factor of 2 lower than the expected flux from the ionising stars. This was not corrected for in these calculations and instead all inferred masses are assumed lower-limits. Since this is systematic across the sample, estimated values are still representative of the sample distributions.

The distribution of N_{ly} is shown in Fig. 4.12, top. The log N_{ly} values range from 46.92 to 49.77, corresponding to single ionising stars of spectral type B0.5V and O5V, respectively (Panagia, 1973, Table II), with a mean of 48.4, corresponding to an O9V star. This distribution further confirms the expectation from C06 that many of the bubbles are the result of ionisation and stellar winds from late O and early B type stars. This was also found by Beaumont and Williams (2010) for a selection of the C06 bubbles, however, for the bubbles that are common between this sample and theirs, there is a higher estimate of N_{ly}, due to not assuming the near distance from the kinematic distance estimates.

Weidner et al. (2010) found that for star cluster masses exceeding 10^2 $M_\odot$, random sampling from the stellar initial mass function is highly unlikely, and that the star cluster mass is related to the mass of the most massive star it hosts. For a star cluster of 10^2 $M_\odot$, the corresponding most massive star has a mass of ~ 8 $M_\odot$. Since a single O9V star has a mass of ~ 19 $M_\odot$ (Weidner and Vink, 2010), and assuming that the majority of the ionising flux within each HII region is the result of the most massive star within the cluster, we can hence determine the minimum star cluster mass required for each HII region.

Figure 4.10, bottom-left, shows the distribution of N_{ly} values by group. There is no apparent association between group and the region's number of Lyman photons, with the most massive regions distributed across all groups except Group 5. The largest value of log N_{ly} in Group 5 is 48.9, which corresponds to a single O6.5V star. From Weidner and Vink (2010), an O6.5V star has a spectroscopic mass of $\sim$ 30 $M_\odot$, which would require a cluster of $\sim 10^3$ $M_\odot$ to produce (Weidner et al., 2010), this is hence the upper mass limit for Group 5. Group 1 contains only intermediate- to high-mass regions, with a lowest log N_{ly} value of 48.2, which is between an O9V and an O9.5V star, corresponding to a minimum cluster mass of $\sim 10^{2.5}$ $M_\odot$ for this group. In both Groups 1 & 5, there appears to be a break in the parameter space between intermediate and high mass clusters, and lower and intermediate mass clusters, respectively. However, splitting these groups each into two subgroups at their highest merge levels on the dendrogram in Fig. 4.3 does not delineate the groups by their respective N_{ly}. Although the group p-values for these subgroups are all $> 92\%$, it would not be the cluster mass in this case that warrants separating out

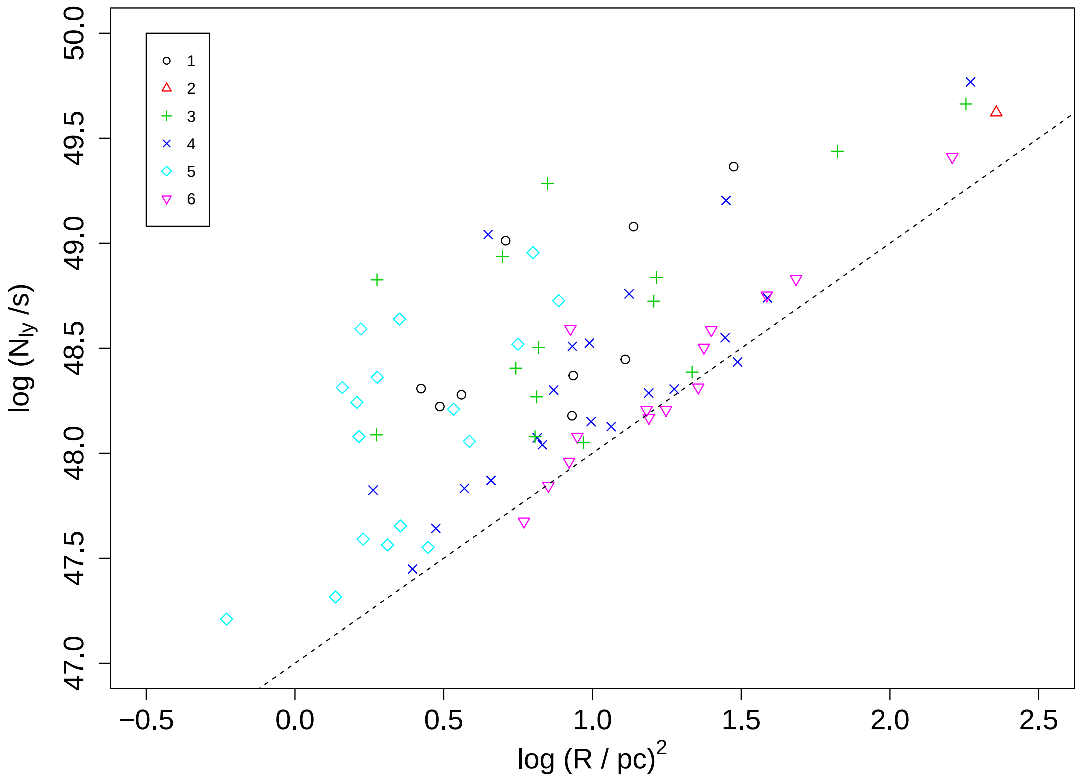


Figure 4.13: Physical effective radii of each HII region versus the number of ionising photons. Colours and symbols correspond to the groups identified from hierarchical analysis. The dashed line indicates a constant surface brightness of $1{\cdot}10^{47}$ photons / s per pc^2.

the two groups. Instead, these two groups are taken as representing their respective N_{ly} distributions, with no high mass regions in Group 5 and no low mass regions in Group 1. The remaining Groups 3, 4 & 6 all show a large spread in N_{ly} values, with Groups 4 and 6 each containing one region outside of their respective group's box-plot tails.

Figure 4.13 shows each region's N_{ly} values versus physical surface area. There is a clear power-law cutoff in the parameter space, indicating a limiting surface brightness of the MAGPIS data. For reference, the dashed line in Fig. 4.13 corresponds to a surface brightness of $1{\cdot}10^{47}$ photons/s per pc^2. Colours/symbols of points match those from the MDS plots in Fig. 4.6 and represent the identified groups. All but one of the regions within Group 6 appear to be at the detection limit of the survey, suggesting that for this group, the noise may be interfering with what we identify as the 'edge' of the HII regions. This is a possible explanation as to why Group 6 appears isolated in Fig. 4.3, not merging with the rest of the data until the largest height on the dendrogram. The fact that these objects have been distinctly grouped by the shape analysis method suggests that regions in the other groups that are also close to the limiting surface brightness are still at a sufficient signal/noise ratio that the edge has been correctly identified. Hence, any inferences made regarding the regions in Group 6 may not be as credible as those from the rest of the groups. This is because there may be a larger error in the determined diameter of the regions, and the regions may in fact belong in different groups if their edges were better defined.

4.3.4 Dynamical Age

Having determined N_{ly} for each H II region, their dynamical ages can be determined from the following equation (Spitzer, 1968; Dyson and Williams, 1980, see Sect. 2.3.3 for further details):

$$t_{dyn} = \left(\frac{4\,R_s}{7\,c_s}\right)\left[\left(\frac{R_{\text{HII}}}{R_s}\right)^{7/4} - 1\right] \tag{4.4}$$

where R_s is the radius of the Strömgren sphere ($= 3\,N_{ly}/4\pi n_0^2 \alpha_B)^{1/3}$, with n_0 the ambient particle number density, taken as $10^3\,\text{cm}^{-3}$, and $\alpha_B = 2.6 \times 10^{-13}\,\text{cm}^3\,\text{s}^{-1}$ is the hydrogen recombination coefficient to all levels above the ground level, c_s is the isothermal sound speed in the ionised gas ($= 11\,\text{km}\,\text{s}^{-1}$; Bisbas et al., 2009) and R_{HII} is the observed radius of the H II region (see Sect. 2.3.1 for further details on the Strömgren radius). Equation 4.4 is the result of analytical models that assume pressure equilibrium between the ionised and neutral shocked gas. Whilst this assumption is reasonable for standard H II regions, external pressure has a larger influence for (ultra- and hyper-) compact H II regions. Raga et al. (2012b,a) addresses this by considering the inertia of the shocked gas that is pushed out by the H II ionization front and present an updated model. The analytical solution of their model is approximately equal to Eq. 4.4 for a 'cold' surrounding medium. This further assumption is hence made for the ambient environment when using Eq. 4.4.

The distribution of dynamical ages for the sample of H II regions is shown in Fig. 4.12, bottom panel. The median value is at ∼ 0.6 Myr and only nine regions have an age greater than 1.6 Myr. This age distribution is on average lower than the one recently obtained by Palmeirim et al. (2017), whose ages were determined by comparing calculated ionised gas pressures to isochrones of 1D simulations by Tremblin et al. (2014), by ∼ a factor of two. The likely reason for this is the assumptions made here about the constant, cold ambient density (and to a lesser extent, the constant electron temperature assumed when calculating the ionising flux), which may vary with environment across the regions.

Figure 4.10, bottom right, shows the distribution of dynamical age by group. 84% (16) of the regions older than 1.1 Myr are assigned to Groups 4 and 6, with Group 6 only containing regions with ages > 0.5 Myr. Group 5 shows the smallest spread in ages, only hosting regions < 0.5 Myr old (22% (17) of all regions, 51% of regions < 0.5 Myr old). Groups 3, 4 & 6 host regions with ages exceeding the tails of their respective groups. These regions are discussed in more detail in the next subsection. The distribution of ages by group is similar to that of the effective radius. This is because the dynamical age has a stronger dependence on radii than ionising flux, hence regions with a large radius remain as outliers in the age distribution.

Despite the associations outlined here, the distributions of radius, ionising photons and dynamical age do not match the MDS scores of the regions along axis 2. This means that the 30% variation in the distance matrix that axis 2 represents cannot be associated to any one physical parameter investigated here. It is likely due to another feature of the curvature distributions that is not clear from the EDF plots. Since the A-D test is sensitive to the tails of the distributions it may be some subtle differences inherent there. Nevertheless, the absence of low- and high-mass regions in two groups, the result that one group only has small, young regions and that one group is at the surface brightness limit of the survey has all been revealed by the shape analysis method presented here. This shows good evidence that there is a link between shape and physical properties of the regions/environments. It is also further encouragement for developing these results towards a training set for future data sets.

4.3.5 Outliers

The apparent outlier from the hierarchical clustering is the single region in Group 2. HII region G012.429-00.049 is the most distant region in the sample (with a heliocentric distance of 22.6 kpc), and from its line of sight radial velocity, is outside the Solar circle and hence does not have a any kinematic distance ambiguity (V_{lsr} = -18.4 km s^{-1}). This region has the sixth largest angular diameter in the sample, which - combined with its far distance - leads to it having the largest spatial diameter in the sample (~ 30 pc). This meant that this HII region had by far the largest number of interpolation points along its boundary. This may be the reason for this HII region having a large A-D test statistic score when compared to all other regions. However, given its size and distance, this HII region is the oldest in the sample at 5.4 Myr. It may simply be because of how evolved the region is that it has been selected as an outlier from the rest of the sample. We see from the region image in Appendix B that the MIR bubble is not in a complex, and the radio continuum emission at the

1σ boundary extends further than the 8 μm shell. This is not common to the sample and suggests a potential projection contamination with another radio source. In the statistical study of Spitzer bubbles by Hou and Gao (2014), this HII region, from Chini et al. (1987) and Lockman (1989), is matched to the same MWP bubble as in this work. There is also a separate UCHII region within this bubble, listed in Wood and Churchwell (1989).

Further outliers have been identified from Fig. 4.10, assuming that the region shapes are linked to the physical parameters discussed. Region G027.476+00.179 in Group 4 has a determined age of 4.2 Myr and also the largest value of N_{ly}. This far exceeds the interquartile range (IQR) of the respective properties of regions in Group 4. The distance of this HII region is quoted as 12.6 kpc, which is the far kinematic distance and is a very secure assignment (Jones and Dickey, 2012). Similarly, an outlier in Group 3, region G019.629-00.095, has an age of 4.3 Myr, which was also determined with the far kinematic distance of 11.7 kpc and is another very secure distance assignment (Sewilo et al., 2004). The other outlier in Group 3, region G025.397+00.033, has an age of 2.0 Myr at a distance of 17.3 kpc, which is outside of the Solar circle. Since in each of these cases, there is either a good resolution to or no KDA, it is suggested that these could be examples where the incorrect assumptions have been made regarding the ambient density or temperature. For example, the images of region G025.397+00.033 in Appendix B show an elongated area of radio emission, extending away from the centre of the MIR bubble. If a lower density of $\mathrm{n} = 500\,\mathrm{cm}^{-3}$ is assumed for this region, its age is then calculated as 1.4 Myr, which is just within the upper extreme of that group's age values. Group 6 hosts the second oldest region with the fifth highest N_{ly}, G035.649-00.053 (N68 from C06), which at a far kinematic distance of 10.4 kpc is calculated as 4.6 Myr old. As discussed previously, Group 6 has been identified as regions on the limit of surface brightness detection, hence there is a larger error in the diameter determination of these regions which would influence the calculated dynamical age.

4.3.6 Applications

When testing the shape analysis functionality of the method, regular shapes such as circles, ellipses and regularly perturbed closed curves were compared to the observed sample. In each case, the test regions were ordinated far away from the observed regions on the MDS plots, and were grouped in their own branches on the dendrograms. Hence, the first proposed application of this analysis method would be to test the efficacy of synthetic observations of numerical simulations of HII regions. This is the main focus for Chap. 5. Preliminary results showed that modern simulations do indeed produce synthetic images that are comparable to the observed data. Hence, for varying initial conditions and at different simulated ages, we can determine when the synthetic regions appear in the groups that we identify. Furthermore, we can take advantage of these controlled parameters to better quantify the groups and find out when a region would move between groups. We have already seen evidence for regions moving systematically between groups from our resolution testing. Ultimately this would help lead to a supervised classification scheme of HII regions based on their shapes, whereby each class is associated to a specified range of physical properties.

A further use of the synthetic observations is to consider different projections of a region at a given stage, to see how this affects the obtained shape of the region. Although in this chapter, a variety of HII region shapes were considered, it remains to be seen from this alone whether or not the line of sight angle to an individual source would have a large influence on its shape. Throughout the testing of the analysis, it was noted that a region was more likely to move between groups if its shape changed in a definite manner, such as more detail being captured at a lower contour level, or a higher spline resolution. We can hence speculate that if no greater amount of fine detail was revealed by a change in projection, then the region should remain in the same group when viewed at different angles.

There are various other diffuse astronomical objects that share common morphologies, such as planetary nebulae, supernova remnants, giant molecular clouds, and cold molecular clumps. If the shape data of these objects can be extracted in a systematic manner, like what was carried out for the HII regions in this work, this analysis method could be applied to these objects. Due to the curvature shape descriptor and the non-parametric statistical tests used in this method, the regions do not necessarily have to be closed contours.

4.4 Conclusions

An unbiased shape analysis of a sample of 76 HII regions was performed here by analysing the curvature distributions of the regions boundaries. These were obtained systematically at a constant signal level above the background noise in each image from the MAGPIS radio continuum data. By applying hierarchical clustering and multi-scale bootstrap re-sampling to the data, six groups were obtained and verified, delineating HII regions of similar morphologies. This was confirmed by visual inspection of the images, and quantitatively by the ordinance technique of multidimensional scaling. 97% of the variation in the shape data is represented by three principal coordinate axes, 60% of which is likely due to the amount of high curvature points along a region boundary, i.e. the level of curvature perturbation. Investigation into the association of physical parameters and the group assigned by the methodology revealed the following results:

There was little association between the region position in the Galaxy and assigned group, with objects at varying Galactocentric distance and Galactic latitude appearing in all groups. We found that one group contained regions with Galactic longitude $< 30^\circ$ and another had only regions with $l > 28^\circ$. However, the remaining three groups contained regions with the full distribution of l values. This suggests that for this section of the Galactic Plane, HII region shape is homogeneously distributed across Galactic latitude, with a small preference for those at similar Galactic longitudes to share a common morphology.

One of the six groups contained only small (<1.6 pc), young (<0.5 Myr) HII regions. This was not exclusive, with 59% and 51% respective completeness. It does show that the size-and-shape information contained within the curvature distributions and the statistical clustering of the data reflects this common feature for this group. There was also a preference for this group to contain only low- to intermediate-mass ionising clusters (maximum of $\sim 10^3\,\mathrm{M}_\odot$). There was a further preference for ionising mass in another group that contained only intermediate- to high-mass clusters (minimum mass $\sim 10^{2.5}\,\mathrm{M}_\odot$). There was no further preference for regions of a given ionising mass to be placed in any particular group. Using these results, five outliers in the sample were discussed and whether there were possible projection effects or incorrect assumptions made as to the ambient density or electron temperatures. In each case there was either good resolution for or no kinematic distance ambiguity.

Another of the identified groups was distinctly separated from the rest of the data by the analysis. It was revealed to contain regions that were at the surface brightness detection limit for the survey. This suggested that the noise may be interfering with the extraction of the boundary of these regions, and that the shape of these objects is less accurately represented than in the rest of the data. For a deeper survey, these regions may in fact belong to a different group in the analysis. The groupings obtained remain present if this (or any individual) group is removed from the data set. The group associations also remain if a random number of regions are removed from the data. This further illustrates that common morphological features were readily identified by the method.

It was evident from testing the methodology that certain selection choices affect the resulting group structure from the hierarchical process. Lowering the threshold level used to define the ‘edge’ of the HII region resulted in higher curvature portions of the shape to be included in its descriptor. In some cases this was due to the noise in the images, in others it can be attributed to structure of comparable scale to the resolution used. This had a much larger influence on the group results than increasing the threshold value, concurring with the MDS result of the number of high-curvature values along the boundary having a large influence on the groupings. Increasing the threshold produced generally more smoothed edges with less defined detail. The p-values (from the bootstrapping) associated with the groups obtained in each case were lower than those obtained for the 1σ value used. For the future work, the optimum threshold level to use will be investigated, for varying noise profiles that can be readily controlled with synthetic observations. The synthetic data will also be used to test whether line of sight projection has a significant influence on the observed planar shape.

The largest source of error for the shape analysis was due to the sampling resolution used to obtain the curvature distributions, which is determined for each region from its distance. Using the spatial sampling resolution corresponding to the image resolution at the far distance (0.54 pc at 19.2 kpc) removed any bias that is attributed to nearby regions. The results show no association between heliocentric distance and group at this resolution. However, as with the threshold levels, it was found that changing the regions distance by ±20% (which in turn changes the sampling resolution) did affect the groups from the hierarchical process. Due to the errors associated with kinematic distance determination, this is a limiting factor of the analysis process. 18% of group assignments changed when the distances are reduced, which corresponds to reducing the sampling resolution, and 53% of assignments changed when the resolution is increased. It was noted that changes to the shape descriptor of individual regions would inevitably lead to the region being grouped with regions of different morphologies. Furthermore, in some cases, multiple regions from a given group were each relocated to the same new group after changing the resolutions, suggesting that the shape descriptors are affected in a common manner by the resolution change. In the following chapter, these inherent deviations in the description of the shape of HII regions, which arise from both selection choices and observational errors, are investigated to see if they can be quantified in a fully systematic way; this would allow for a classification scheme to be constructed from the morphological analysis presented here.

From many levels of abstraction, the results show good evidence for associating HII region shape to the region’s physical parameters, and that shape can be used as an intrinsic measure. With such a large amount of high resolution images readily available, there are many potential applications of this approach to larger samples of HII regions, and other astronomical objects. It is worth reiterating, however, that for diffuse objects, one must have a clear definition for the shape they extract, and execute this in a systematic and repeatable manner.

The next chapter focuses on the comparisons of synthetic observations of HII regions, with given initial conditions and projections, to the observed sample detailed in this chapter. The initial conditions of the simulated HII region are investigated, to see how they affect the extracted shapes. This will also allow for further quantification of the groups, moving towards a supervised classification scheme for HII regions.

Chapter 5

Synthetic Observations

This chapter details the application of the shape analysis methodology to a selection of HII regions from numerical simulations. The synthetic observations of the numerical simulations were provided by Dr Ahmad Ali and published as Ali et al. (2018). The analyses of said synthetic observations, which are presented herein, were carried out by the thesis author.

5.1 Introduction

Since the advance in high powered computing in the latter part of the 20th century, astrophysicists have utilised these tools to perform numerical simulations of all aspects of the Cosmos. From modelling how a star is formed and begins to fuse hydrogen, to galactic formation and evolution, all the way to entire cosmological models that reflect the largest scale structure astronomers have ever observed. In this chapter, the statistical shape analysis method (described in Chap. 3) is applied to a set of Synthetic Observations (SOs) of numerical simulations of HII regions. The shapes of these synthetic HII regions are then compared to those of the MAGPIS sample (which were discussed in Chap. 4). The efficacy of the numerical simulations will hence be tested using the shape analysis method. The different parameters such as noise and projection angle will then be investigated to determine how they affect the identified shape of the HII regions. Finally, we will discern whether such SOs can be used as a training set in a supervised morphological classification scheme of HII regions.

The term 'numerical simulation' covers a broad range of techniques in astrophysics. Essentially, it refers to any kind of analytical or computer model, whose aim is to simulate an astrophysical phenomenon. The analytics may involve gravitational dynamics, when simulating the orbits of planets around a star; or hydrodynamics, when simulating the convective layers of stellar structure. Different simulations with different purposes, hence include many aspects of physical simulation that result in predictions of the object in question. Radiation plays a dominant role in astrophysics. The transport of radiation through the ISM is therefore one of the most fundamental processes to be considered when modelling stellar objects and galactic structures. Analysing the radiation from an object not only tells us about the nature of the radiation source, but also the medium through which it has travelled to reach us. Interstellar dust therefore also plays an important role in the study of radiation, since it scatters and reradiates UV through to IR photons.

In the last decade, a number of radiative transfer (RT) models have been used to generate idealised observations of the numerical simulation they relate to. The extensive review by Steinacker et al. (2013) lists 28 distinct astrophysical RT codes, used previously and at the time of the review. The RT codes work by sampling the simulation at every grid point, for the given dimensionality of the simulations. For each grid point, the RT code samples the probability distribution functions that a photon propagating through that grid point gets scattered, absorbed, or reradiated; depending on the constituents of that space in the grid. Temperatures can thus be calculated and flux images can be generated from any viewing angle the observer specifies. Such synthetic observations are referred to as 'ideal' synthetic observations by Koepferl and Robitaille (2017), who developed a tool called `FluxCompensator`[1]. This tool aims to account for the observational affects when receiving astronomical radiation, i.e., the capabilities of the telescope, optics and detectors. Their pipeline produces realistic SOs by accounting for aspects such as resolution, pixel scale, point spread function and noise profiles. The `FluxCompensator` database predominantly deals with IR observations. Hence, in this work, a bespoke processes for converting the 'ideal' SOs to realistic SOs is constructed, so that they are comparable to the MAGPIS survey.

[1] Available at https://github.com/koepferl/FluxCompensator

Table 5.1: Table 2 from Ali et al. (2018): Initial parameters of the massive star in the numerical simulation.

Parameter	Value
Mass	33.7 $M_\odot$
Luminosity	1.49×10^5 $L_\odot$
Radius	7.59 $R_\odot$
Effective temperature	41 189 K
Ionizing flux ($h\nu \geqslant 13.6$ eV)	7.36×10^{48} s^{-1}

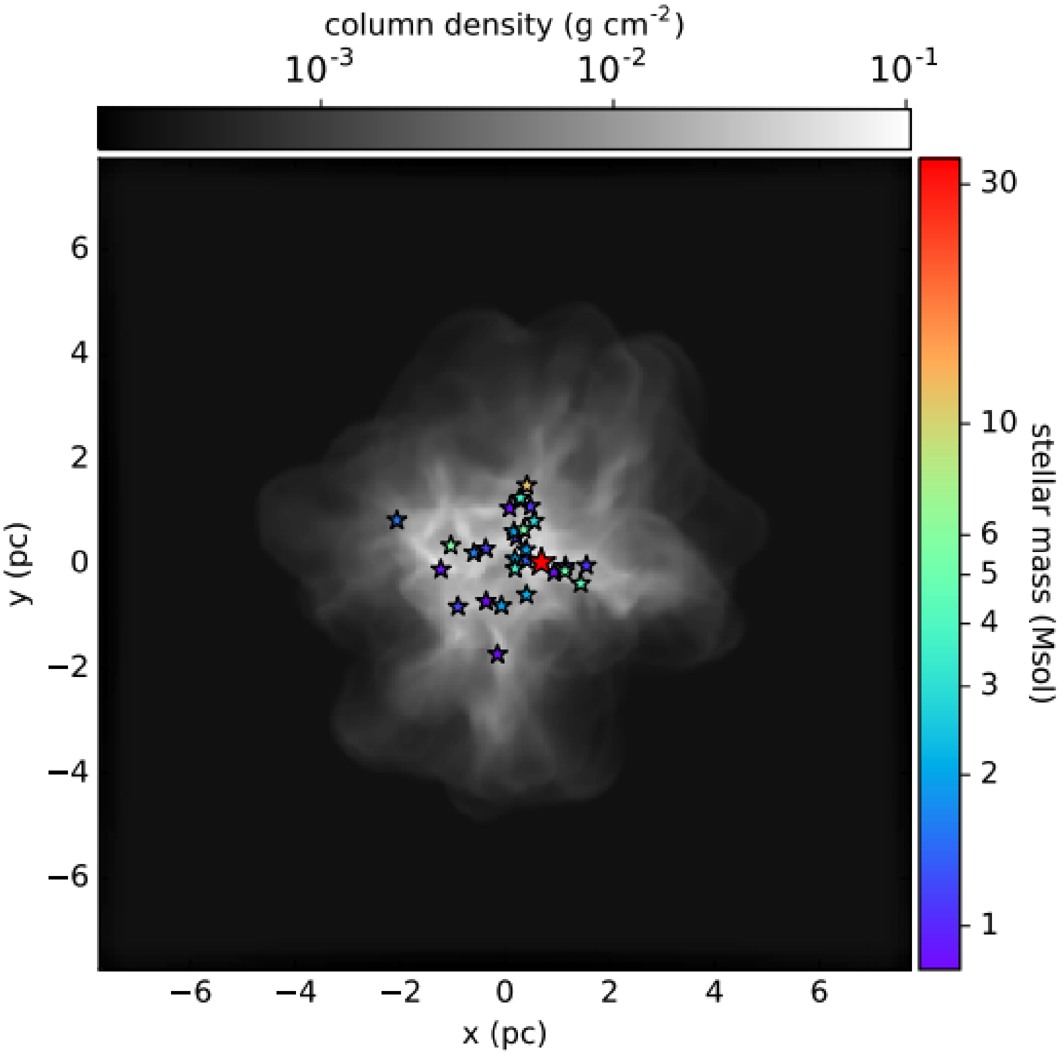


Figure 5.1: Figure 1 from Ali et al. (2018): Positions of stars at the onset of feedback, with stellar mass in colour scale, overlaid on column density in greyscale (both are logarithmic). The most massive star is 33.7 $M_\odot$ in red. The second highest is 11.3 $M_\odot$. The third is 5.7 $M_\odot$. The least massive is 0.82 $M_\odot$.

5.1.1 The Synthetic Observation Data Sample

The synthetic observations of HII regions used in this study are from the numerical simulations of Ali et al. (2018). The simulations are performed using the radiative transfer (RT) and hydrodynamics (HD) code `TORUS` (Harries, 2015). The simulations are comprehensive, including photoionisation balance, thermal balance, radiation pressure, interstellar radiation, hydrodynamics and stellar evolution tracks. The processes and steps these numerical simulations take to arrive at the numerical HII region are detailed in Sect. 2.5. To grossly oversimplify an entirely complex procedure: for the given initial conditions outlined below, a spherical cloud of gas evolves under gravity and turbulence for 75% of the mean free fall time for such a cloud. At this time, stars are added from a random sampling of the Salpeter (1955) IMF, such that the cumulative stellar mass is 10% of the cloud mass. The most massive star is placed at the cloud's most massive clump, with the others spread around, following a probability density function that accounts for the star formation rate. Then, at this stage, the radiation field is switched on and the simulation evolves until all of the mass leaves the grid. Voila, a numerical HII region is born.

The initial conditions for the numerical simulation in Ali et al. (2018) are as follows: The grid is a 3D cube of resolution 256^3 pixels. The spherical cloud has a uniform density inner core that extends to half the sphere radius, with the outer half going as $r^{-1.5}$. The density outside the sphere is 1% of the density at the edge of the sphere. The sphere has a total mass $M = 1000\ \mathrm{M}_\odot$, radius $R = 2.66\,\mathrm{pc}$, and mean surface density $\Sigma = 0.01\,\mathrm{g\,cm^{-2}}$. The size of the grid is approximately 15.5 pc, yielding a resolution of 0.06 pc per pixel. The initial mass, luminosity, radius, effective temperature and ionising photon rate of the massive star are listed in Tab. 5.1. The initial distribution of stars is shown in Fig. 5.1, overlaid on an image of the column density of the cloud. It is at this point that the radiation field is switched on and the simulation evolves. Whilst the initial condition of the cloud is spherical, we see from the position of the most massive star that this is not at the centre of the grid. This, together with the random turbulence distribution, could lead to differences in the shape of the ionised region when viewed from different projections. This will constitute part of the focus later in this chapter.

Figure 5.2 shows the bulk grid properties of the simulation against time. Properties shown are: total mass on grid, mass flux off the grid, maximum density, ionised mass and mass fraction, and ionised volume fraction. We see here that from the onset of feedback ($t = 0$) to 0.6 Myr, there is a steady mass flow, with mass beginning to leave the grid at ∼0.4 Myr. After 0.6 Myr, the overall mass flux begins to decrease. Spikes in the distribution correspond to removal of the clumps. The size

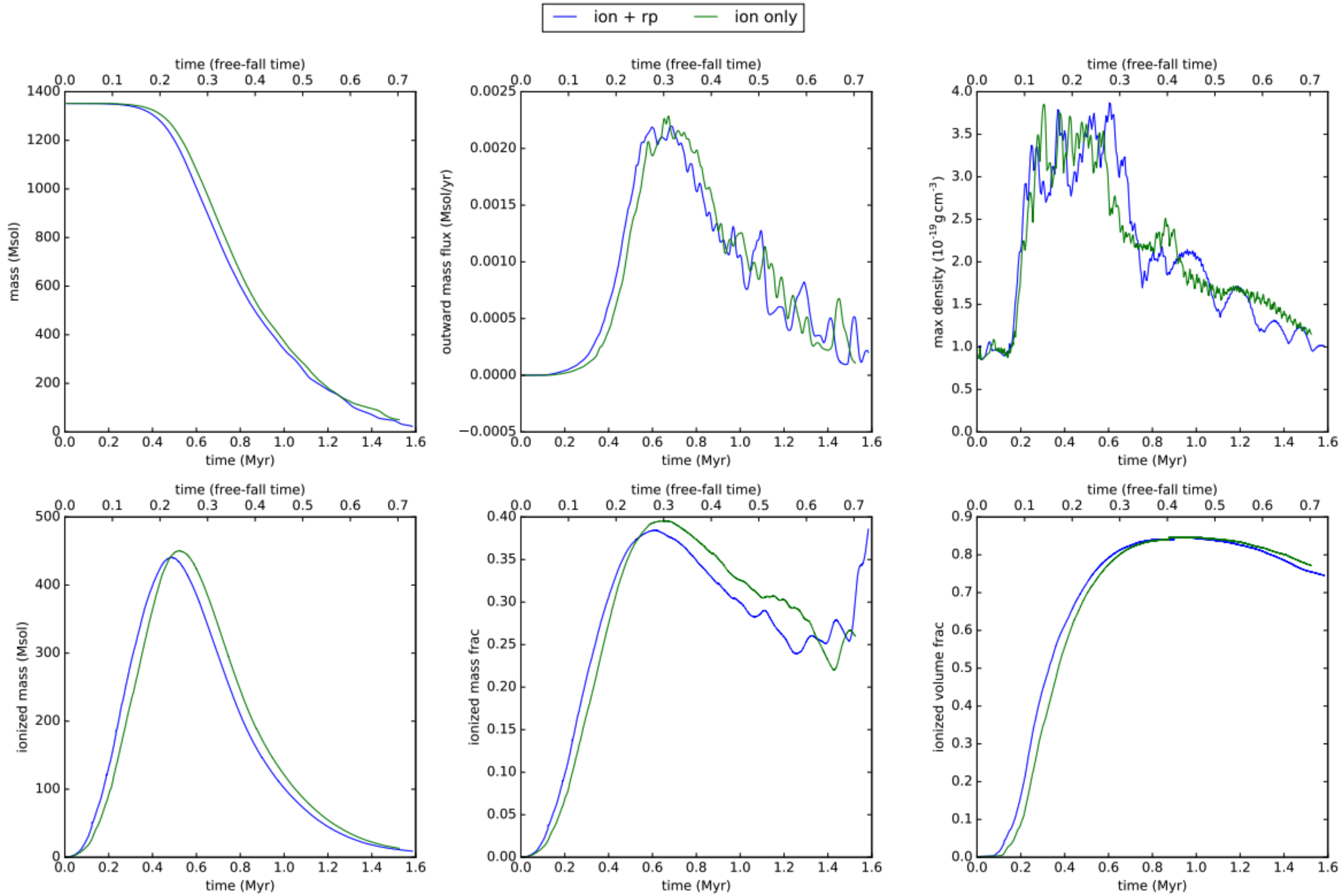


Figure 5.2: Figure 3 from Ali et al. (2018): Bulk grid properties as a function of time, showing total mass, mass flux off the grid, maximum mass volume density, ionised mass, ionised mass fraction, and ionised volume fraction. The blue line is the model with ionisation and radiation pressure; the green line is only ionisation. $t = 0$ corresponds to the onset of feedback.

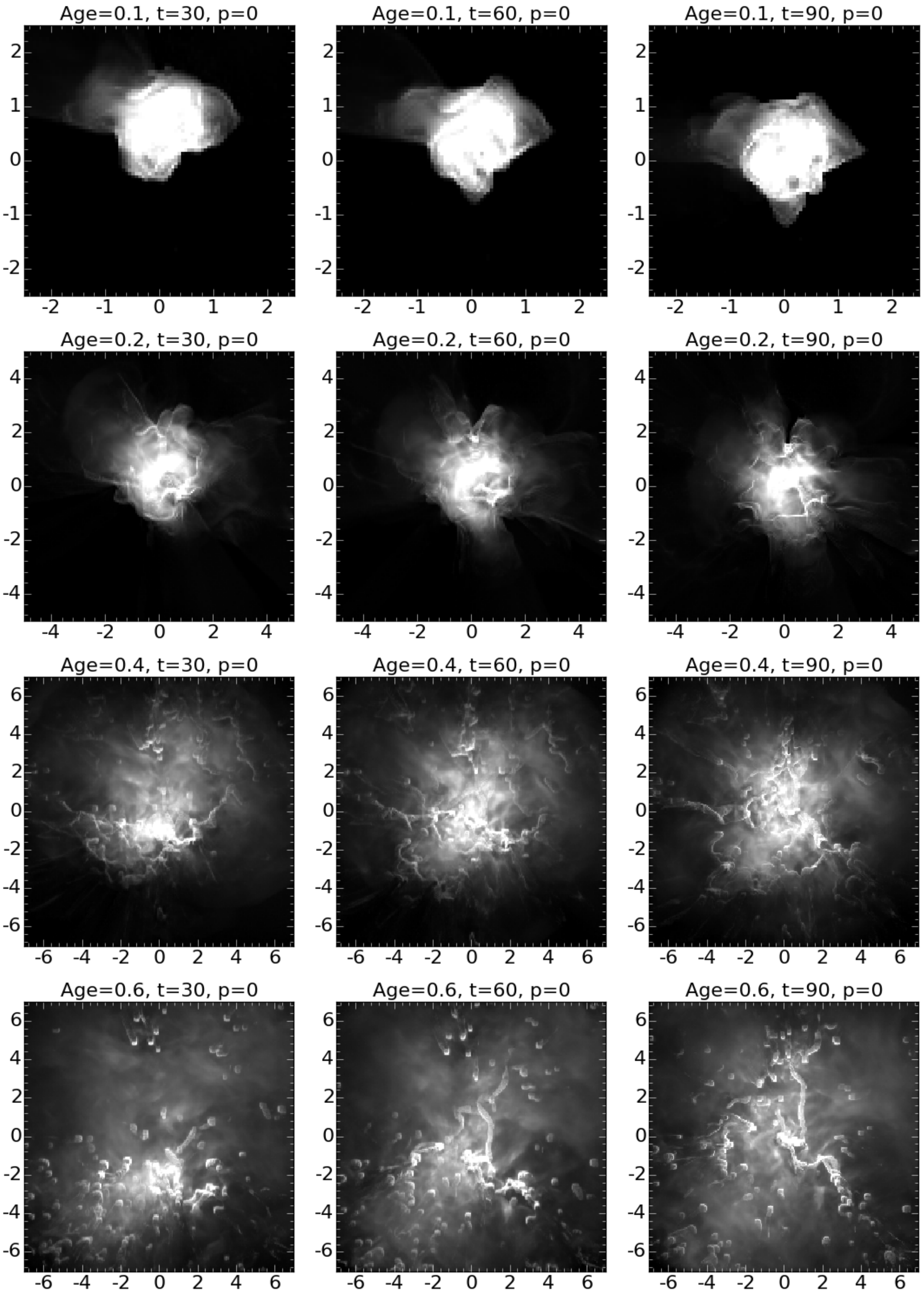


Figure 5.3: Overview of the 20 cm, 1.4 GHz synthetic observations from the numerical simulation in Ali et al. (2018). Three snapshots of varying ϕ projection angle are shown for the four evolutionary time-steps of 0.1, 0.2, 0.4 and 0.6 Myr. Coordinates are shown in pc scale.

of the spikes grows with time, as the densest clumps are the last to leave the grid. By ∼1.6 Myr, or 0.74 $\langle t_{\rm ff} \rangle$, all the mass has left the 15.5 pc^3 grid. The peak value of ionised mass is 440 M$_\odot$ at 0.5 Myr. The peak ionised mass fraction, which is just under 40% of the total mass, is reached at 0.6 Myr. At this time, the fraction of volume ionised is ∼80%, showing that the majority of the neutral gas remains in the small, dense clumps, which resist the ionisation.

The synthetic observations of the numerical simulations were produced using the temperatures, densities, dust properties, elemental abundances and ionisation fractions that were calculated and evolved during the RHD model with both photoionisation and radiation pressure. Snapshots can be taken at given simulation ages and from any ϕ and θ spherical viewing angles. Simulation ages of 0.1, 0.2, 0.4 and 0.6 Myr were considered in this work. For each of these respective ages, 18, 22, 19 and 18 projections were provided, resulting in 77 synthetic observations of the numerically modelled HII region.

Figure 5.3 shows 12 example SOs of the 20 cm radio continuum emission produced by the simulations. Three example projections are shown for each of the evolutionary stages. In this example, the ϕ angle is kept fixed and the θ angle is 30, 60 and 90 °. The axes for each image are in parsecs. These, and the other SOs, provide the basis of the following shape analysis and statistical clustering. Before we delve into said analysis, we need to consider how to extract the shape of the synthetic HII regions, in a manner complementary to the MAGPIS HII region shape extraction.

5.2 Shape Extraction

The first purpose of investigating the SOs in this work is to test the efficacy of the simulations by comparing them to the MAGPIS observations; and then look at how different parameters affect the shape. The extraction of the shape of the HII regions hence needed to be performed in the same manner as it was for the MAGPIS observations. Therefore, the 'ideal' SOs needed to be processed into realistic SOs, with properties matching the observational MAGPIS sample. The first step to achieving this was to convert the intensity units of the SOs from MJy/sr to those used in the MAGPIS observations, Jy/beam (see Sect. 2.4.1 for details of the MAGPIS beam size). The solid angle conversion followed 1 [deg^2] = $(\pi/180)^2$ [sr], and the flux conversion was simlpy 1 Jy = 10^{-6} MJy. Assuming the SOs are at a distance of 6.2 kpc results in an angular pixel scale of 2$''$, also matching that of the MAGPIS data. The reason for performing this conversion was so that Gaussian noise could be added to

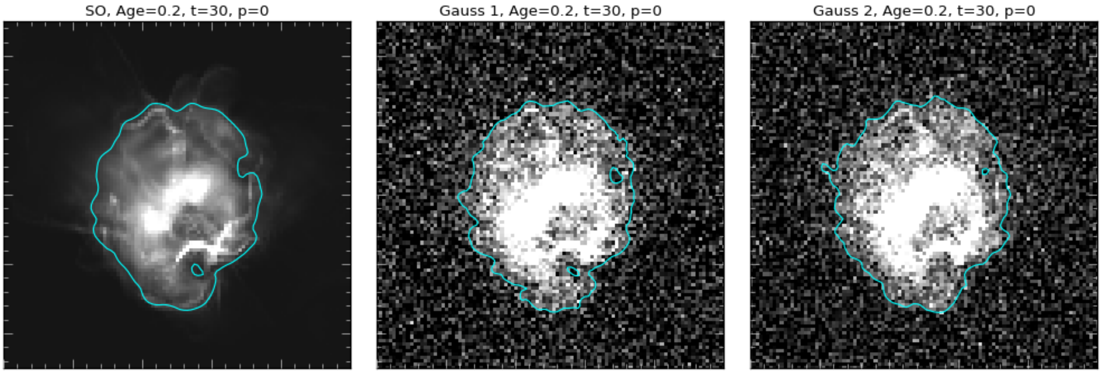


Figure 5.4: Comparison images of an example SO projection at 0.2 Myr. Right: Original SO. Middle: Same SO with random Gaussian noise added to each pixel value. Left: As middle but with a different Gaussian distribution used.

the SOs in order to perform the contouring procedure to identify the 'edge' of the HII regions, as was carried out for the MAGPIS data.

5.2.1 Gaussian Noise Profiles

For the MAGPIS data, the shape of each HII region was extracted using the image contouring procedure outlined in Sect. 3.1.3. After applying sigma clipping to all of the pixel values, which removed the signal from the ionised emission, the mean and standard deviation were taken from the remaining clipped values. This provided the contour value to apply to the original images to systematically define the HII region boundaries. In order to carry out this procedure on the SOs, noise was introduced to the SOs by taking a random value from a Gaussian distribution with mean, μ, and standard deviation, σ, from one of the MAGPIS tiles, after sigma clipping. The standard probability density function for the Gaussian distribution was used:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-\mu)^2 \big/ 2\sigma^2} \tag{5.1}$$

Figure 5.4 shows an example of one of the 0.2 Myr SOs. The left panel shows the original SO, the middle and right panels show the same SO with the random Gaussian noise added to each pixel. The middle panel has its Gaussian profile taken from the MAGPIS tile centred at l = 12.43°, b = -0.04°, with μ = 0.000959 Jy/beam and σ = 0.000389 Jy/beam. The right panel's Gaussian noise profile follows the MAGPIS tile centred at l = 30.25°, b = 0.24°, with μ = 0.000830 Jy/beam and σ = 0.000345 Jy/beam. The contours for the two SOs with Gaussian noise are determined by taking the clipped mean plus 1σ, the same way in which each contour was calculated for the each of the HII regions in the MAGPIS data. The level for the

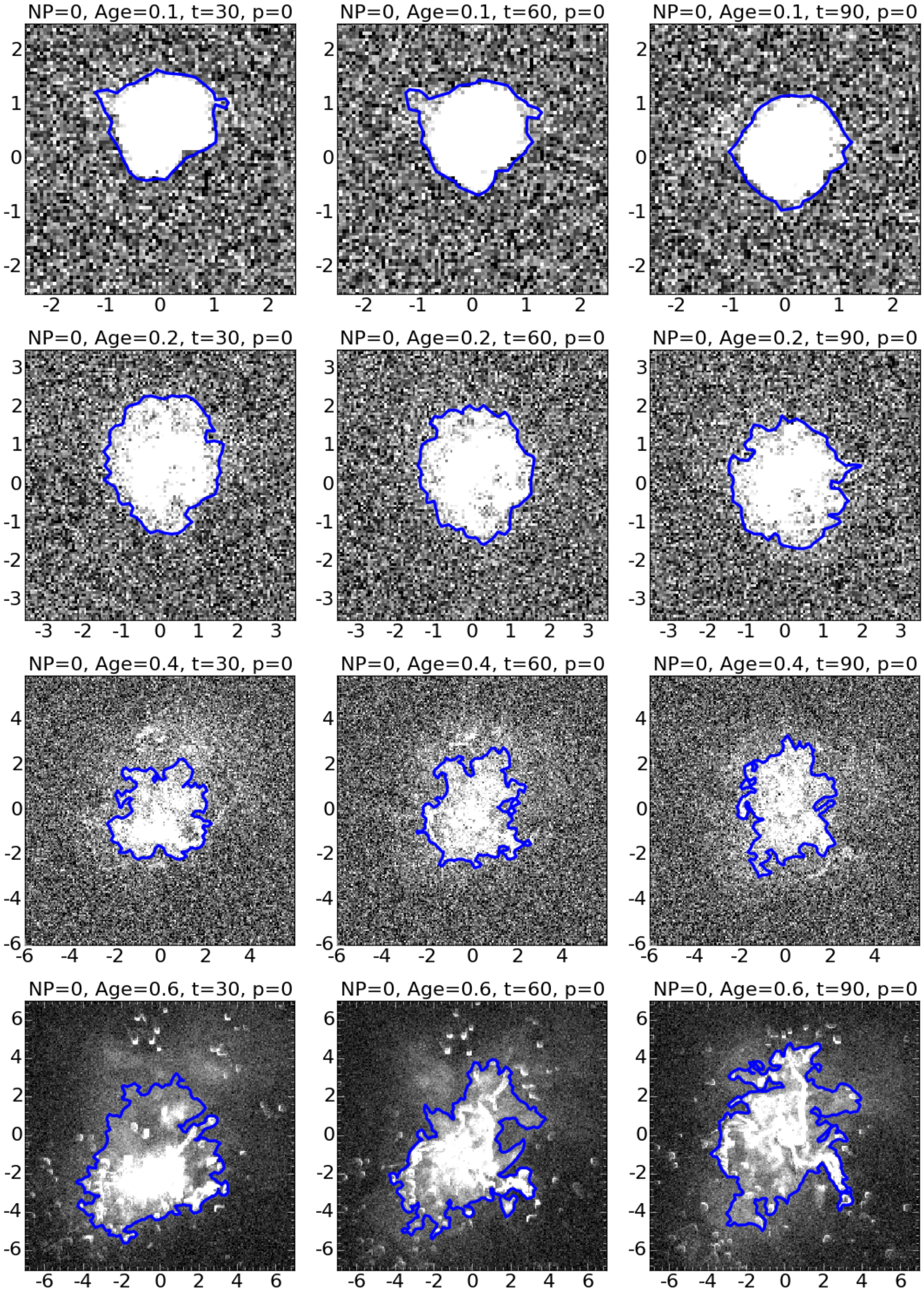


Figure 5.5: Overview of the addition of random Gaussian noise to the same SOs shown in Fig. 5.3. Contours are shown at a constant level of 1σ plus the clipped mean of the values from the Gaussian distribution introduced.

contour of the original SO in the left panel is the same as the middle panel, since the pixel vales of the SO had already been converted to match that of the MAGPIS data. There is not much visual difference between the three contours. The original SO has a much smoother contour as expected. Small perturbations along the boundaries of each of the Gaussian profile SOs are noted, they are approximately the same size. Whilst some extrusions along the boundary appear in the SOs with the Gaussian profiles, the intrusion on the original SO on the upper right side is smoothed out by the Gaussian noise.

The conversion of MJy/sr to Jy/beam was suitable for the 0.1 and 0.2 Myr SOs. However, for the 0.4 and 0.6 Myr SOs, using the actual conversion factor resulted in the emission from the SOs being drowned out by the introduction of the Gaussian noise from the MAGPIS data distributions. This is explained by referring back to the bulk properties of the simulations in Fig. 5.2, which shows that after 0.4 Myr, mass starts to leave the grid of the simulation. Therefore, the integrated radio continuum intensity at 20 cm, which is used to produce the SO, is less than should be expected if all the mass were considered. The intensity of these older SOs in Jy/beam was hence artificially boosted by a factor of ~ 4 for the 0.4 Myr SOs and ~ 70 for the 0.6 Myr SOs. This was the minimum increase required for each to yield a single contoured central region, representative of the original SO, before the noise was introduced.

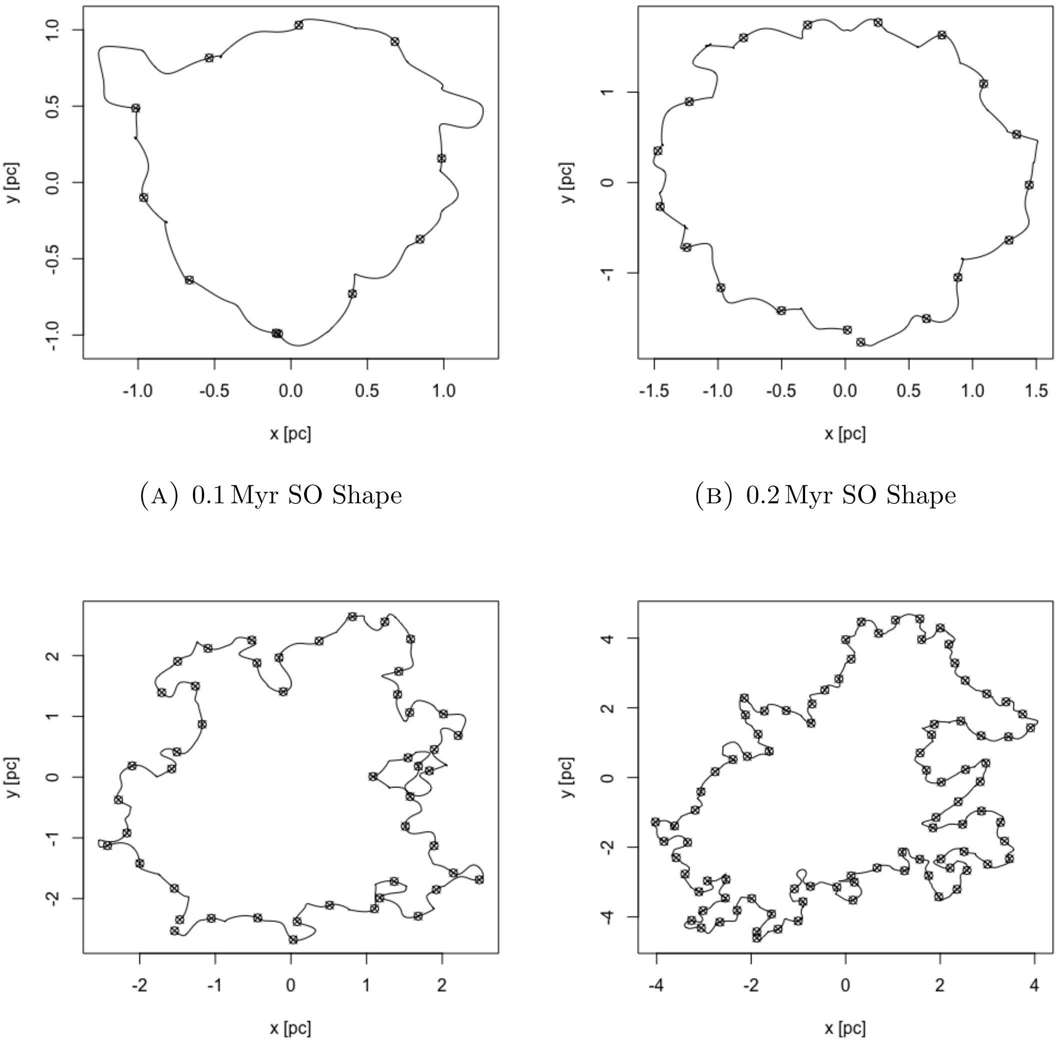
Figure 5.5 shows a summary of the addition of a Gaussian noise profile to the same 12 example SOs shown in Fig. 5.3. The images for all 77 SOs with this Gaussian noise profile are shown in Appendix C. The noise profile used was that of the middle panel in Fig. 5.4, with the two younger SOs having the actual intensity conversion factor used and the older two have their emission boosted as described above. It is clear from this summary that as the age of the SOs increases, so does the amount of perturbation along the HII region boundaries. It appears that the contoured shape changes more with each projection for the 0.4 and 0.6 Myr regions than for the 0.1 and 0.2 Myr regions (however, larger perturbations are noted for some of the different projections for the earlier ages in the full overview in Appendix C). The axes of the plots are given in spatial parsec scale, hence the effective radius of each HII region boundary also increases with the age of the SO. Only one Gaussian noise profile was used for all of the 77 SOs, since the shape did not to change by a substantial amount with the introduction of noise with different Gaussian distributions from the MAGPIS tiles. This was tested using the shape comparison method, resulting in lower T_{AD} test scores than for any of the pairwise scores achieved thus far in this work. Having the same noise distribution also meant that the contour level applied to each SO was the same, since this is calculated from the noise itself - distinguishing the signal of the HII region. Having now defined the systematic extraction of the shape of the SOs, conducive to that of the MAGPIS data shape extraction, we can now begin to compare the shapes of each.

5.2.2 Shape comparison: SO & MAGPIS

The first test of the SO data was to directly compare them to the MAGPIS observational sample from Chap. 4. With this in mind, the further steps to extract and compare the shapes of the regions were carried out in the same manner as it was for the MAGPIS data in the preceding chapters (full details in Chap. 3 & Sect. 4.2). To summarise these steps: interpolation splines were fitted to the region boundaries from the SO Gaussian profile images, with interpolation knot intervals of $\sim 0.54\,\mathrm{pc}$ (see Fig. 5.6); then the curvature values were calculated at each knot. The curvature empirical distribution functions (EDFs) were then statistically compared pairwise, using the Anderson-Darling (A-D) test statistic, providing the shape-distance between all regions. After applying a Euclidean distance transformation to the A-D test scores, hierarchical clustering was performed on the distance matrix of HII region shape distances using Ward's agglomerative method. The resulting hierarchical structure was then investigated using the dendrogram graphical representation.

In this primary investigation of how the shapes of the SO HII regions compare to those of the MAGPIS HII regions, the 12 example SOs from Fig. 5.5 were considered, along with the 76 MAGPIS HII regions from the previous chapter. Figure 5.7 shows the resulting dendrogram from the hierarchical clustering. The branches of the SOs are highlighted by colours that correspond to the age of the SO: 0.1 Myr in red, 0.2 Myr in pink, 0.4 Myr in blue and 0.6 Myr in green. The first clear result is that the shapes of the SO HII regions are grouped in amongst the MAGPIS HII regions, with none appearing as outliers. By ensuring that the shape of the SO HII regions was extracted and quantified in the exact same way as the MAGPIS sample, this result conclusively shows that the numerical simulations are producing HII regions that are representative of what we observe in our Galaxy.

The next point to note from Fig. 5.7 is that the three different projections considered for each SO age are not all grouped together in the dendrogram. The 0.1 Myr projections are close to each other on the dendrogram, however, one of them belongs to a different parent group than the other two. Similarly, for the 0.2 and 0.4/,Myr projections, two each belong to the same parent group, with the third positioned a few groups away, respectively. The most spread in group allocation is seen for the 0.6 Myr projections. These results display that for each of the SO projections,



(a) 0.1 Myr SO Shape

(b) 0.2 Myr SO Shape

(c) 0.4 Myr SO Shape

(d) 0.6 Myr SO Shape

Figure 5.6: Boundaries of an example synthetic observation HII region for each of the four ages in the sample (those in the middle column of Fig. 5.5). Points signify the approximately equally spaced interpolation spline knots, where the curvature was calculated.

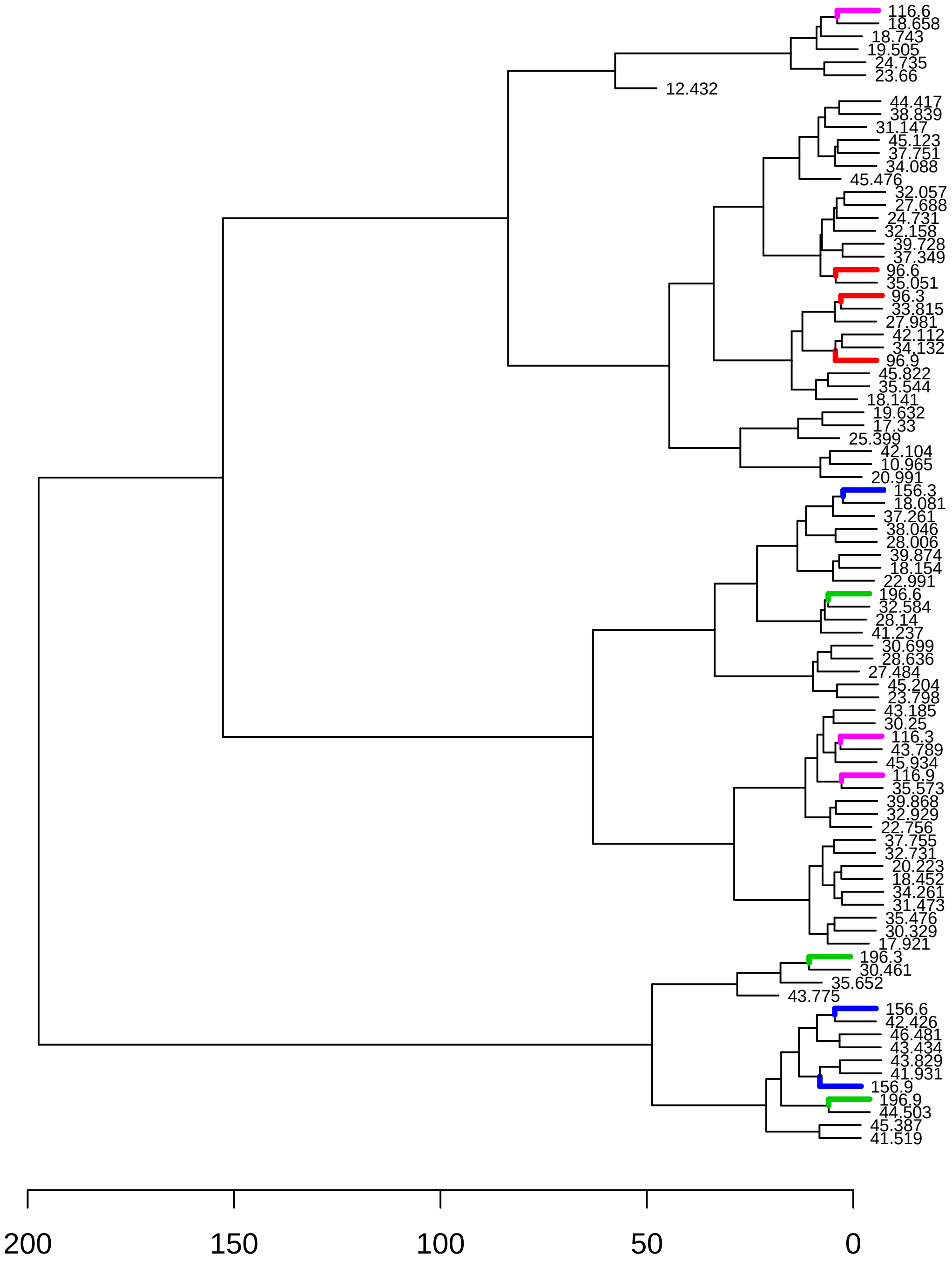


Figure 5.7: Dendrogram of the MAGPIS sample of Hii regions from Chap. 4 with the 12 example SOs with added Gaussian noise (shown in Fig. 5.5). The dendrogram represents the results from applying hierarchical clustering of the shape data of each Hii region. The branches of the 12 SOs are coloured by their age: 0.1 Myr in red, 0.2 Myr in pink, 0.4 Myr in blue and 0.6 Myr in green. The horizontal axis represents the height computed from the agglomerative clustering method.

Table 5.2: Summary of the different MAGPIS Noise Profile (NP) properties that the SOs were inserted into.

NP	l [°]	b [°]	μ [mJy/beam]	σ [mJy/beam]
1	20.99	0.09	0.105	0.263
2	12.43	-0.04	0.945	0.367
4	20.22	0.11	0.329	0.222
5	30.25	0.24	0.829	0.343
6	41.93	0.04	0.144	0.265
7	43.76	0.06	0.105	0.263

there is a MAGPIS HII region that shares the most similar shape, such that the projections are having a significant affect on the identified shape of the region. This will be further investigated and discussed later in this chapter.

5.2.3 MAGPIS Noise Profiles

Whilst the results of the previous subsection show that the numerical simulations are producing well representative SOs, the introduction of noise to the SOs in order to extract the shape can be developed further. The distribution of noise from radio interferometry images does follow a Gaussian distribution, however, it is not completely homogeneous across the image tiles. Artefacts from the reduction process and emission from fore- and background sources each contribute to the inhomogeneity of the noise distributions. In order to see how much of an affect the observational noise distributions from MAGPIS has on the shape of the SOs, the SO data was directly insterted into the MAGPIS tiles; into an area of the image tile deemed to contain only image noise and no signal. An example of this is shown in Fig. 5.8, where one of the 0.2 Myr projections (upper-right) is inserted to the MAGPIS tile containing HII region G030.252+00.053 (lower-left). The inhomogeneity of the image noise can be seen clearly by the gradient of noise distribution across the tile. G030.252+00.053 is at a distance of 4.5 kpc, has an effective spatial radius of 0.8 pc and a dynamical age of 0.1 Myr (Anderson et al., 2014; Campbell-White et al., 2018). The spatial pixel scale of the SOs is 0.06 pc per pixel. Since it was assumed that the HII region in the SO is at a distance of 6.2 kpc, which corresponds to the same MAGPIS angular pixel scale of 2″ per pixel; the smoothing factor applied to the contours was uniform for both the MAGPIS and the SO HII regions. As before, the contours shown are calculated from the clipped mean plus one standard deviation of the entire tile, which identifies both HII regions well.

Figure 5.9 shows an overview of an example SO projection for each of the four ages that have been inserted into six different MAGPIS tiles. The example SOs

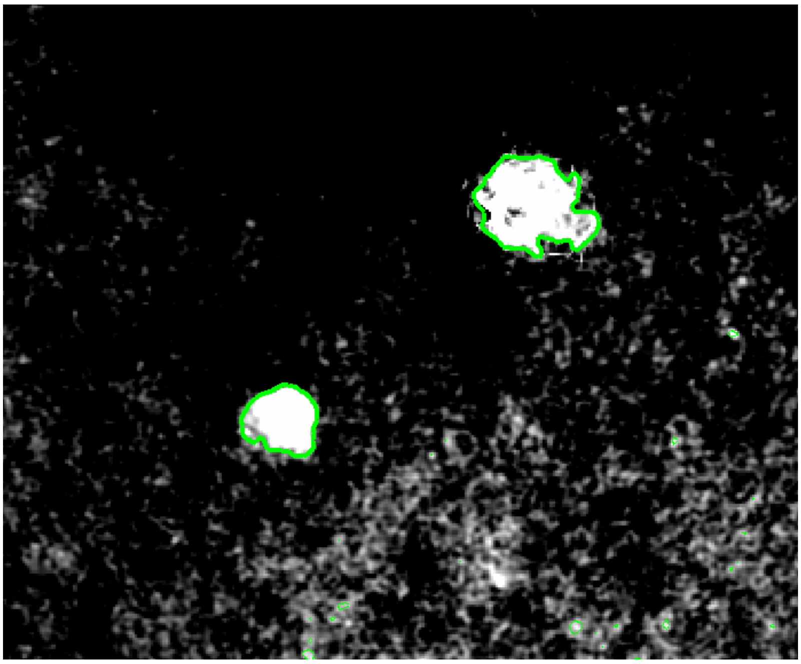
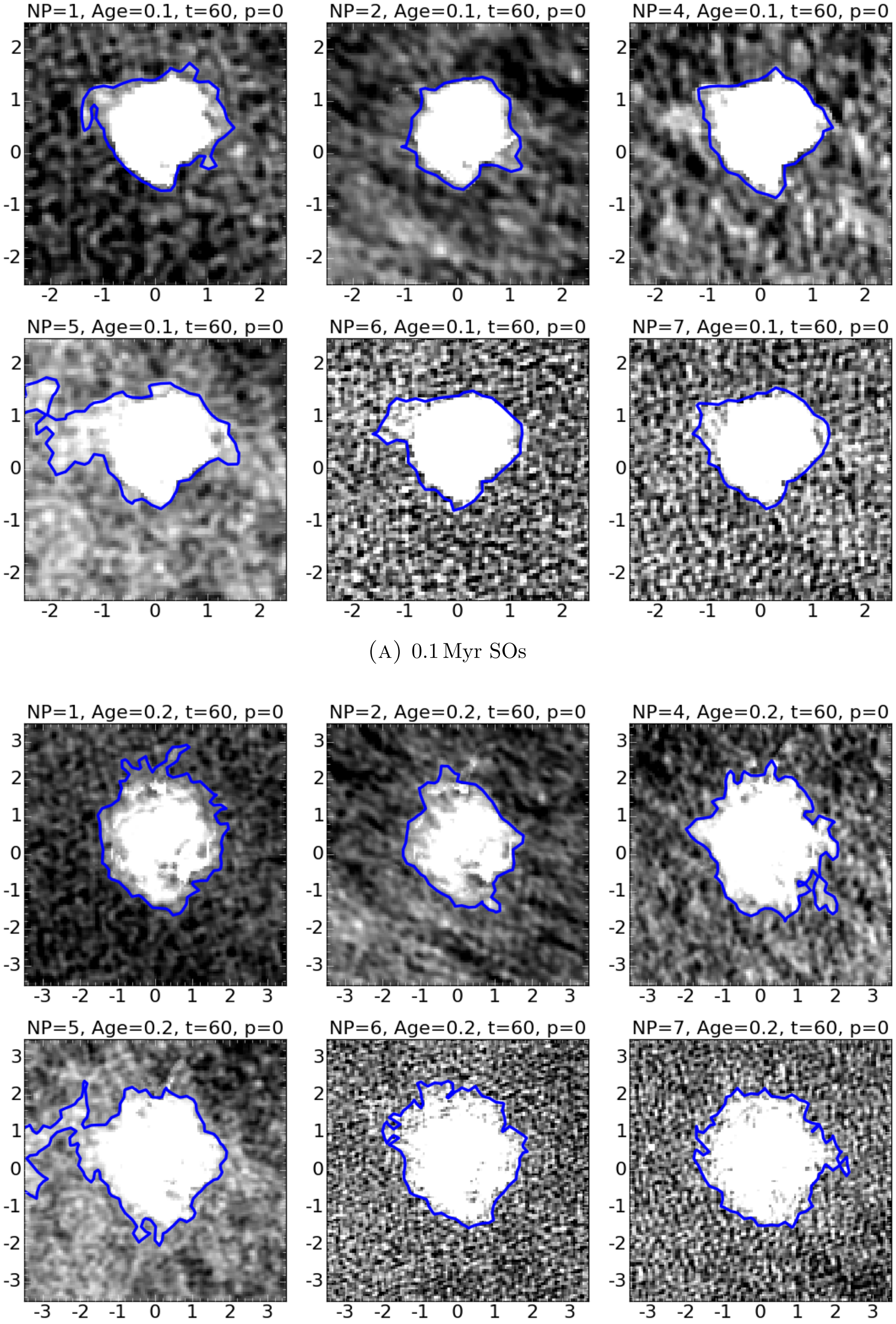


Figure 5.8: Example MAGPIS image tile and H ii region G030.252+00.053 (lower-left). One of the 0.2 Myr SOs has been inserted to the tile (top-right). Contours shown are that obtained from the image tile using the clipped mean plus one sigma.

used for each age are those in the middle column of Fig. 5.3 (the same ones used in Fig 5.6 to exemplify the interpolation spline fitting to the boundary contours). The MAGPIS tiles that the SO were inserted into will be referred to as the noise profile (NP). Details of which MAGPIS tile each NP represents is given in Tab. 5.2. The means and standard deviations were calculated from the sigma clipped MAGPIS tiles, with the contour level applied in each of the images then taken as the mean plus one standard deviation, as before. In the example images shown for NPs 1, 2, 4 & 5 at 0.4 and 0.6 Myr, part of the MAGPIS H ii region can be seen in the bottom right corner. Also, as with the Gaussian noise examples shown previously, the 0.4 and 0.6 Myr SOs have had their emission boosted by the same amount as before (factor of ∼4 and ∼70, respectively). This enabled the central region to be contoured appropriately; accounting for the loss of mass from the simulation grid.

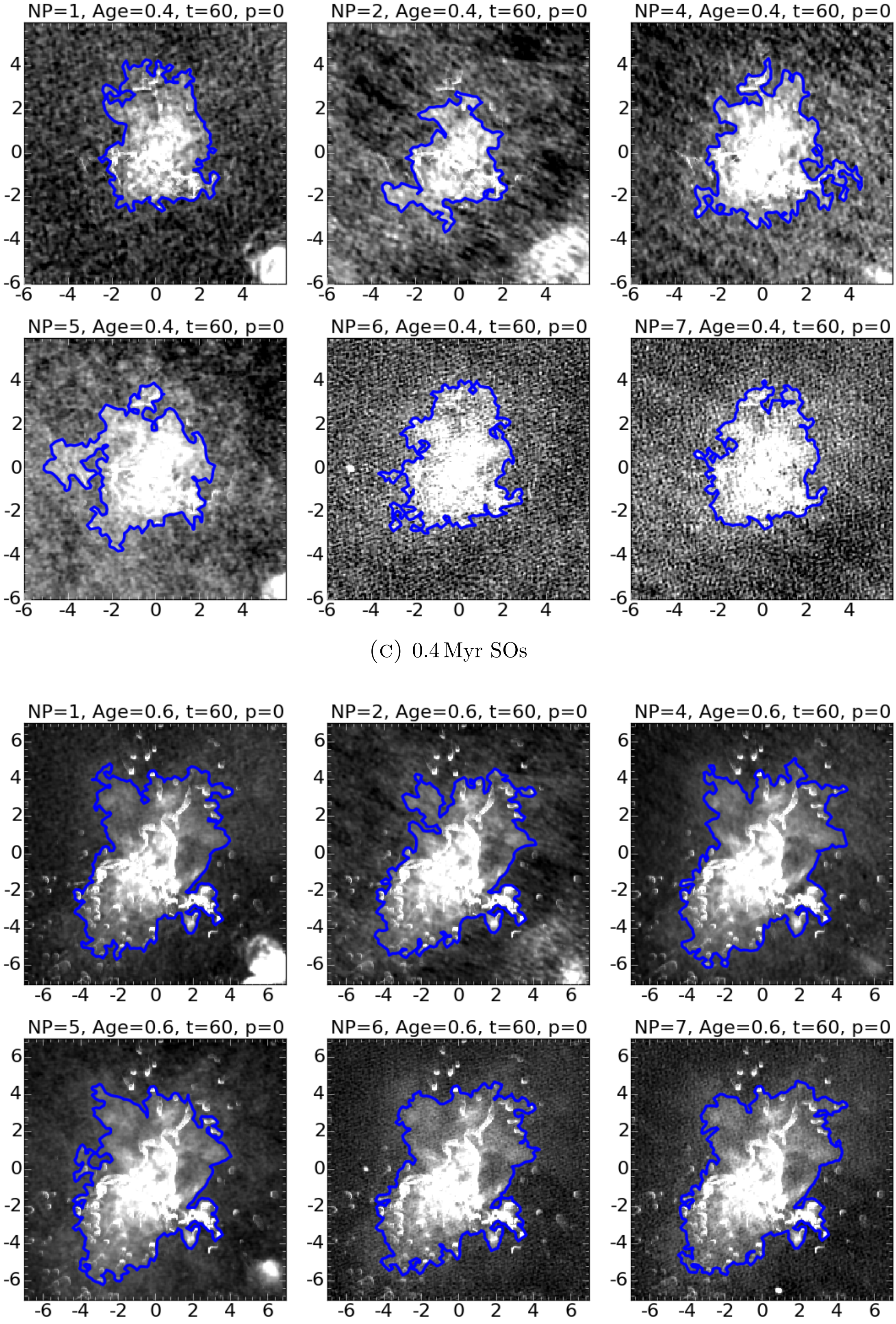
The change in the noise structure and intensity is apparent from Fig. 5.9, with NPs 6 and 7 appearing to have the most ‘salt and pepper’ like noise, similar to that seen in the random Gaussian distributions used previously. The examples shown for the 0.6 Myr SOs appear to have the least dissimilarities by group, although this may be because the regions themselves are larger and it is harder to visually discern the differences along the boundaries. It is worth remembering here that whilst these contours represent the systematically defined region boundaries, the shapes that are compared in the subsequent analyses are quantified from the shape landmarks,



(A) 0.1 Myr SOs

(B) 0.2 Myr SOs

Figure 5.9: SOs inserted into six different MAGPIS tiles to show how different noise profiles affects extracted shape. For each age, the same projection is shown, along with the 1σ plus clipped mean contours.



(c) 0.4 Myr SOs

(d) 0.6 Myr SOs

Figure 5.9: Continued from previous page

which are given by the interpolation spline fitting (see Fig. 5.6). Since we are still interested in the direct comparison of these SOs to the MAGPIS HII regions, the spline sampling remained at ~ 0.54 pc. Therefore, each of the boundaries shown here were under-sampled, with small perturbations along the contour smoothed out by the splines, with the level of smoothing proportional to the spatial size of the regions. Nevertheless, the differences in the contoured HII region shapes seen here, resulting from the different NPs, will enable investigation into how such changes in NP affect the final comparison of the shape. This will be discussed in detail in Sect. 5.3.

For the first three ages, NP 5 results in a spurious extrusion on the left side of the HII region. This is an example of where there may be underlying signal in what was thought to be only background noise, and thus it is having a clear affect on the identified shape. For the rest of the analysis carried out, we excluded NP 5 and considered the remaining five profiles. Figure 5.10 shows the resulting dendrogram from applying the clustering algorithm to the 76 MAGPIS HII regions, along with the 20 example SO HII regions (from Fig. 5.9), whose shapes were extracted from the five NPs detailed here. As with the previous dendrogram of the Gaussian noise profiles, the branches of the SO HII regions are coloured by their age: 0.1 Myr in red, 0.2 Myr in pink, 0.4 Myr in blue and 0.6 Myr in green. Figure 5.10 shows that all of the 0.4 Myr SO HII regions belong to the same parent branch, along with three of the five 0.6 Myr SO regions. The other two 0.6 Myr regions are paired together in a separate group. There are only a few examples of where SOs of the same age are paired together in this manner. As with the Gaussian noise profiles, most of the SO HII regions are being paired with one of the MAGPIS HII regions. This further confirms the efficacy of the simulations for producing representative SOs.

5.3 Discussion

Having shown that the insertion of the SOs into different NPs from the MAGPIS data leads to HII region shapes that are mathematically similar to what we observe in the MAGPIS sample; we can now further investigate the properties of the SOs and how they affect the obtained shapes. Whilst only SOs from one set of initial conditions are considered here, i.e. star cluster mass, ambient density, etc.; introduction of the MAGPIS NPs essentially expands the number of observations at each simulation time-step. For the 77 projections, across the four ages, with five NPs, we arrive at a sample of 385 individual HII regions. These variables are hence the

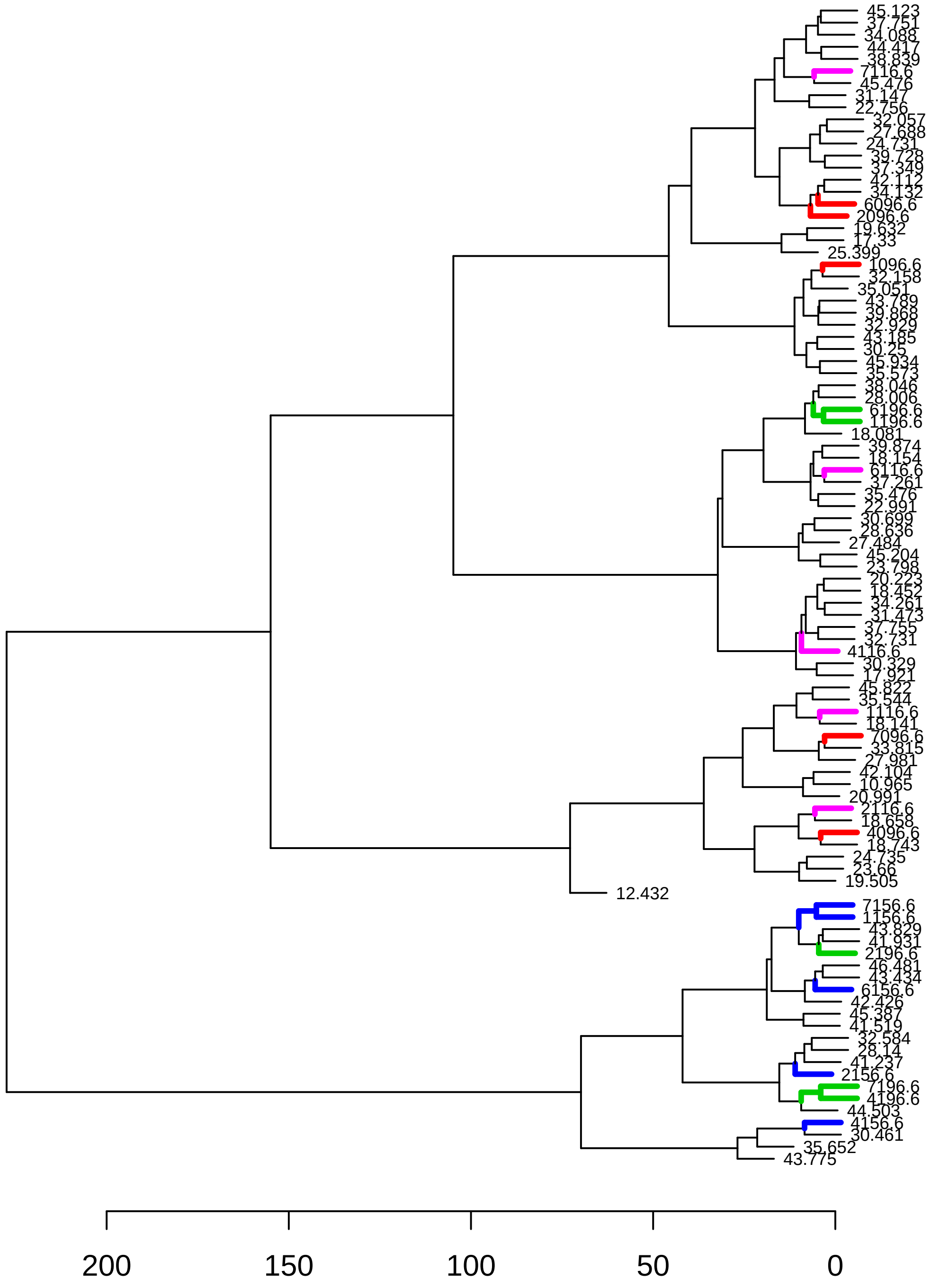


Figure 5.10: Dendrogram of the MAGPIS sample of Hii regions from Chap. 4 with the 20 example SOs that had been inserted into the MAGPIS NPs (shown in Fig. 5.9, excluding NP 5). The dendrogram represents the results from applying hierarchical clustering of the shape data of each Hii region. The branches of the 20 SOs are coloured by their age: 0.1 Myr in red, 0.2 Myr in pink, 0.4 Myr in blue and 0.6 Myr in green.

parameters that will be investigated further in this section. The ultimate aim of this investigation is to determine whether the SOs could be used as a training set for supervised classification of HII regions via their shapes, hence we need to understand how each of these parameters is affecting the obtained shapes and defining the groups.

5.3.1 Hierarchical Clustering of the SO Sample

Figure 5.11 shows the dendrogram resulting from applying the hierarchical clustering method to the shape distances of the 385 SO HII regions. As with the previous dendrograms presented in this chapter, the branches of the SOs are coloured by their age: 0.1 Myr in red, 0.2 Myr in pink, 0.4 Myr in blue and 0.6 Myr in green. It looks like from this dendrogram that there is a clear divide between groups containing the 0.1 and 0.2 Myr regions (early-type regions, to return to a familiar nomenclature in astronomy) and those containing the 0.4 and 0.6 Myr regions (late-type regions). This is displayed concisely in Fig. 5.12 by a bar chart of number of regions of a given age for the three distinct groups. These groups are obtained by cutting the dendrogram at a height of e.g. 300. In this situation, Group 1 contains mostly early-type regions with some late-type regions. Group 2 hosts exclusively early-type regions and Group 3 hosts mainly late-type regions with some 0.2 Myr regions being included. In terms of grouping the HII regions purely by age, the dendrogram in Fig. 5.11 shows that only a few of the smaller groups are all the same colour, that is, regions of the same age. A cut on the dendrogram, resulting in six groups is shown by the dashed red boxes. Each of these six groups appear to host either entirely late- or entirely early-type regions, with only a few exceptions of mixing. How often groups host SO regions of only one age can be investigated by taking a cut lower down the dendrogram, and considering more groups.

Figure 5.13 shows the ages and effective radii of the SO HII regions across 20 groups from the dendrogram. Here, the mean number of region per group is ∼20. It is clear that even with this many groups, it is still most common for there to be a mix of both early-type regions and both late-type regions belonging to each group. The few exceptions are group 8, with mostly 0.2 Myr regions and one 0.6 Myr region; Group 13 has only 0.4 Myr regions; and Groups 17 and 18 contain only 0.6 Myr regions. Increasing the number of groups further (to e.g. 40 groups), results in the same pattern of the grouping of early- and late-type regions and not differentiating the respective individual ages. This result shows that there is a lot of similarity between the shapes of the early-type and late-type SO HII regions, an observation we can also make from considering the interpolation splines which define the shape

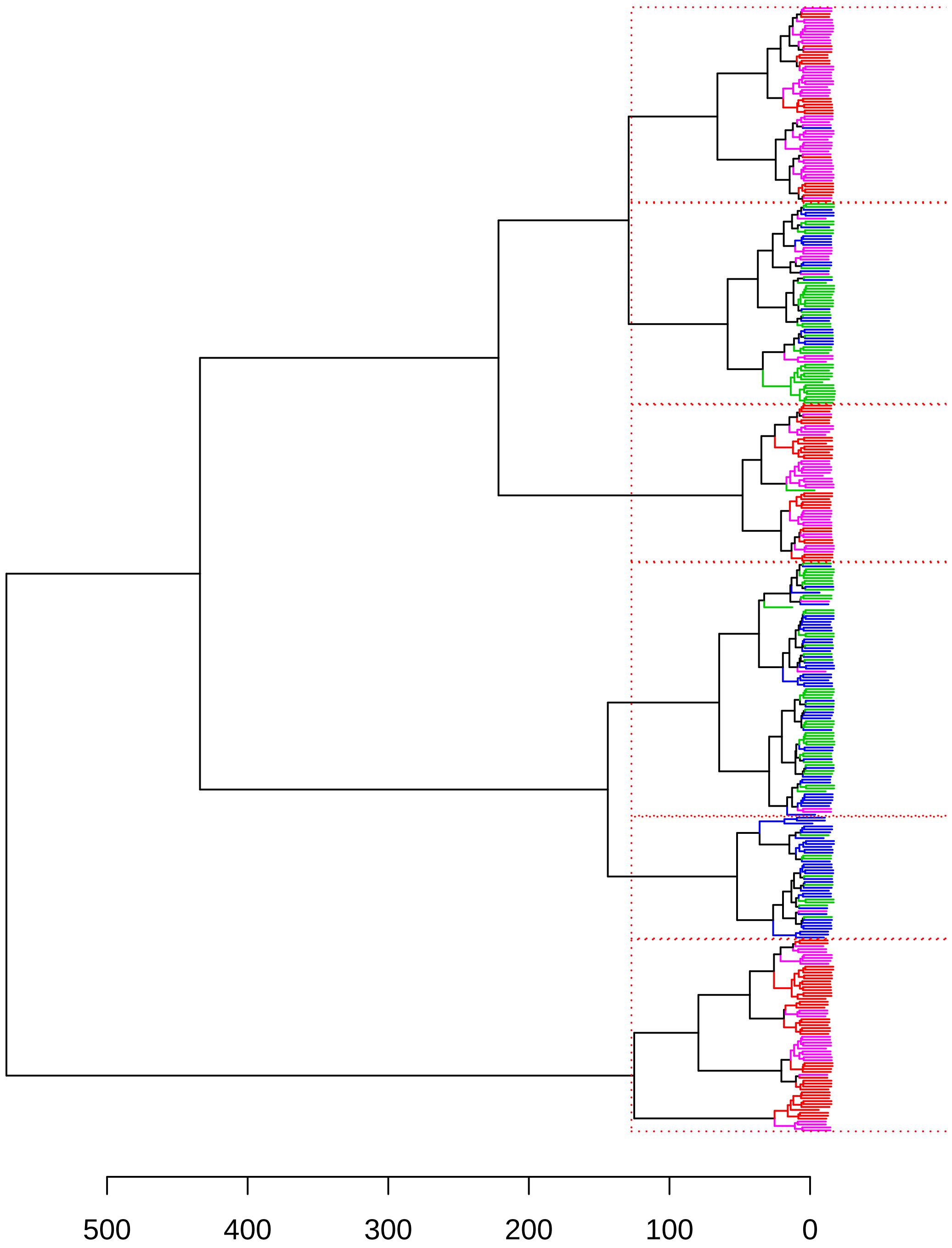


Figure 5.11: Dendrogram resulting from applying hierarchical clustering to the shape data of the sample of 385 SO HII regions inserted into MAGPIS NPs. As with the previous dendrograms in this chapter, the branches are coloured by their age: 0.1 Myr in red, 0.2 Myr in pink, 0.4 Myr in blue and 0.6 Myr in green. Six groups are delineated by the dashed red boxes.

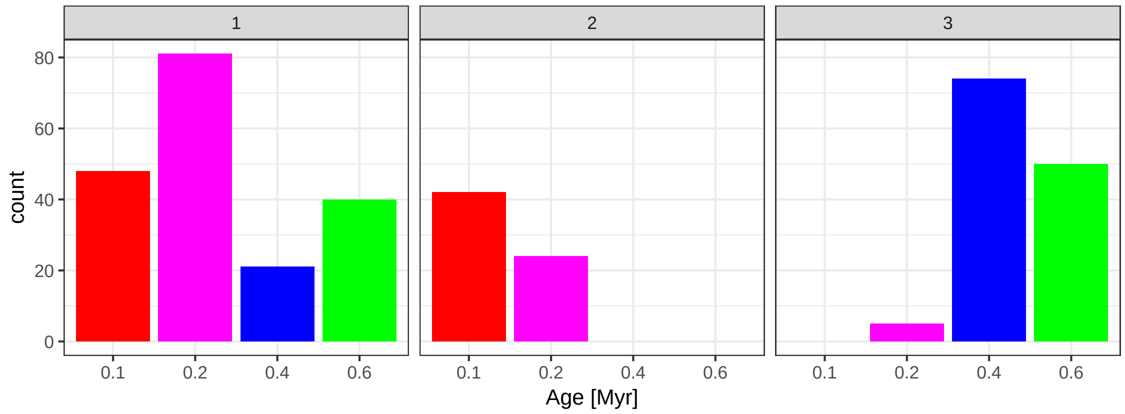


FIGURE 5.12: Bar chart showing the ages of the SOs that have been assigned to the first three groups in the dendrogram in Fig. 5.11, at a cut height of e.g. 300.

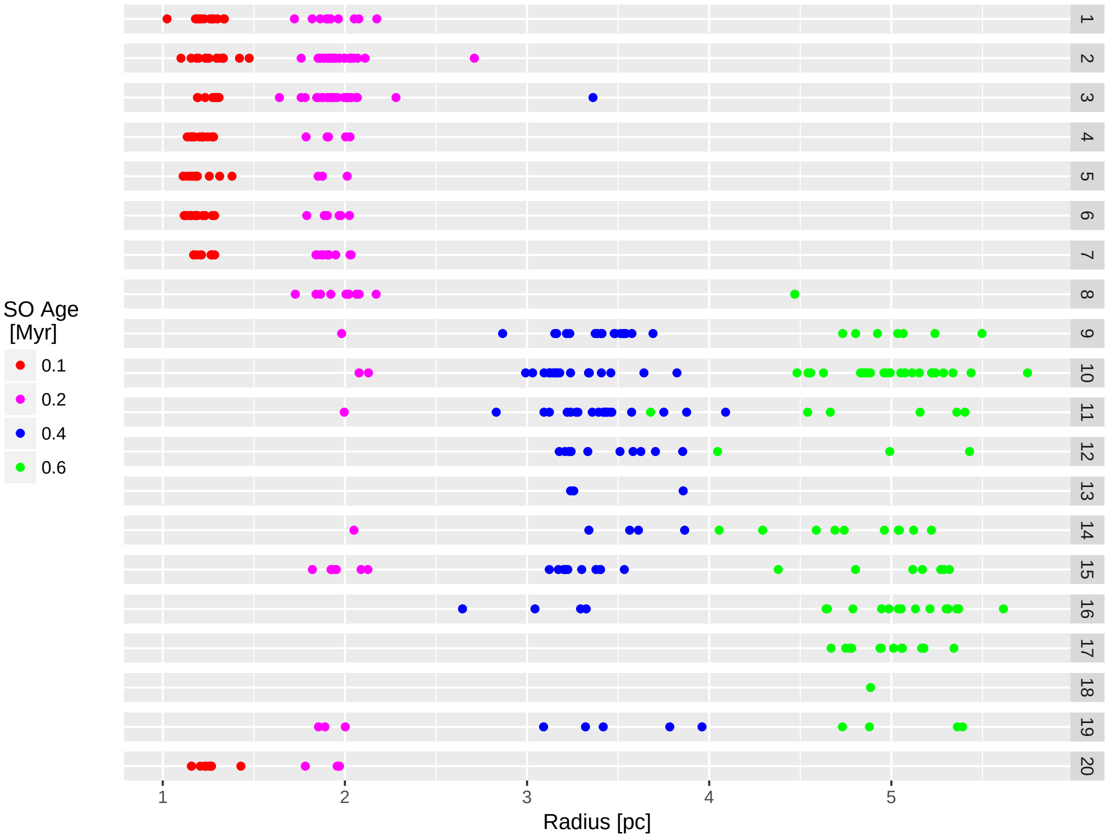


FIGURE 5.13: Distribution of SO effective radii and age for 20 groups from the hierarchical clustering of the shape data.

in Fig. 5.6.

A noteworthy point here is that the late-type regions are those which had their emission artificially boosted before being inserted to the MAGPIS tiles, to account for the loss of mass from the simulation grid. Whilst this could be the defining factor for the differentiation between these regions shapes, we do still see associations between some of the early-type regions and these late-type regions. Referring to the overview of HII region shapes from the purely Gaussian distributions in Appendix C, it can be seen that even with the more uniform noise, certain projections from the 0.2 Myr SOs feature boundaries with more perturbations. Is is likely these examples that are grouped with the late-type regions, based upon what we already know from how the grouping procedure works. Furthermore, from Fig. 5.13, there is both cross over and distinction in the obtained groups between the 0.4 and 0.6 Myr shapes. For the purpose of further investigations, it is maintained that the late-type regions are thus representative of their Galactic counterparts. This will be returned to in future work when considering simulations from a larger grid.

Figure 5.13 also shows that there is not a clear distinction in region radii by group, apart from that seen in the main split in early and late type regions (possessing small and large radii, respectively). In fact, some of the groups that host both early- and late-type regions display a large spread in region radii. The results from the MAGPIS data in Chap. 4 showed that one of the identified groups was exclusively young, small regions. The fact that this result is not seen here could reaffirm the notion that those regions from the previous chapter had similar shapes because of their young ages, and not purely their small sizes.

The remaining parameters to investigate, if and how they have an affect on the HII region shapes are the noise profile and projection. Figure 5.14 shows the distribution of NPs across the six groups identified by the boxes on the dendrogram. It appears that there is no clear preference for regions belonging to a given NP to be placed in a particular group. Group 2 has slightly more regions from NP 1 than the other NPs. Group 4 has fewer regions from NP 4 than any of the other NPs, whilst Group 5 shows the opposite result. Considering the ages along with the NPs, the regions from NP 6 that appear in Group 2 are only 0.2 Myr old, and all but one of the regions from NP 1 in Group 6 are 0.6 Myr old. The majority of NPs in each group, however, are associated with at least two of the ages, again, split by early- and late-type ages. Similar results to these are obtained by considering each of the θ and ϕ projection angles. There is no clear preference for a given observation angle to result in regions being assigned to the same group.

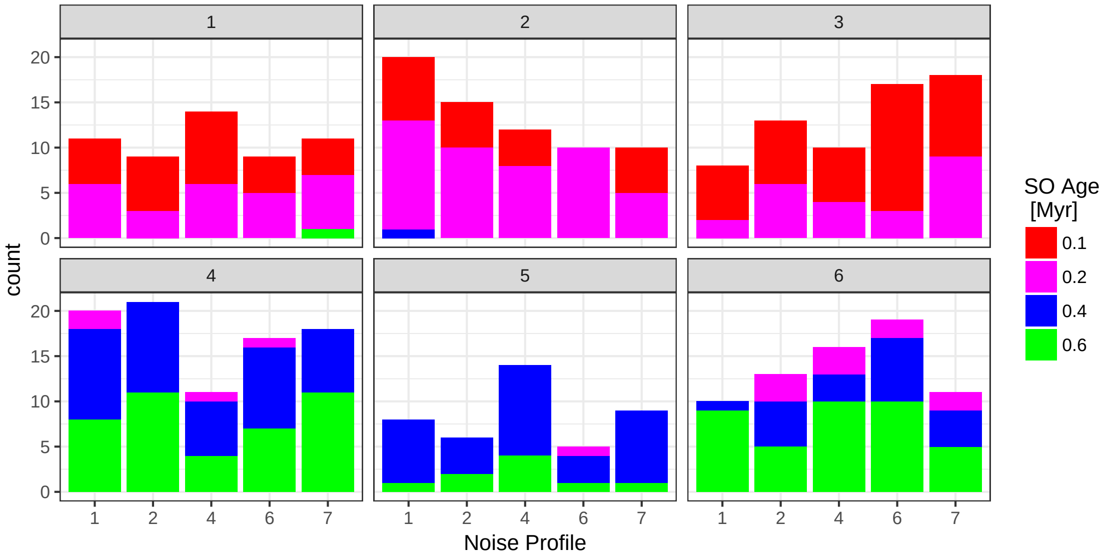


Figure 5.14: Bar chart showing the distribution of MAGPIS noise profiles hosting the 385 SOs, for the six groups delineated in Fig. 5.11.

Another way to discern whether the NP or the projection angle has more of an affect on the shapes and obtained groups, is to consider for a given NP, how many projections of each age are grouped together. Conversely, for a given projection angle, how many NPs for each age are grouped together. These results are shown in Fig. 5.15. For the six groups described previously, the top panel shows for a given projection angle and age, how many of the five NPs are placed in a given group. The bottom panel shows for a given NP and age, how many of the 18-22 different projections are grouped together.

We see from the top panel in Fig. 5.15 that it is most likely for only one or two of the NPs for a given age and projection to be grouped together, suggesting that the NP is having a large influence on the extracted shape of the HII region. For the examples where three or four of the NPs are grouped together, there is no preference for this to occur in a given group. There is only one situation where all five of the NPs are grouped together, that is for one of the 0.6 Myr projections in Group 4. The bottom panel shows some differences for the number of projections for a given NP and age that are grouped together. Group 1 has on average, five projections for a given NP grouped together. In Group 2, there is a preference for many of the 0.2 Myr projections to be grouped together. In Group 3, the average is again five projections, however, NP 6 for the 0.1 Myr regions includes 14 of the different projections. For the late-type regions, Group 4 shows the highest average number of projections per NP, followed by the 0.4 Myr regions in Group 5 and the 0.6 Myr regions in Group 6. The average number of given projections per group is

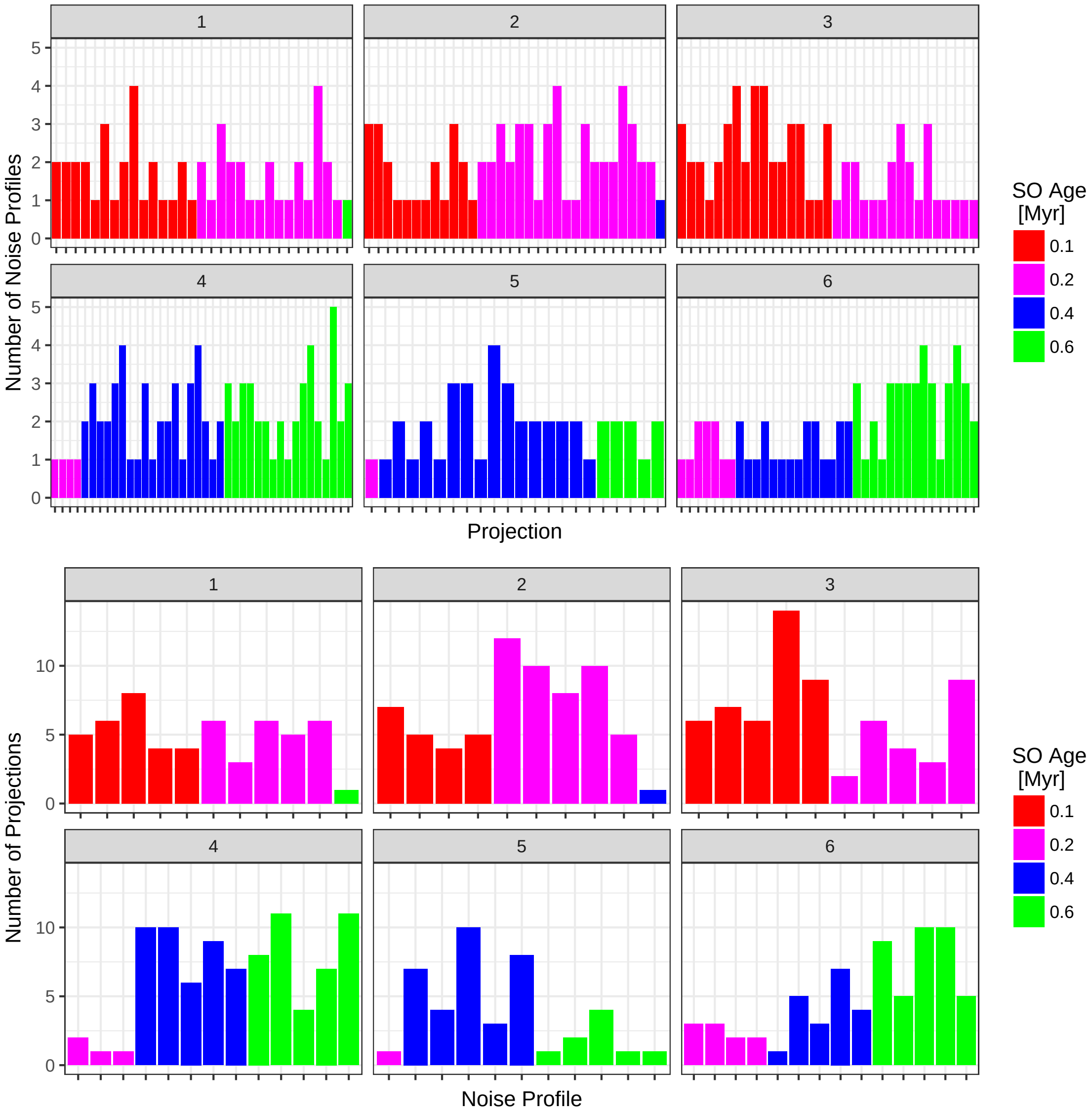


Figure 5.15: Bar charts showing the respective influence of changing the NP or the projection of the SO. Top: Distribution of different NPs for fixed age and projection. Bottom: Distribution of different projections for fixed age and NP.

only slightly higher for the early-type regions. This is an interesting result since the early-type regions are much more spherically symmetric than the late-type regions. This again shows that the curvature method for representing the region shapes is robustly quantifying the regions based on their boundaries.

The result that, in some groups, many of the different projections of a given age and NP are grouped together shows that changes in the viewing angle may not affect the shape of the HII region substantially. Whilst this may be due to the initial condition of spherical symmetry throughout the numerical simulations, this is still a result that can only be achieved via study of the SOs. As much as we would like to place a telescope at the other side of the Galaxy, it looks like we are still at least a few years away from that kind of technology (tens of thousands of years in fact, at the current orbital speeds of satellites). Since we do only have the one vantage point from Earth, the requirement for observational classification remains to be able to conduct this in conjunction with the information we can collect. Therefore, having the different projection angles of the SOs essentially provided different observations of HII regions that share the same initial conditions, but can be thought of as evolving differently due to differences in ambient density or ISM structure along our line of sight. However, investigating further how the noise structure of our observations affects the shape we observe is an important aspect towards refining an observational morphological classifier. The next subsection will look at this in more detail.

5.3.2 Investigating How Noise Affects Shape

The manner by which the HII region shapes were extracted from both the MAGPIS tiles and the SOs was by analysing the background noise from the radio continuum images. By removing the radio signal and setting a threshold level that was above the remaining noise profile, this enabled the boundary of each region to be defined by the contouring procedure. This led to a systematically defined data sample, whereby the HII regions were each extracted, such that their signal levels should be consistent across the Galactic Plane. Nevertheless, one could argue that if you took one of the HII regions and placed it in a different area of the Galaxy, defining the boundary from the background radio noise in the vicinity could lead to the shape being different. This was thus what the SOs allowed for, by doing exactly that. We have seen in the previous section that these different NPs, defined from the MAGPIS tiles, are having an affect on the shape. The question here is can we use the statistical tools employed in this work to quantify the affect each of the different noise profiles are having on the HII region shapes?

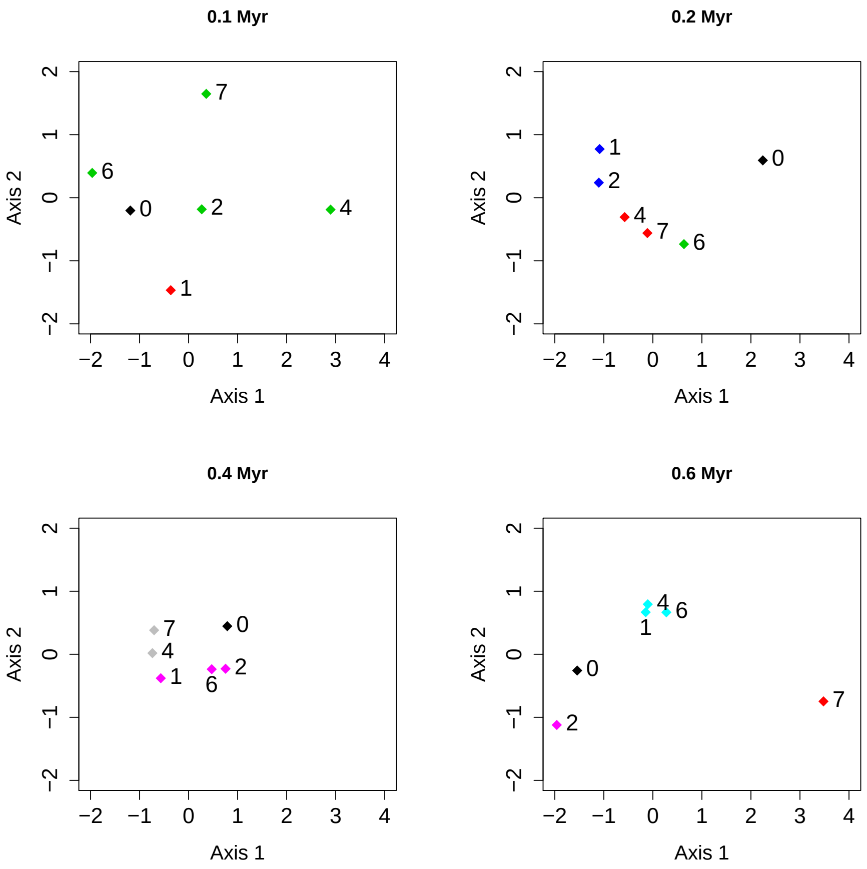


Figure 5.16: Multidimensional Scaling ordination plots for the shape distances of four example SO ages and projections. Labels signify NP, with 0 representing the shape obtained from the random Gaussian distribution. Point colours correspond to which group the SO HII region shapes were allocated to in the hierarchical clustering of the full dataset.

In an attempt to answer this question, let us return to the ordinance technique of multi-dimensional scaling (MDS), which was used in Chap. 4 to check that the hierarchical clustering was properly defining groups based upon the regions shapes. To recap: MDS reduces the dimensionality of an input distance matrix to a number of orthogonal principal coordinates. The eigenvectors of which, give the ordination and the eigenvalues give the relative importance of that axis for representing the data variation. In Chap. 4, it was found that there was a correspondence between the amount of high curvature points along the region boundaries and the scores along axis 1 of the MDS ordination, and surmised that the variation along axis 2 was also directly resulting from features of the curvature distributions. Here, MDS is applied to the distance matrix of HII region shapes, for a given SO across the different NPs, to both visually and quantitatively see how the resulting mathematical shapes compare.

Figure 5.16 shows the MDS ordinations for four of the SOs, one at each age. The numbering of the HII region shape's points on the graphs corresponds to the NP of the SO (from Sect. 5.2.3), with NP = 0 corresponding to the random Gaussian distribution (from Sect. 5.2.1). In each of the MDS plots, only the six HII region shapes shown were compared pair-wise using the A-D test. These results are hence showing only the differences the NPs have on the shapes. In each of the four instances, for increasing SO age, axis 1 of the MDS accounts for 64, 76, 73 and 80% of the shape variability, respectively. Axis 2 accounts for 23, 18, 16 and 15%, respectively. Therefore, these two axes are sufficient for investigating the shape differences accordingly.

In each of the four plots, the Gaussian noise profile 0 shape is ordinated away from the other five NPs. For 0.1 Myr, all of the shapes are spread over the ordination plot. For 0.2 Myr, NP 0 is ordinated away from the origin, at approximately an equal distance from each of the other NPs. A similar looking distribution is seen for the 0.4 Myr data, with a tighter association. For the 0.6 Myr data, NPs 1, 4 and 6 are ordinated very close to one another, with NPs 0, 2 and 7 ordinated away. In each case, the points are coloured by which of the six groups the shape was sorted into from the dendrogram in the previous section. As expected, those ordinated close together are assigned to the same group in the larger data set. NP 7 in the 0.6 Myr data is the one example of that age assigned to a group comprising otherwise only early-type regions. As we can see here, it has the furthest distance from the other points. Another interesting note is that, for the 0.1 Myr shapes, the shape from NP 2 is ordinated close to the origin of the coordinate system. This means that this shape represents the 'average' of the sample and the other shapes are all differing with respect to this shape.

Unfortunately, the take away message from these MDS ordinations is that the affect the different NPs have on the underlying shape of the SO HII region is not a systematic affect across the ages. There is a slight ordination similarity between the distribution of the MAGPIS NPs with respect to the Gaussian shape in the 0.2 and 0.4 Myr plots. However, each of the respective NPs by age are not behaving in the same manner in each of the MDS plots; neither with respect to the other MAGPIS NPs, nor the Gaussian noise only shapes. The example shown for the 0.1 Myr SO shows a large spread in the MDS ordination, yet this is an example where four of the five NPs are put in the same of the six groups from the hierarchical clustering of the entire dataset. On the other hand, NP 1 of the 0.4 Myr data is ordinated close to NPs 4 and 7 but is placed into the same group as NPs 2 and 6. It should be remembered here, however, that the hierarchical groupings denoting the colours are for a much larger sample, such that many of the other projections will be influencing the groupings due to the agglomerative procedure.

Another use of the MDS ordination technique is to further investigate how the different selection choices for defining the HII region shapes affect the resulting spread in the shape data. That is, the initial sigma level used when extracting the boundaries and the spline knot spacings for controlling the spatial resolution of the shape landmarks. So far, for all of the SO shape data, only the 1 sigma contour level was considered, along with the 0.54 pc spline knot spacing. This was for the purpose of directly comparing the SOs to the previous results from the MAGPIS data. However, it was suggested there that the SOs could provide a better test set for determining how these two selection variables affect the resulting shape.

When rerunning the MDS of the shape data using contours with 0.8 and 1.2 sigma above the mean value, the relative positions of the ordinations changes, but the overall spread in the data remains. This suggests that, as with the NPs, varying the initial sigma level is not having a systematic affect on the shape data. When rerunning the MDS with different spline resolutions, decreasing the spline interval (hence increasing the spatial resolution) results in a larger spread along the MDS axes. This was expected as more features and points for comparison are captured with a higher resolution. Furthermore, as found in the previous chapter, the amount of high curvature points corresponds to the score along axis 1 of the MDS. Increasing the spline interval (decreasing the resolution) results in a smaller spread in the MDS ordination. Whilst this may seem like a favourable result, what is actually happening here is the level of smoothing along the boundaries is increased, resulting in fewer features along the curves. Decreasing the sampling resolution too much also becomes redundant for the smaller diameter regions. As found with the previous tests of this nature, carried out in Chap. 4, the 1 sigma, 0.54 pc interval seems to be a good intermediary between extremes. The most important aspect here is that the shape extraction and quantisation remains consistent for the sample.

5.3.3 Using the SO Sample as a Training Set

Although a systematic affect that the different noise profiles are having on the SO HII region shapes was not found, the SOs still possess much better constraints than the Galactic observations. It was also shown here, via the shape analysis method, that they are explicitly representative of the Galactic HII regions. Considering the sample of 385 SOs, it is known with certainty, which evolutionary stage each SO is at. The different projections and noise profiles essentially provide many more observations of each given age. This is similar to what is observed along the Galactic plane, with no preference for HII regions of a given age to be at a given place along the Plane (Anderson et al., 2014). Ultimately, for a training set, SOs of differing initial conditions are required. Using the data that was available here, however, it was investigated whether the SO data sample can be used to infer the ages of the MAGPIS observational sample.

The initial conditions of the numerical simulations, which produce the SOs analysed here, involve ionisation and feedback from a star of mass 33.7 $M_{\odot}$, with an ionising photon rate of $N_{ly} = 7.36 \times 10^{48}$ s^{-1}, or log N_{ly} = 48.86 (Ali et al., 2018). Since the MAGPIS observational sample analysed in the previous chapter cover a range of masses ($\sim$ 17 - 45 $M_{\odot}$), a mass limited sample from these HII regions was considered for use in the following test. The limit was $48.3 < \log N_{ly} < 48.8$. This corresponds to a mass range between 23 and 34 $M_{\odot}$ (Weidner and Vink, 2010), and was around the mean of the normally distributed values for the MAGPIS sample (Fig. 4.12). The lower limit was taken to account for the fact that as the regions evolve in the numerical simulations, mass leaves the grid and the resulting SOs may in fact be representative of regions of lower masses. This resulted in a subsample of 26 MAGPIS regions, with ages ranging between 0.1 Myr and 1.9 Myr. Whilst this age range of the MAGPIS subsample covers a considerably larger range than the SOs used in this study, there is not a one to one correspondence between the two ages used. The ages from the SO snapshots begin with $t = 0$ when feedback starts in the simulation. Whereas, the ages calculated for the MAGPIS sample are only an estimation of the dynamical ages. These estimates involve assumptions regarding the surrounding ISM and require accurate distances to the regions. The dynamical age then considers the observed expansion of the ionisation front with respect to the theoretical Strömgren radius. Therefore, for the purpose of this test, the respective age ranges and distributions of both the SOs and MAGPIS sample are considered, to see how they compare.

Using the data set of 385 SOs described in Sect. 5.3.1, the 26 mass limited MAGPIS HII region shapes were included with the SO sample. Where the MAGPIS HII regions would be placed in relation to the SOs in the resulting group structure was then investigated. Figure 5.17 shows the dendrogram resulting from the hierarchical clustering procedure for the SO training data and the MAGPIS target data. Branches are coloured as before - 0.1 Myr in red, 0.2 Myr in pink, 0.4 Myr in blue and 0.6 Myr in green - with the addition of the MAGPIS HII regions in cyan. The introduction of the MAGPIS regions has changed the ordering of the six delineated groups, with mostly late-type regions shown in the top two groups and

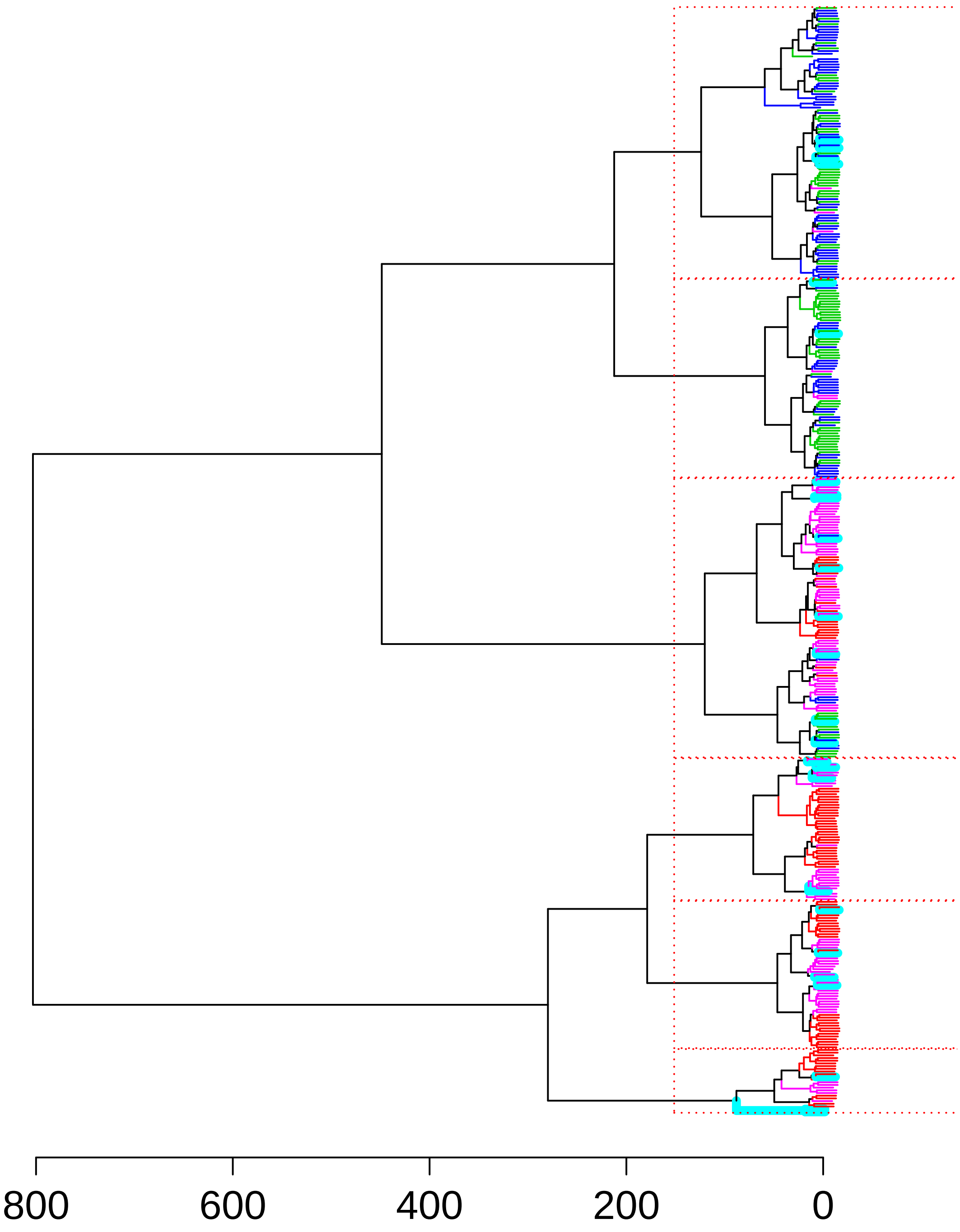


Figure 5.17: Dendrogram resulting from applying hierarchical clustering to the shape data of the sample of 385 SO Hii regions along with 26 of the MAGPIS Hii regions. The branches are coloured by their age for the SO data: 0.1 Myr in red, 0.2 Myr in pink, 0.4 Myr in blue, 0.6 Myr in green; and the MAGPIS Hii regions are in cyan.

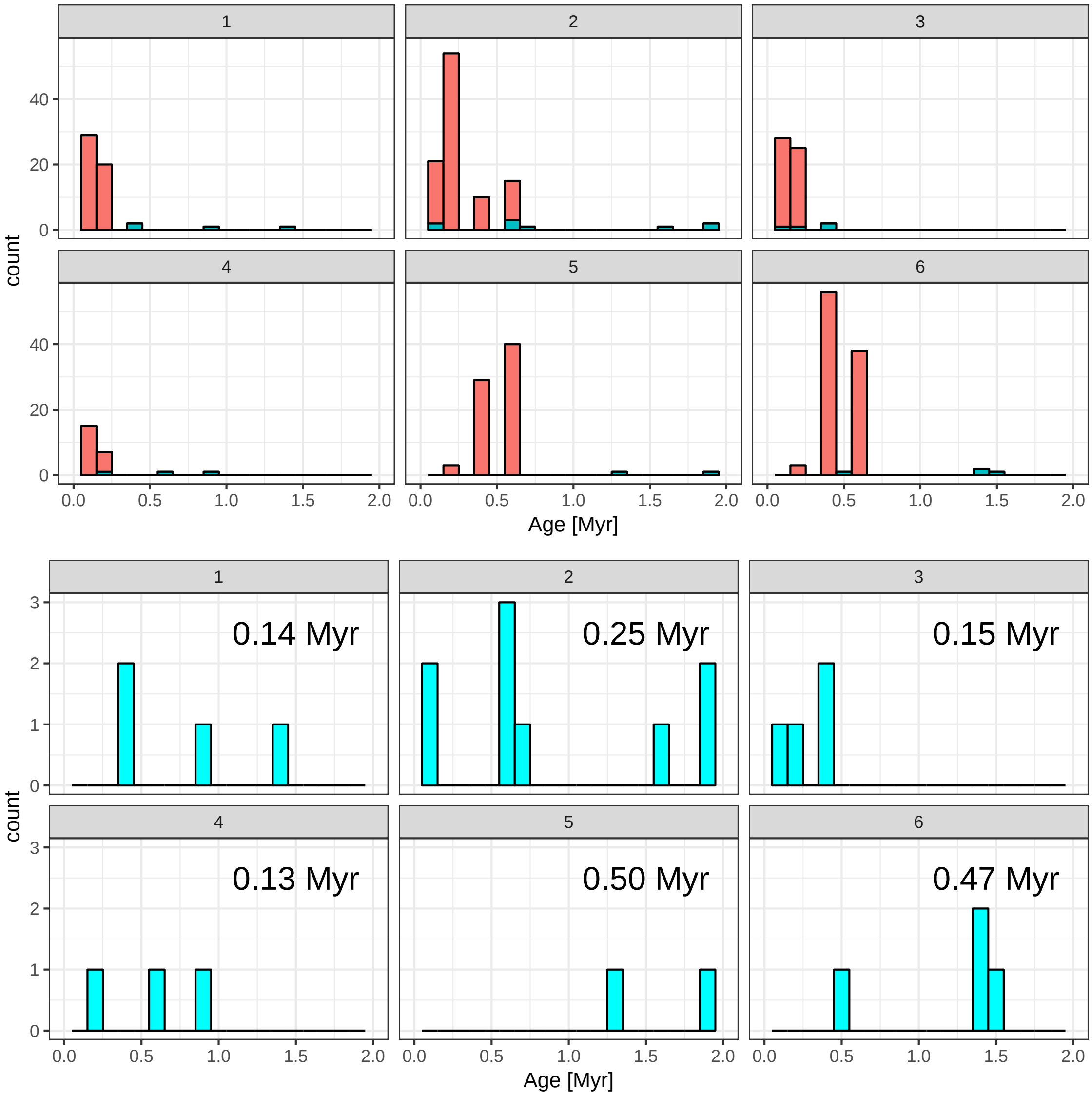


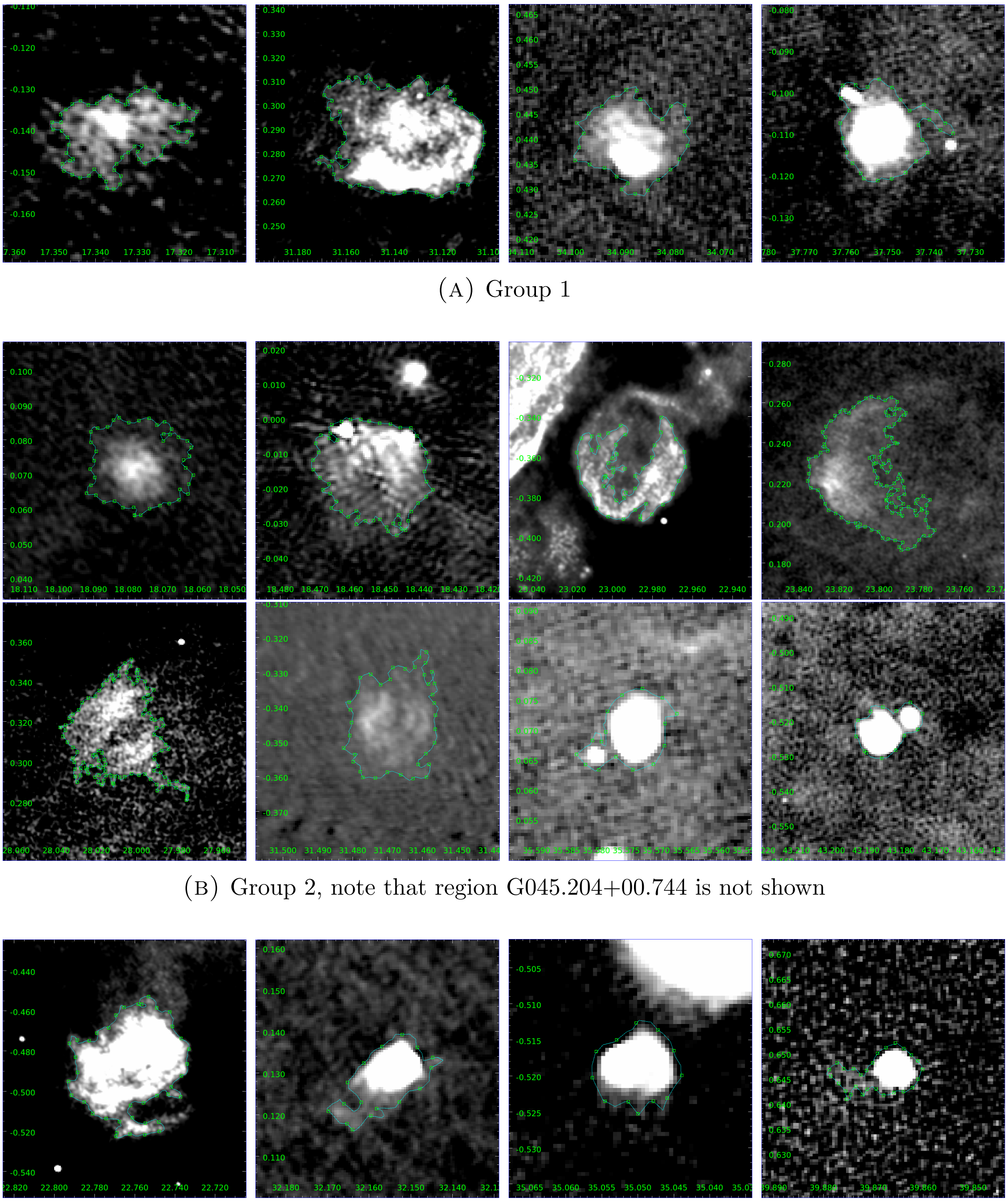
Figure 5.18: HII Region ages for six groups identified from Fig. 5.17. Top: Ages of the SO HII regions (in orange) and MAGPIS HII regions (in cyan). Bottom: Ages of the MAGPIS HII regions, with the average age of the SO HII regions shown in each group.

mostly early-type regions in the rest. As explained in Sec 3.3.2, the ordering of the final groups is arbitrary. It is the resulting group associations and hierarchy that matter. The MAGPIS Hii regions appear slotted in to the SO Hii regions. The same result of good correspondence that was seen with the data the other way around in Sect. 5.2.2. There are two MAGPIS regions, paired together, joined to the bottom group, These two may represent slight outliers since they join to the adjacent group at a substantial height. These will be discusses in more detail later in the section.

Figure 5.18 shows the respective ages of the SO and MAGPIS Hii regions that are grouped into the six groups in Fig. 5.17. The upper panels include both the SO regions, in orange, and the MAGPIS regions, in cyan. The bottom panels show only the MAGPIS regions, since they are hard to gauge in the upper panels. The upper panels give the overview of the SO data ages, a result which has been discussed previously in the section. There is a clear split between early- and late-type regions, with Group 2 showing the most cross over between ages. The ages indicated in each panel of the lower plots are the mean ages of the SO data in each respective group. The MAGPIS Hii regions within each group are then shown clearly, along with this stated mean SO age. Following from the respective age discrepancies, mentioned previously, the MAGPIS Hii regions can be relatively distinguished as early- and late-types by considering those with age less than 1 Myr to be early-type and those greater than 1 Myr to be late type.

Group 1 hosts exclusively early-type SOs with a mean age of 0.14 Myr. 75% of the MAGPIS regions assigned to this group are also early-type. Group 2 has the largest mix of early- and -late type SOs, but with majority 0.2 Myr SOs and a mean age of 0.25 Myr. Two thirds of the MAGPIS regions in Group 2 are also early-type, with the remaining late type, showing a good agreement with the SO data. Group 3 has exclusively early-type regions for both the SO and MAGPIS data. The same result is seen for Group 4. Group 5 hosts majority late-type SOs, with a mean age of 0.5 Myr. The two MAGPIS regions assigned to this group are also late-type. 75% of the MAGPIS regions in Group 6 are late-type, in good agreement with the mean age from the SOs. Each of these results are promising for the prospect of using the SOs as a training set for supervised classification. We see here that even with only one parameter of investigation, the evolutionary stage of the regions, we have good agreement between the SOs and MAGPIS observed sample.

In addition to these results Fig. 5.19 shows an overview of the MAGPIS Hii region images and shapes that were assigned to each SO group. We can see from Fig. 5.19 that the MAGPIS regions sorted into each of the test groups, appear to share similar morphological features. This reaffirms the notion that the shape anal-

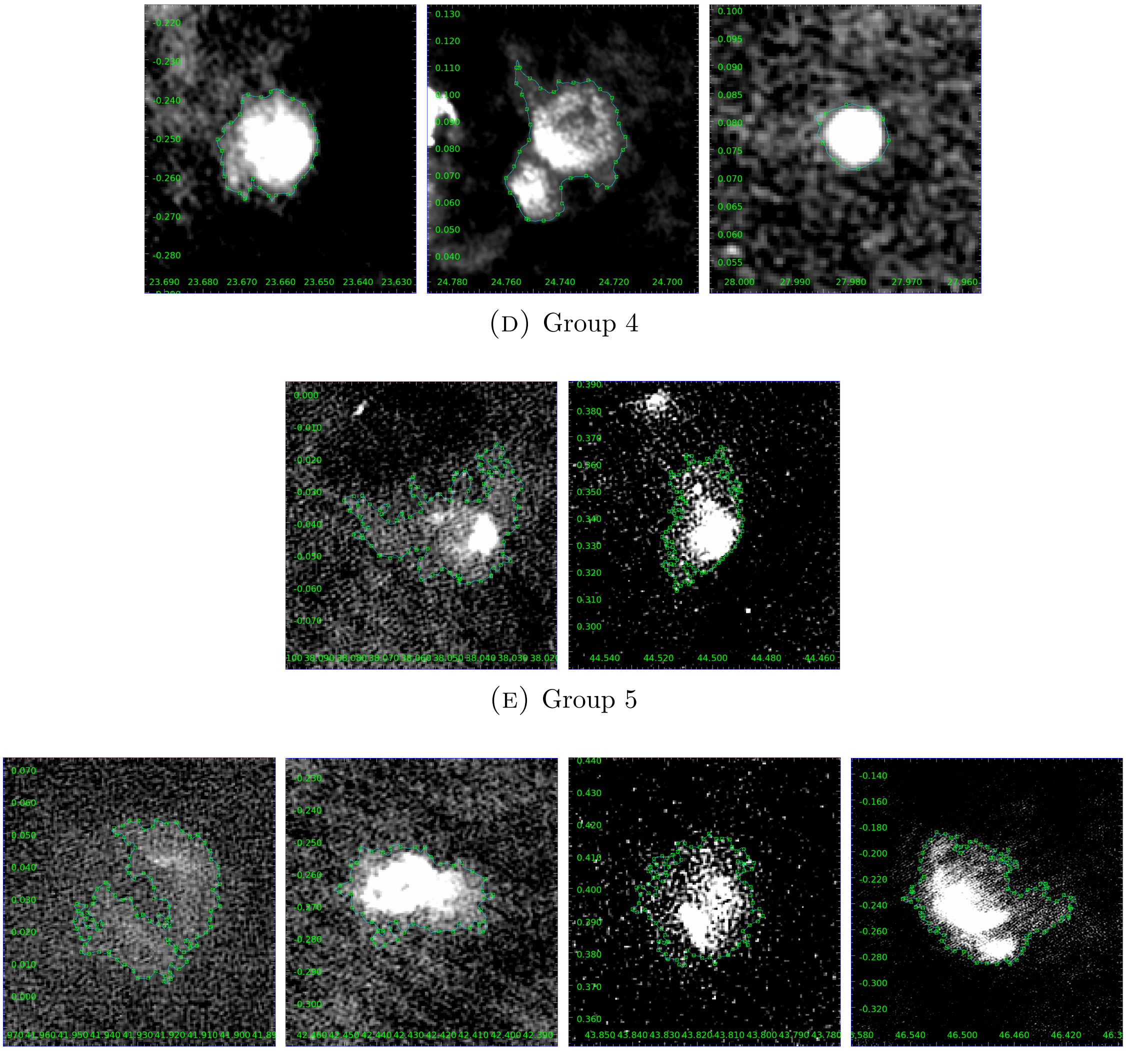


(a) Group 1

(b) Group 2, note that region G045.204+00.744 is not shown

(c) Group 3

Figure 5.19: Summary of the MAGPIS Hii regions that were assigned to each of the training groups in Fig. 5.17.



(D) Group 4

(E) Group 5

(F) Group 6

Figure 5.19: Continued from previous page

ysis and statistical methods employed here are performing as expected. Group 2 is the largest group, showing the most visual differences between MAGPIS regions. This would be the first group to split if the cut was made lower on the dendrogram in Fig. 5.17, which would separate the more uniform regions from the more perturbed. The first two MAGPIS regions shown in the images for Group 4 are those located at the edge of the dendrogram. They appear to each host at least one tight point of inflection, which would result in a large outlier in the curvature distributions. This is the likely cause for why they are slightly apart from the rest of the data. The comparatively smooth sections along the rest of these region's boundaries are likely why they were grouped to the other MAGPIS region, and the corresponding SO regions in Group 4.

The idea of using the SOs as a training set in supervised morphological classification of HII regions would require an input sample with many varied initial conditions and known parameters. It has been shown throughout this chapter that the SOs produced by Ali et al. (2018) are well representative of their Galactic counterparts. Furthermore, these final results show that SOs of different ages could be used to predict whether a Galactic HII region is early- or late-stage. With different initial masses and ambient densities, the parameters of each training set group could be further refined, and this investigation could be repeated with a correspondingly larger sample of Galactic HII regions. This, and other potential applications of this work, are elaborated upon in Chap. 6.

5.4 Conclusions

The synthetic observations of an HII region produced in the numerical simulations of Ali et al. (2018) were analysed using the shape analysis and statistical clustering methodology developed in this work. The numerical HII region was the result of photoionisation and feedback of a 34 $M_\odot$ star, in a 1000 $M_\odot$ cloud. 77 SOs were provided, comprising four evolutionary snapshots (0.1, 0.2, 0.4 and 0.6 Myr), and multiple viewing projection angles. After the addition of random Gaussian noise, following the distribution of observational noise from one of the MAGPIS tiles, the shapes of the SO HII regions were extracted in the same manner as they were for the MAGPIS sample. The shape analysis results provided resounding confirmation of the efficacy of the numerical simulations. When considering the 76 MAGPIS HII regions from Chap. 4 and 12 example SO HII regions, across the four ages, the SO HII regions were placed in amongst the MAGPIS HII regions, in the resulting dendrogram from the hierarchical clustering procedure.

This result was also found when directly inserting the SO regions to different MAGPIS tiles. By using five MAGPIS noise profiles for the 77 SOs, there were essentially 385 distinct observations of the numerical Hii region, at the given ages and projections. As with the shapes of the Hii regions using the random Gaussian noise distribution, those from the MAGPIS NPs were grouped in amongst the MAGPIS Hii regions, with the majority of synthetic regions paired with one of the MAGPIS regions. This suggested that the different projection angles and noise profiles were having a significant impact on the regions shapes. Such that the SO Hii regions of the same age were not exclusively grouped with each other in the hierarchical clustering.

When considering the hierarchical clustering of the 385 SOs that had been inserted into the MAGPIS tiles, the following results were obtained:

- The resulting hierarchy showed a clear divide between early- (0.1 and 0.2 Myr) and late-type (0.4 and 0.6 Myr) regions. This divide was not exclusive by age, with a low cut on the dendrogram (resulting in many groups) still producing groups that mostly possessed a mix of both early- and late-type regions. Whilst this may be due to how the late-type region's emission had to be artificially boosted to account for mass leaving the simulation grid, the results for the late-type regions still show the same cross over as the early-type regions. Furthermore, these late-type regions were still shown to be representative of the MAGPIS observational sample, even with this boosting. However, this is a point to return to in future work with simulations from a larger grid.

- There was no further association between the identified groupings and SO region radii, apart from the main split between the early- (small radius) and late-type (large radius) regions. This suggests that the result obtained in Chap. 4, pertaining to one group hosting exclusively small regions could in fact be due to those regions all being young Hii regions.

- There was no strong preference for SO regions from a given noise profile, nor given projection angle, to be assigned to specific groups. In terms of which of these parameters has more of an affect on the shape - for a given SO age and projection angle, not many of the five NPs were grouped together in the hierarchical clustering, on average only two NPs would be grouped together. For a given age and NP, however, there were more instances where numerous different projections were grouped together. This result was consistent for both the early- and late-type SO Hii regions, even though the early-type regions appeared to be more spherically symmetric.

The multidimensional scaling ordinance technique was used in order to try to find a systematic affect the different noise profiles had on the SO Hii region shapes. For an example projection from each of the four ages, the shapes from the different NPs, along with the one extracted from the purely Gaussian noise profile, were compared using the shape analysis method. In each case, the Gaussian NP was ordinated away from the five MAGPIS NPs. However, the ordination for each of the MAGPIS NPs with respect to each other and the Gaussian shape was not systematic across the ages. The MDS did show systematic affects for how the shape of the Hii regions is extracted. Different initial contour levels changed the ordination in the MDS axes, but not the relative scores along the axes. Whilst different shape resolutions changed the scores along the axes but not the relative ordination positions. Higher resolutions corresponded to a larger spread in MDS scores, showing that as more detail is considered, the variances in shape as a result of the different NPs is more profound.

The results in this work have shown that the SOs considered here are conclusively morphologically representative of the Galactic Hii regions we observe in radio continuum surveys. To determine whether the SOs could potentially be used to construct a training set for supervised classification of Hii regions, via their shapes, a mass limited sample of the MAGPIS Hii regions were considered along with the 385 SOs. These results showed that there was good correspondence between respective early- and late-type Hii regions from each sample. This suggests that there is a lot of potential for the utilisation of the SOs to construct such a training set. For the SOs considered in this work, access was only available for one set of initial conditions. It was hence only associations between the evolutionary stages of the respective synthetic and observed sample that was investigated here. For a larger, diverse SO sample, of varying initial masses and ambient densities, across the different evolutionary stages, the results shown here suggest that predictions could be made as to the nature of Galactic Hii regions; based upon how their shapes compare to those of the simulations. This would be the next step to consider, in future work.

Chapter 6

Summary & Conclusions

This chapter covers the overall summary and conclusions from the analysis and results of this Thesis. Future work and applications of the methods presented in this Thesis are also outlined.

6.1 Summary

Introduction (Chapter 1)
A brief historical account of HII regions was given here, along with the overall aims of the Thesis and its structure. HII regions are diffuse nebulae of ionised hydrogen, excited by the extreme ultraviolet emission from massive stars. Due to the embedded nature of massive star formation, and their great distances, there are many observational difficulties involved when investigating such stars. HII regions, via their infrared and radio emission, highlight the location of these massive stars. Furthermore, due to their relatively short lifetimes, HII regions indicate where Galactic massive star formation is occurring.

The overall aim of the work presented herein, is to determine whether statistical shape analysis of observational data can be used as a way of investigating HII region and massive stellar properties. Development of a shape analysis and statistical clustering method would also allow for the synthetic observations of simulated HII regions to be quantitatively compared to their observational counterparts.

Theory, Observations & Simulations (Chapter 2)
In this chapter, the background theory relating to HII regions, their massive stellar sources, and the interstellar medium in which they are found was covered. Details were provided as to how Galactic structure is observed via the ISM and how GMCs are linked to massive star formation. The contentiousness of such formation was outlined. The theoretical interpretation of HII regions covered photoionisation, thermal, and dynamical affects. The equations governing the structure, source luminosity and dynamical evolution were explained, including how we observe HII regions at radio wavelengths.

Following this, an overview of Galactic Plane surveys was provided. The MAGPIS observational data, from which the HII region radio continuum data was taken, was detailed here. Also included was how the intensity units from the MAGPIS data were converted into a flux density, for use in calculations. The WISE HII Region Discovery Catalogue was introduced; and an overview of how distances to HII regions were determined was provided. Finally, the concept of numerical simulations were introduced. This covered the physical processes modelled in order to simulate the cloud, massive star, and resulting HII region, of which the synthetic observations were provided for analysis in this work. The objective here was to provide a working understanding of how the synthetic HII regions from these models are generated. Analogous to the outline given previously for the the MAGPIS HII region images.

Shape Analysis & Statistical Methods (Chapter 3)
This chapter introduced the concept of shape from both an astronomical and mathematical perspective. Morphological investigation of astronomical objects has proved to be a successful endeavour, particularly for galaxy classification. Mathematical shape concerns the geometric representation of an object, to which affects such as translation, rotation and scaling are invariant. Size-and-shape preserves scale as a measure. In this chapter, the systematic procedure for the extraction of shape from the MAGPIS HII region radio continuum data was detailed and justified based on examples shown in this chapter and the literature.

Boundaries of HII region shapes were extracted via image contouring of the radio emission at a determined signal level above the background noise. Interpolation splines fitted to these boundaries then allowed for the local curvature values to be calculated at equidistant points along the boundary. The curvature distribution for each HII region thus represented the shape-descriptor. The pairwise shape-distance between descriptors was then computed using the Anderson-Darling non-parametric test statistic. Hierarchical clustering was then performed on the distance matrix of A-D test scores. This resulted in a dendrogram representation of the HII region shapes, from which groups could be identified.

Radio Continuum Observations (Chapter 4)

Following the methods outlined in the previous chapter, this chapter details the statistical clustering to a selection of Galactic HII region shapes from the MAGPIS radio continuum data. A sample of 76 HII regions was identified from the image contouring procedure. Although there were more HII regions within the coverage of the MAGPIS survey, a stringent selection criteria was enforced in order to test the concept of the shape analysis method on a clean sample. Six groups were identified from the hierarchical clustering of the shape data. Visual inspection of the shapes belonging to resulting groups confirmed that the shape analysis was working as intended. The identified groups possessed different amounts of local curvature variation along the boundaries. Similarities between groups originating from the same parent branch on the dendrogram were also noted. This was further quantitatively confirmed by delineation of the curvature distributions for each group, and by the ordinance technique of multidimensional scaling (MDS).

There was little association between the region position in the Galaxy and assigned group, with objects at varying Galactocentric distance and Galactic latitude appearing in all groups. One group contained regions with Galactic longitude $< 30^\circ$ and another had only regions with $l > 28^\circ$. However, the remaining three groups contained regions with the full distribution of l values. One of the six groups contained only small ($<$1.6 pc), young ($<$0.5 Myr) HII regions. This was not exclusive, with 59% and 51% respective completeness. There was also a preference for this group to contain only low- to intermediate-mass ionising clusters (maximum of $\sim 10^3\,\mathrm{M}_\odot$). There was a further preference for ionising mass in another group that contained only intermediate- to high-mass clusters (minimum mass $\sim 10^{2.5}\,\mathrm{M}_\odot$). There was no further preference for regions of a given ionising mass to be placed in any particular group. Five outliers in the sample were discussed and whether there were possible projection effects or incorrect assumptions made as to the ambient density or electron temperatures; since in each case, there was either good resolution for or no kinematic distance ambiguity.

One of the identified groups was distinctly separated from the rest of the data. It was revealed to contain regions that were at the surface brightness detection limit for the survey. The groupings obtained remain present if this (or any individual) group is removed from the data set. The group associations also remain if a random number of regions are removed from the data. Further to this, it was evident that certain selection choices affect the resulting group structure from the hierarchical process. Lowering the threshold level used to define the 'edge' of the HII region resulted in higher curvature portions of the shape to be included in its descriptor; which had a much larger influence on the group results than increasing the threshold value. Increasing the threshold produced generally more smoothed edges with less defined detail. The largest source of error for the shape analysis was due to the sampling resolution used to obtain the curvature distributions, which is determined for each region from its distance. Due to the errors associated with kinematic distance determination, this is a limiting factor of the analysis process.

Synthetic Observations (Chapter 5)
This chapter detailed the application of the shape analysis methodology to a selection of HII regions from the numerical simulations of Ali et al. (2018). The numerical HII region was the result of photoionisation and feedback of a 34 $M_{\odot}$ star, in a 1000 $M_{\odot}$ cloud. From which, 77 synthetic observations (SOs) were provided, comprising four evolutionary snapshots (0.1, 0.2, 0.4 and 0.6 Myr), and multiple viewing projection angles. The shapes of the synthetic HII regions were extracted in the same manner as they were for the MAGPIS HII region shapes. The shape analysis results provided excellent confirmation of the efficacy of the numerical simulations. When considering the MAGPIS HII regions from Chap. 4 and 12 example SO HII regions with a Gaussian noise distribution, the SO HII regions were grouped in amongst the MAGPIS HII regions in the hierarchical clustering procedure. The same was found when directly inserting the SO regions to different MAGPIS tiles. By using five MAGPIS noise profiles for the 77 SOs, there were essentially 385 distinct observations of the numerical HII region.

From the hierarchical clustering of the 385 SOs that had been inserted into the MAGPIS tiles, the resulting hierarchy showed a clear divide between early- (0.1 and 0.2 Myr) and late-type (0.4 and 0.6 Myr) regions. There was no further association between the identified groupings and SO region radii, apart from the main split between the early- (small) and late-type (large) regions. There was no strong preference for SO regions from a given noise profile, nor given projection angle, to be assigned to specific groups. In terms of which of these parameters has more of an affect on the shape - for a given SO age and projection angle, not many of the five noise profiles were grouped together in the hierarchical clustering. For a given age and noise profile, however, there were more instances where numerous different projections were grouped together.

The MDS ordinance technique was used in order to try to find a systematic affect the different noise profiles had on the SO HII region shapes. For an example projection across the evolutionary stages, the shapes from the different noise profiles, along with the one extracted from the purely Gaussian noise profile, were compared. In each case, the Gaussian shape was ordinated away from the five MAGPIS noise profile shapes. Although, the ordination for each of the MAGPIS noise profile shapes with respect to each other and the Gaussian shape was not systematic. MDS did show systematic affects for how the shape of the HII region is extracted. Different initial contour levels changed the ordination in the MDS axes, but not the relative scores along the axes. Whilst different shape resolutions changed the scores along the axes but not the relative ordination positions. Higher resolutions corresponded to a larger spread in MDS scores, showing that as more detail is considered, the variances in shape as a result of the different noise profiles are more profound. All of these influences are to be considered in future applications of the shape extraction.

To determine whether the SOs could potentially be used to construct a training set for supervised classification of HII regions, via their shapes, a mass limited sample of the MAGPIS HII regions were considered along with the 385 SOs. These results showed that there was good correspondence between respective early- and late-type HII regions from each sample.

6.2 Overall Conclusions

Overall, the results presented in this Thesis show good evidence for the association of HII region shape to the region's physical parameters, and that shape can be used as an intrinsic measure of astronomical data. The significant caveat to this, as seen from the results and discussion, is that the definition of shape for such observations involves a number of selection choices; such as the signal level used to determine the boundaries and the resolution considered. However, as long as the extraction and representation is systematic across samples, many of the resulting effects can be quantified and constrained.

In terms of the observational sample considered, the results showed that for this section of the Galactic Plane, HII region shape is homogeneously distributed across Galactic latitude, with a small preference for those at similar Galactic longitudes to share a common morphology. Associations were also found between both the

Hii region age and the mass of the ionising source with the assigned group. The shape analysis method also identified regions that were at the surface brightness detection limit of the MAGPIS survey. As a proof of concept, these are all promising results for future potential applications.

The shape analysis method provided a means of quantitatively comparing synthetic observations of Hii regions to their observational counterparts. In this case, the SOs from Ali et al. (2018) were shown to be thoroughly representative of the MAGPIS Hii regions. The SOs allowed for investigation into different viewing projection angles, a concept that is only conceivable via the use of such simulations. For these data, projection was found to have less of an affect on the resulting shape than the distribution of noise around the Hii region signal. Associations between the respective age distributions of the SO and MAGPIS Hii region samples were also found in this work. The use of such SOs as a training set for supervised morphological classification of Hii regions is hence justified, and will be pursued in future work. For a larger training set, of varying masses and ambient densities, across the different evolutionary stages, the results shown here suggest that it should be possible to make predictions as to the nature of the Galactic Hii regions, based upon how their shapes compare to those of the highly representative simulations.

6.3 Future work

The foremost proposed application of this shape analysis research is to increase the number of parameters investigated by using SOs from numerical simulations of differing initial masses, ambient density and temperatures. If similar groupings that are shown in this investigation are seen for a more diverse sample, we will be one step closer to a thorough classification scheme. For example, it could be that the groups from the hierarchical procedure still differentiate between early- and late-type Hii regions, but also by intermediate mass and high mass ionising sources, a result that was observed for the Galactic regions in Chap. 4. Furthermore, with a different initial condition set-up, and a larger simulation grid, later evolutionary stages could be considered, increasing the parameter space further in this respect. In addition to this, the shape analysis method developed here provides a quantitative measure of how different SOs are from one another. Hence, the method could be applied to the simulation data to see quantitatively how different feedback mechanisms within the simulations affect the resulting structure of the Hii region via the shapes.

For the observational analysis of Galactic Hii regions, a further application would be to compare samples from different radio continuum surveys. The work carried out here regarding a systematic affect of the noise profiles on the HII regions was inconclusive, i.e. if there is a systematic affect, it as not revealed by this investigation. However, comparing different observational data with different noise profiles may lead to progress in this area. Similar to how the same SO with a different noise profile was considered, this could be performed for the same Galactic HII region from different surveys. A survey with complementary coverage to the MAGPIS survey that could be used is The HI/OH/Recombination line survey of the inner Milky Way (THOR) (Beuther et al., 2016). Another application could be to compare the shapes of the candidate HII regions in the WISE catalogue to those that are confirmed to be HII regions. This analysis method could provide justification for reclassifying such candidate objects. In addition to this, whilst we have only considered radio continuum images of HII regions in this work, a full classification scheme based on their morphologies should also consider the MIR data.

The analysis method presented in this Thesis has the potential to be adapted to a machine learning (ML) algorithm. If the hierarchical clustering procedure was able to decide whether to continue with certain groups or reject them based upon predefined criteria, the resulting training set could be more accurate for investigating parameters of the observed samples. For example, late-type regions being assigned to known early-type groups could be excluded from the procedure. The ML process could also consider the MDS investigation of the different NPs on the fly, rejecting any data with large ordination differences. This could be the natural evolution of this work, after the next steps of increasing the sample size of both the SO and Galactic samples.

There are various other diffuse astronomical objects that share common morphologies, such as planetary nebulae, supernova remnants, giant molecular clouds, and cold molecular clumps. If the shape data of these objects can be extracted in a systematic manner as it was here for the HII regions, this analysis method could be applied to these objects. Furthermore, since it was shown here that modern SOs are producing well representative HII regions; we could begin to investigate how observations of different objects and their corresponding simulations compare. An additional example is that HII regions could be compared to supernovae remnants, which are visually similar in the radio continuum images. It would be interesting to discover whether the shape analysis method presented here, that has been shown to be a robust quantisation of the region shapes, could tell the two nebulae apart. Looking further ahead, as telescope and imaging techniques continue to improve, there will be even more high resolution data to analyse and characterise. Hopefully, this kind of shape analysis methodology will continue to prove useful to astronomers, as we discover more and more about the Cosmos.

Appendix A

MAGPIS HII Region Properties

The following table lists the individual properties of all HII regions investigated in Chapter 4. Included are: the WISE (Anderson et al., 2014) and MWP (Simpson et al., 2012) catalogue name of the source, the C06 (Churchwell et al., 2006) name where applicable, the Galactic longitude and latitude, line of sight radial velocity and distances (taken from the WISE catalogue), angular and spatial effective radius, the ionised flux at 1.4 GHz that was determined from within the 1σ boundaries, the number of ionising photons, Strömgren radius, dynamical age and assigned group number from this analysis. Note that the two regions with group 6* are those originally labelled as group 7 in Fig. 4.3 and later merged into group 6.

Table A.1

WISE name	MWP name	C06	l	b	$V_{\rm lsr}$	D	$R_{\rm GC}$	$R_{\rm eff}$		Flux	N_{ly}	$R_{\rm s}$	t_{dyn}	G
			[deg]		[km/s]	[kpc]		[′]	[pc]	[Jy]	[s^{-1}]	[pc]	[Myr]	
G010.964+00.006	1G010965+000116		10.965	0.012	17.7	14.1	6.2	0.90	3.7	0.723	49.08	0.7	0.6	1
G012.429−00.049	1G012432−000410		12.432	−0.041	−18.4	22.6	14.6	2.30	15.1	0.982	49.62	1.1	5.4	2
G017.336−00.146	1G017330−001379		17.330	−0.138	−6.2	17.3	9.7	0.92	4.7	0.097	48.39	0.4	1.4	3
G017.928−00.677	1G017921−006843	N20	17.921	−0.684	39.1	12.8	5.5	0.70	2.6	0.080	48.04	0.3	0.6	4
G018.076+00.068	1G018081+000708		18.081	0.071	58.2	11.8	4.7	0.91	3.1	0.288	48.52	0.5	0.6	4
G018.144−00.281	1G018141−002783		18.141	−0.278	53.9	4.2	4.5	1.85	2.3	6.984	49.01	0.7	0.2	1
G018.152+00.090	1G018154+000992		18.154	0.099	53.0	4.1	4.6	1.32	1.6	0.200	47.45	0.2	0.4	4
G018.451−00.016	1G018452−000152		18.452	−0.015	56.5	11.9	4.8	1.05	3.6	0.485	48.76	0.6	0.7	4
G018.657−00.057	1G018658−000495		18.658	−0.050	44.1	12.5	5.3	0.52	1.9	0.146	48.28	0.4	0.3	1
G018.741+00.250	1G018743+002521		18.743	0.252	19.1	14.2	6.9	0.42	1.8	0.099	48.22	0.4	0.3	1
G019.504−00.193	1G019505−001900	N25	19.505	−0.190	37.8	12.8	5.7	0.79	2.9	0.110	48.18	0.4	0.7	1
G019.629−00.095	1G019632−001189		19.632	−0.119	58.6	11.7	4.8	3.94	13.4	4.028	49.66	1.1	4.3	3
G020.227+00.110	1G020223+001118		20.223	0.112	22.1	13.9	6.8	0.84	3.4	0.083	48.13	0.3	0.9	4
G020.988+00.092	1G020991+000963		20.991	0.096	18.6	14.1	7.0	1.33	5.5	1.396	49.36	0.9	1.0	1
G022.761−00.492	1G022756−004827		22.756	−0.483	74.8	4.8	4.3	1.99	2.8	2.769	48.73	0.6	0.4	5
G022.988−00.360	1G022991−003666		22.991	−0.367	74.1	4.8	4.3	1.95	2.7	1.039	48.30	0.4	0.6	4
G023.661−00.252	1G023660−002527		23.660	−0.253	66.2	11.2	4.9	0.90	2.9	0.224	48.37	0.4	0.6	1
G023.787+00.223	1G023798+002263		23.798	0.226	107.4	9.4	3.8	2.28	6.2	0.745	48.74	0.6	1.9	4
G024.728+00.159	1G024731+001580		24.731	0.158	109.3	6.3	3.7	0.78	1.4	0.110	47.56	0.2	0.3	5

Continued on next page

Table A.1 – continued from previous page

WISE name	MWP name	C06	l	b	V_{lsr}	D	R_{GC}	R_{eff}		Flux	N_{ly}	R_s	t_{dyn}	G
			[deg]		[km/s]	[kpc]		[′]	[pc]	[Jy]	[s^{-1}]	[pc]	[Myr]	
G024.734+00.087	1G024735+000889		24.735	0.089	111.3	6.4	3.7	1.93	3.6	0.818	48.45	0.4	0.9	1
G025.397+00.033	1G025399+000360		25.399	0.036	−14.4	17.3	10.4	1.62	8.2	1.097	49.44	0.9	2.0	3
G027.476+00.179	1G027484+001817		27.484	0.182	34.0	12.6	6.5	3.73	13.7	4.424	49.77	1.2	4.2	4
G027.682+00.076	1G027688+000778		27.688	0.078	100.2	6.0	4.1	0.96	1.7	0.119	47.55	0.2	0.4	5
G027.980+00.080	1G027981+000753		27.981	0.075	76.6	10.3	4.9	0.54	1.6	0.229	48.31	0.4	0.2	1
G027.997+00.317	1G028006+003153		28.006	0.315	92.7	9.4	4.4	2.03	5.5	0.369	48.43	0.4	1.9	4
G028.146+00.146	1G028140+001526		28.140	0.153	89.2	5.4	4.4	1.70	2.7	0.287	47.84	0.3	0.7	6
G028.638+00.194	1G028636+002033		28.636	0.203	103.8	7.5	4.0	1.80	3.9	0.412	48.29	0.4	1.1	4
G030.252+00.053	1G030250+000547		30.250	0.241	65.2	4.5	5.0	0.59	0.8	0.096	47.21	0.2	0.1	5
G030.311−00.215	1G030323−002072		30.329	−0.209	106.1	7.3	4.2	0.91	1.9	0.153	47.83	0.3	0.4	4
G030.468+00.394	1G030461+004237		30.461	0.424	57.7	3.8	5.4	3.53	3.9	1.334	48.21	0.4	1.1	6*
G030.690−00.258	1G030699−002560		30.699	−0.256	98.5	8.4	4.4	2.17	5.3	2.713	49.20	0.8	1.1	4
G031.138+00.285	1G031147+002879	N54	31.147	0.288	104.7	7.3	4.3	1.89	4.0	1.192	48.72	0.5	0.9	3
G031.470−00.344	1G031473−003459		31.473	−0.346	88.8	9.0	4.7	1.12	2.9	0.477	48.51	0.5	0.6	4
G032.057+00.077	1G032057+000783		32.057	0.078	96.3	7.2	4.4	0.94	2.0	0.263	48.06	0.3	0.4	5
G032.152+00.131	1G032158+001306		32.158	0.131	95.0	6.1	4.5	0.78	1.4	0.740	48.36	0.4	0.2	5
G032.582+00.001	1G032584+000057	N56	32.584	0.006	77.4	9.4	5.1	1.09	3.0	0.163	48.08	0.3	0.8	6
G032.733+00.209	1G032731+002120		32.731	0.212	16.1	13.2	7.7	0.45	1.7	0.030	47.64	0.2	0.4	4
G032.928+00.606	1G032929+006055		32.929	0.605	−38.3	19.2	13.0	0.45	2.5	0.293	48.95	0.7	0.3	5

Continued on next page

Table A.1 – continued from previous page

WISE name	MWP name	C06	l	b	V$_{\rm lsr}$	D	R$_{\rm GC}$	R$_{\rm eff}$		Flux	N$_{ly}$	R$_{\rm s}$	t$_{dyn}$	G
			[deg]		[km/s]	[kpc]		[′]	[pc]	[Jy]	[s^{-1}]	[pc]	[Myr]	
G033.813−00.150	1G033815−001494	N60	33.815	−0.149	50.0	10.8	6.0	0.41	1.3	0.040	47.59	0.2	0.2	5
G034.089+00.438	1G034088+004405		34.088	0.441	32.6	11.8	6.8	0.69	2.4	0.219	48.41	0.4	0.4	3
G034.133+00.471	1G034132+004724		34.132	0.472	34.6	11.7	6.7	0.40	1.4	0.586	48.83	0.6	0.1	3
G034.256+00.136	1G034262+001267		34.261	0.136	54.0	3.5	5.8	2.08	2.1	10.758	49.04	0.7	0.2	4
G035.051−00.520	1G035051−005195		35.051	−0.519	48.0	10.7	6.2	0.39	1.2	0.215	48.31	0.4	0.1	5
G035.467+00.004	1G035476+000027		35.476	0.003	58.6	10.0	5.8	1.08	3.1	0.169	48.15	0.4	0.8	4
G035.543+00.006	1G035544+000131	N67	35.544	0.013	57.5	10.1	5.9	0.87	2.5	0.218	48.27	0.4	0.5	3
G035.571+00.071	1G035573+000703		35.573	0.070	47.6	10.7	6.2	0.48	1.5	0.455	48.64	0.5	0.1	5
G035.649−00.053	1G035652−000348	N68	35.652	−0.035	51.9	10.4	6.1	4.21	12.7	2.853	49.41	0.9	4.6	6
G037.259−00.083	1G037261−000809		37.261	−0.081	40.8	10.8	6.5	0.81	2.6	0.122	48.07	0.3	0.6	4
G037.344+00.684	1G037349+006876		37.349	0.688	45.6	10.5	6.4	0.60	1.8	0.176	48.21	0.4	0.3	5
G037.750−00.110	1G037751−001098	N70	37.751	−0.110	49.7	10.1	6.2	0.87	2.6	0.374	48.50	0.5	0.4	3
G037.754+00.560	1G037754+005577		37.755	0.560	18.3	12.2	7.6	0.38	1.4	0.054	47.82	0.3	0.2	4
G038.045−00.034	1G038046−000318		38.046	−0.032	58.3	9.5	5.9	1.57	4.3	0.268	48.31	0.4	1.3	4
G038.840+00.495	1G038839+004990		38.839	0.499	−43.2	18.2	12.8	0.58	3.1	0.041	48.05	0.3	0.8	3
G039.728−00.396	1G039728−003965		39.728	−0.397	58.3	9.1	6.0	0.48	1.3	0.253	48.24	0.4	0.1	5
G039.864+00.645	1G039868+006467		39.868	0.647	−40.9	17.6	12.4	0.46	2.4	0.128	48.52	0.5	0.4	5
G039.873−00.177	1G039874−001737		39.874	−0.174	60.0	9.0	5.9	0.82	2.1	0.110	47.87	0.3	0.5	4
G041.235+00.367	1G041237+003647		41.237	0.365	71.3	5.4	5.5	1.84	2.9	0.375	47.96	0.3	0.8	6

Continued on next page

Table A.1 – continued from previous page

WISE name	MWP name	C06	l	b	V$_{\rm lsr}$	D	R$_{\rm GC}$	R$_{\rm eff}$		Flux	N$_{ly}$	R$_{\rm s}$	t$_{dyn}$	G
			[deg]		[km/s]	[kpc]		[′]	[pc]	[Jy]	[s^{-1}]	[pc]	[Myr]	
G041.512+00.021	1G041519+000375	N79	41.519	0.037	17.7	11.6	7.7	2.06	7.0	0.602	48.83	0.6	2.2	6
G041.929+00.030	1G041931+000351	N80	41.931	0.035	20.7	11.3	7.6	1.45	4.8	0.193	48.31	0.4	1.5	6
G042.103−00.623	1G042104−006230	N82	42.104	−0.623	66.0	7.7	5.8	1.81	4.1	1.389	48.84	0.6	0.8	3
G042.111−00.444	1G042112−004451	N83	42.112	−0.445	53.4	8.9	6.2	0.49	1.3	0.181	48.08	0.3	0.2	5
G042.434−00.275	1G042426−002691		42.426	−0.269	61.5	8.1	5.9	1.23	2.9	0.715	48.59	0.5	0.5	6
G043.185−00.525	1G043185−005229		43.185	−0.523	59.1	8.1	6.0	0.55	1.3	0.713	48.59	0.5	0.1	5
G043.432+00.516	1G043434+005189		43.434	0.519	−12.8	13.6	9.5	1.06	4.2	0.104	48.21	0.4	1.3	6
G043.774+00.057	1G043775+000606	N90	43.775	0.061	70.5	6.1	5.7	1.37	2.4	0.153	47.68	0.2	0.7	6*
G043.792−00.122	1G043797−001171		43.789	−0.116	43.3	9.2	6.6	0.56	1.5	0.064	47.65	0.2	0.3	5
G043.818+00.395	1G043829+003957		43.829	0.396	−10.5	13.3	9.3	1.26	4.9	0.216	48.50	0.5	1.4	6
G044.418+00.535	1G044417+005356		44.417	0.536	−55.1	18.4	13.8	0.47	2.5	0.042	48.08	0.3	0.6	3
G044.501+00.332	1G044503+003373		44.503	0.337	−43.0	16.6	12.2	1.29	6.2	0.246	48.75	0.6	1.9	6
G045.118+00.144	1G045123+001453		45.123	0.145	56.7	7.7	6.2	1.19	2.7	3.881	49.28	0.8	0.3	3
G045.197+00.740	1G045204+007443		45.204	0.744	−35.0	15.5	11.3	1.17	5.3	0.177	48.55	0.5	1.6	4
G045.391−00.725	1G045387−007155	N95	45.387	−0.715	52.5	8.0	6.3	1.69	3.9	0.276	48.17	0.4	1.2	6
G045.475+00.130	1G045476+001340		45.476	0.134	55.6	7.7	6.2	1.00	2.2	1.747	48.94	0.6	0.3	3
G045.825−00.291	1G045822−002892		45.822	−0.289	61.2	6.6	6.0	0.71	1.4	0.336	48.09	0.3	0.2	3
G045.933−00.403	1G045934−004012		45.934	−0.401	63.9	5.9	6.0	0.68	1.2	0.071	47.32	0.2	0.2	5
G046.495−00.241	1G046481−002389		46.481	−0.239	57.2	4.7	6.1	3.66	5.0	2.094	48.59	0.5	1.4	6

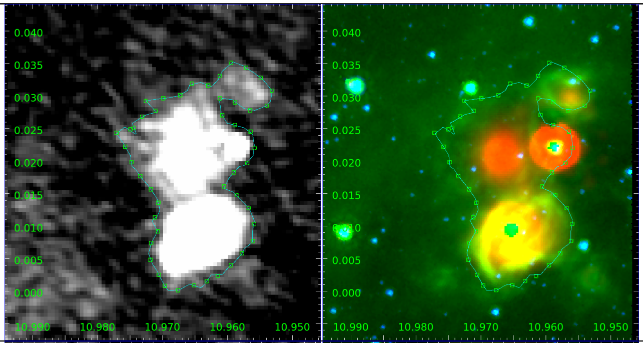
Appendix B

MAGPIS Images by Group

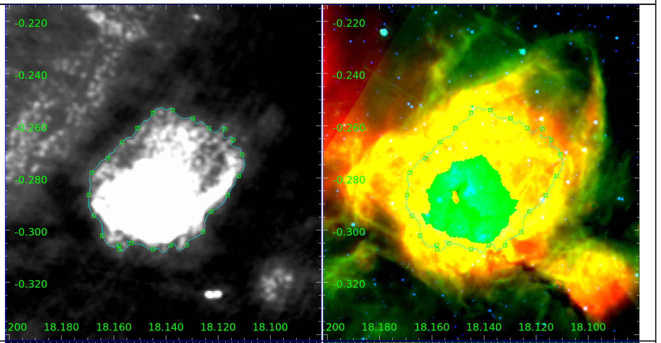
Hii region images for the six identified groups shown at varying scales, coordinates are Galactic l and b, North is up, East is left. 1.4 GHz radio continuum MAGPIS image shown left in grey-scale, with the identified boundary at 1σ above the noise. Points along the boundary represent the location of the interpolation spline knots, where the curvature was determined. The average spacing of these knots is 0.54 pc. The mid infrared RGB *Spitzer* images are shown right with the same boundary overlaid. The blue and green channels are the 4.5 μm and 8.0 μm IRAC bands, respectively, the red channel is the 24 μm MIPSGAL band. Note that saturated pixels in the red channel are shown as transparent. For Group 6, the two regions with * after their name are the regions originally labelled as group 7 on the dendrogram in Fig. 4.3, later merged with group 6 following the bootstrapping results.

Group 1

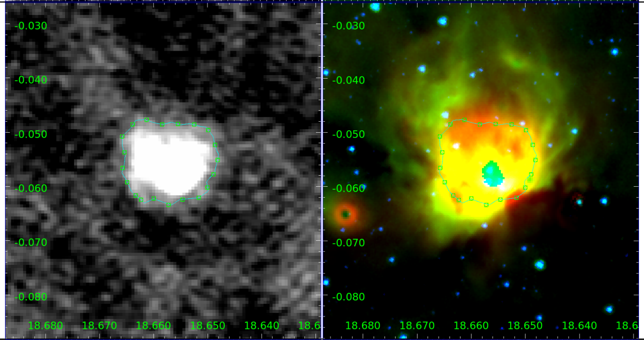
G010.964+00.006



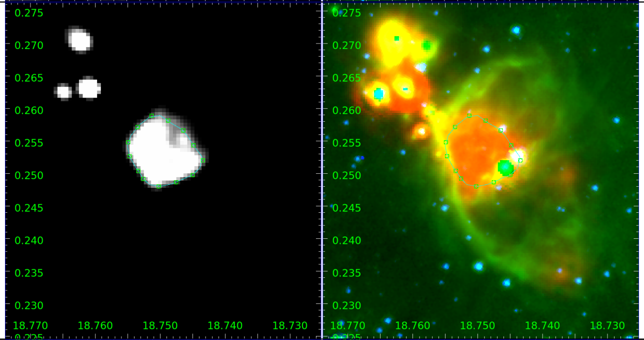
G018.144−00.281



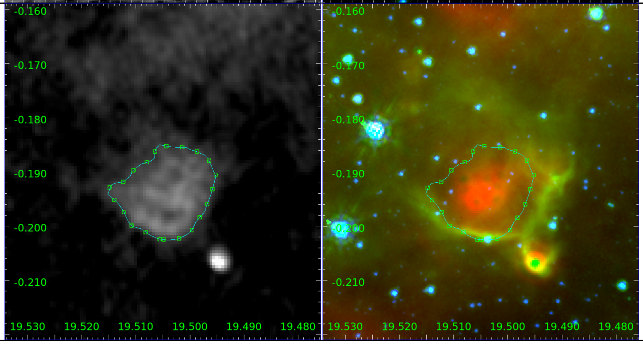
G018.657−00.057



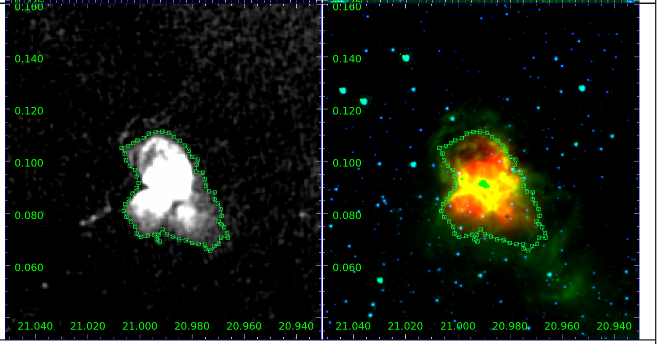
G018.741+00.250



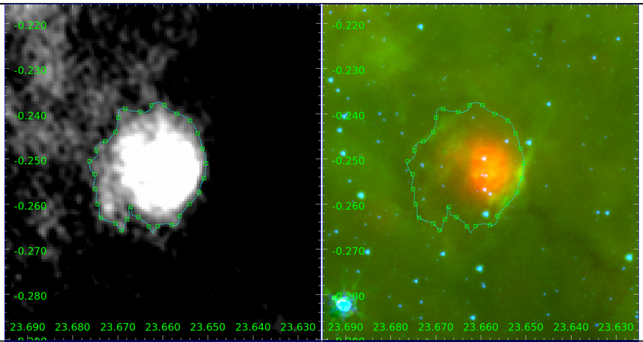
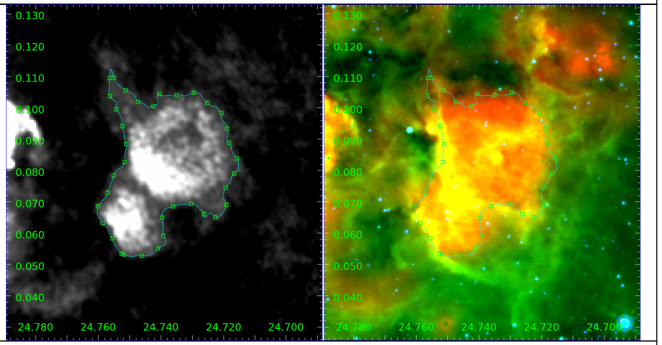
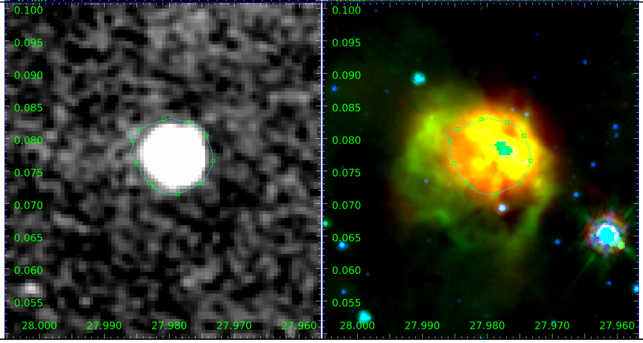
G019.504−00.193



G020.988+00.092

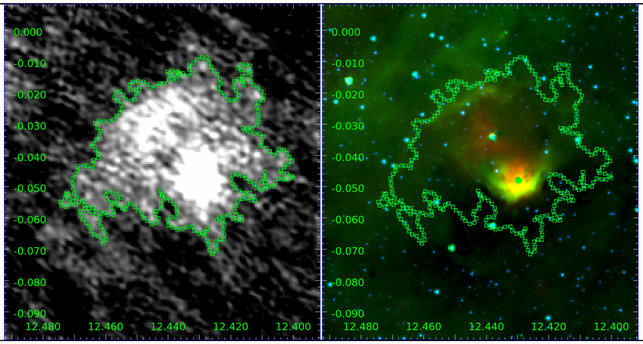


Continued on next page

Group 1			
G023.661−00.252	 -0.220 -0.230 -0.240 -0.250 -0.260 -0.270 -0.280 23.690 23.680 23.670 23.660 23.650 23.640 23.630	G024.734+00.087	 0.130 0.120 0.110 0.100 0.090 0.080 0.070 0.060 0.050 0.040 24.780 24.760 24.740 24.720 24.700
G027.980+00.080	 0.100 0.095 0.090 0.085 0.080 0.075 0.070 0.065 0.060 0.055 28.000 27.990 27.980 27.970 27.960		

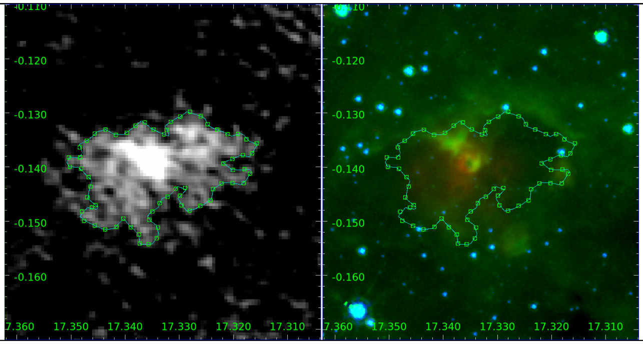
Group 2

G012.429−00.049

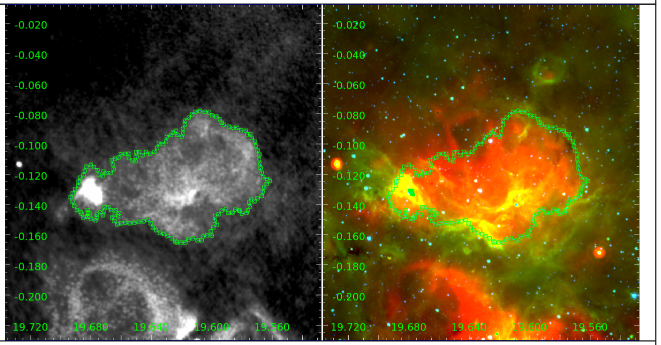


Group 3

G017.336−00.146



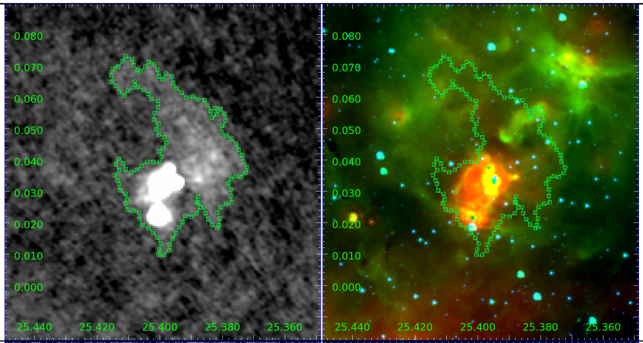
G019.629−00.095



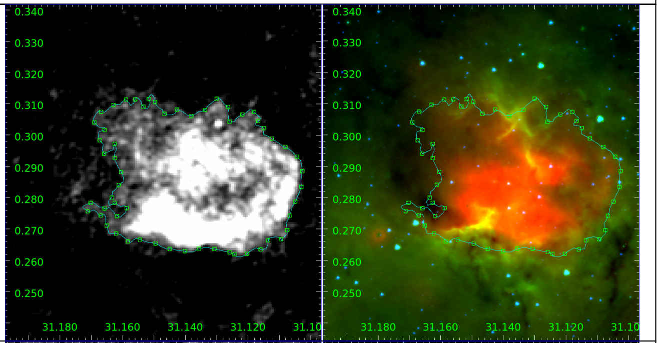
Continued on next page

Group 3

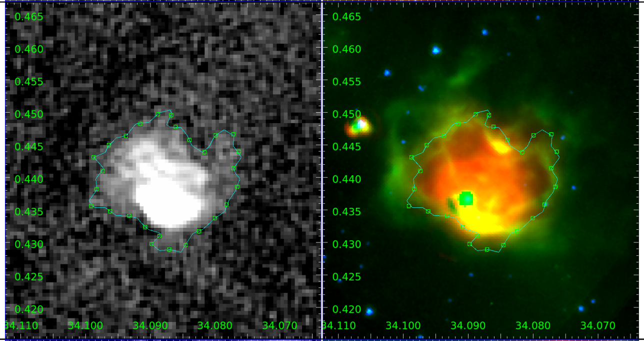
G025.397+00.033



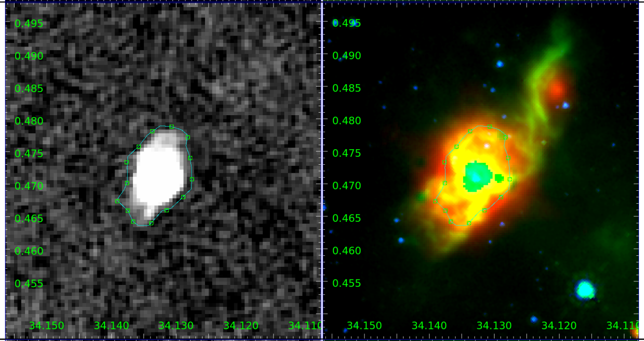
G031.138+00.285



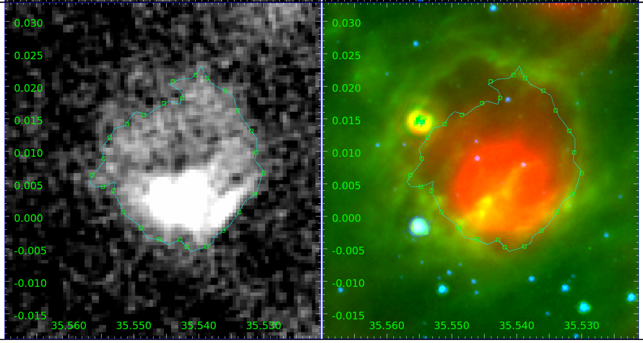
G034.089+00.438



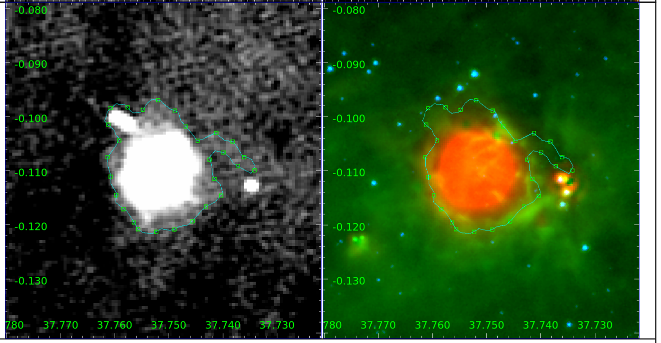
G034.133+00.471



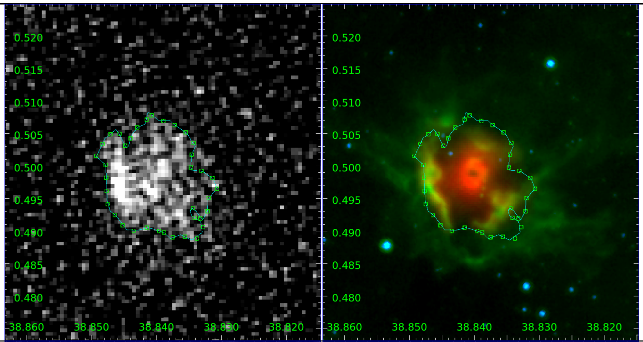
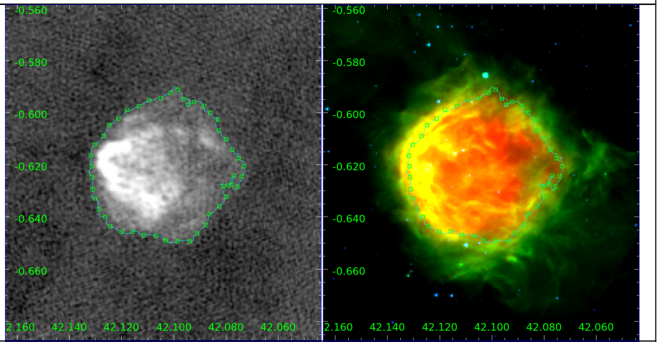
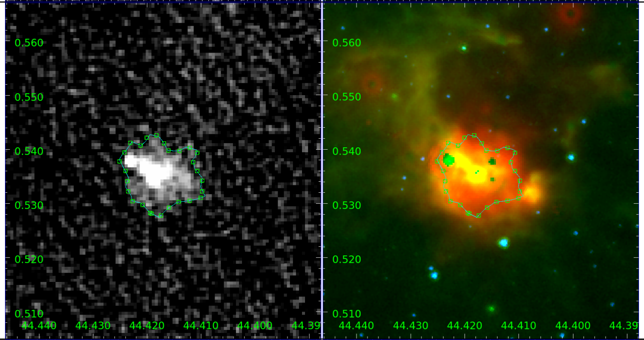
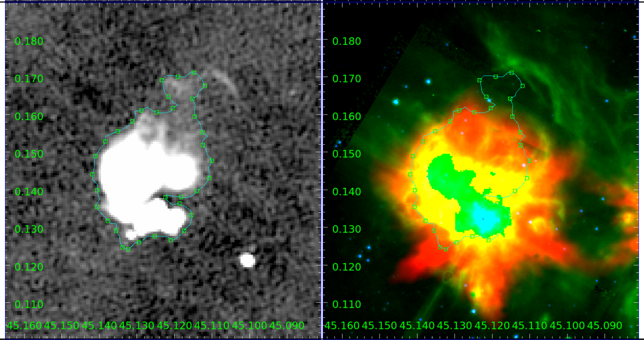
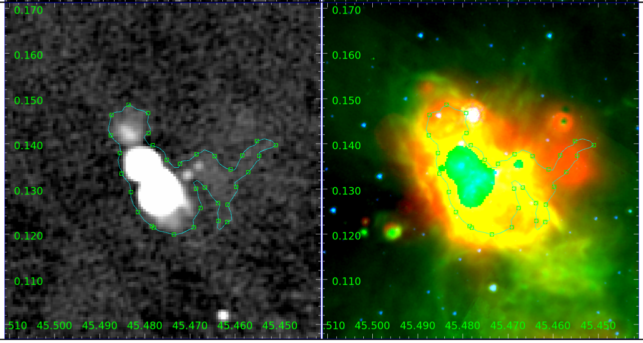
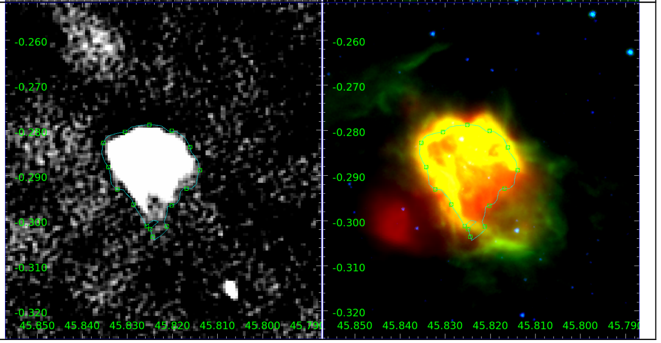
G035.543+00.006



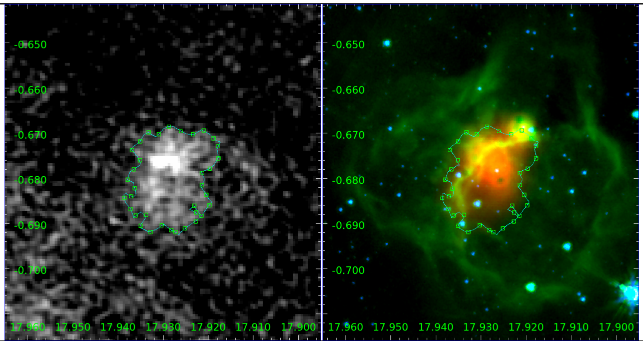
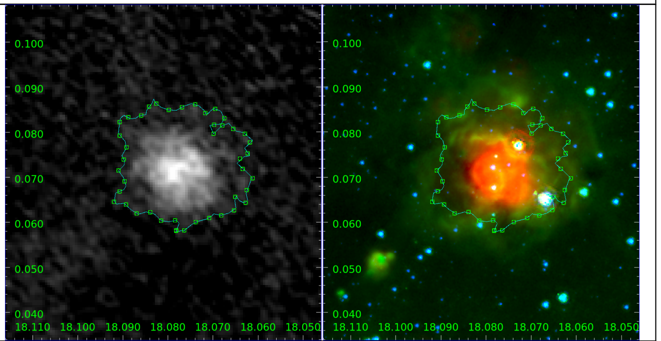
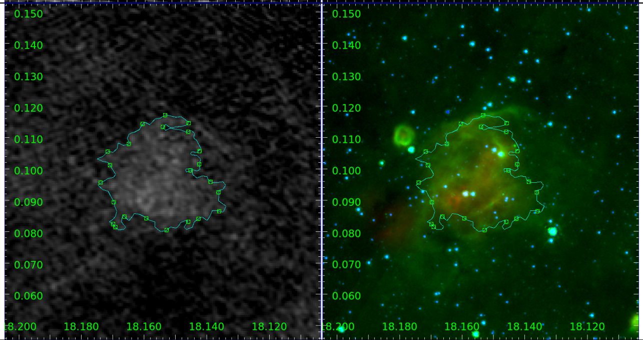
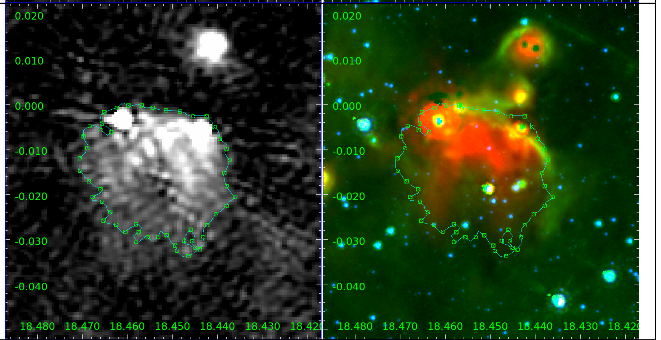
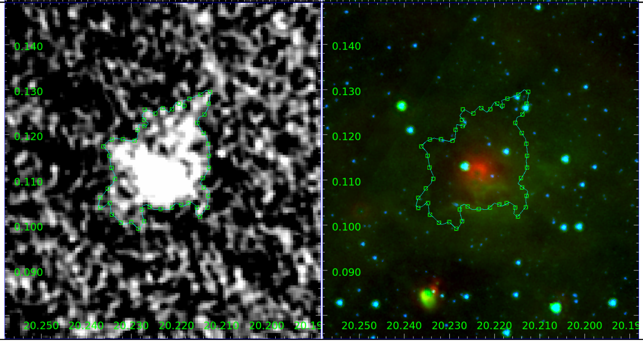
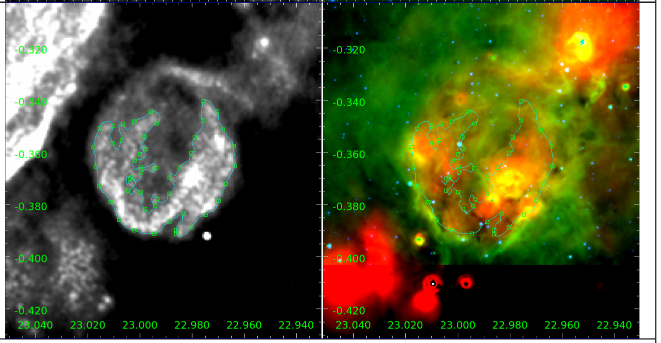
G037.750−00.110



Continued on next page

Group 3			
G038.840+00.495		G042.103−00.623	
G044.418+00.535		G045.118+00.144	
G045.475+00.130		G045.825−00.291	

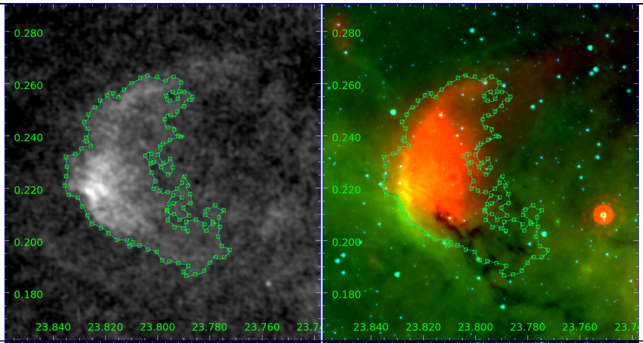
Group 4

G017.928−00.677		G018.076+00.068	
G018.152+00.090		G018.451−00.016	
G020.227+00.110		G022.988−00.360	

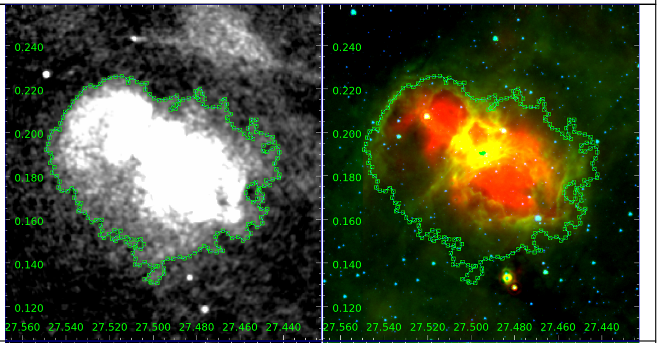
Continued on next page

Group 4

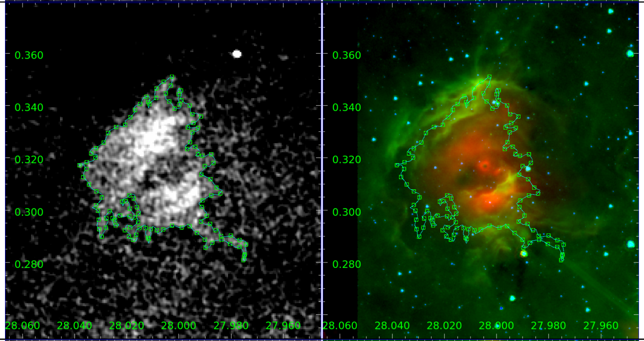
G023.787+00.223



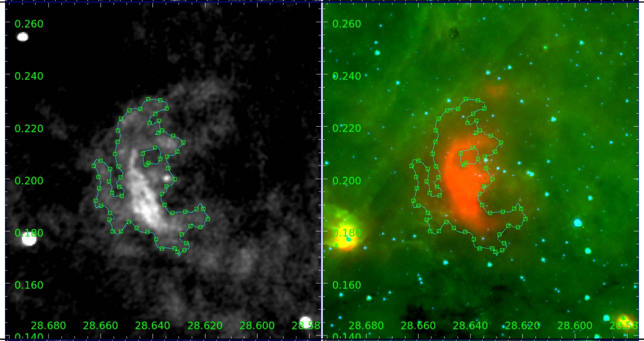
G027.476+00.179



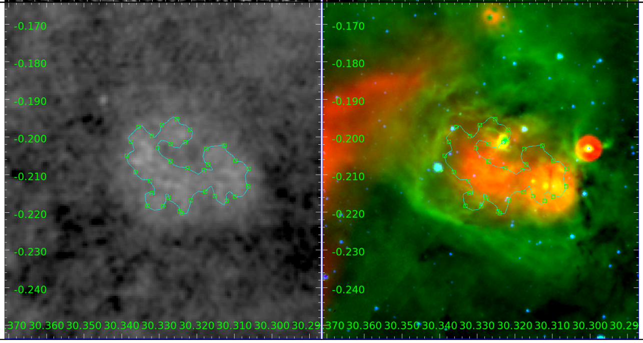
G027.997+00.317



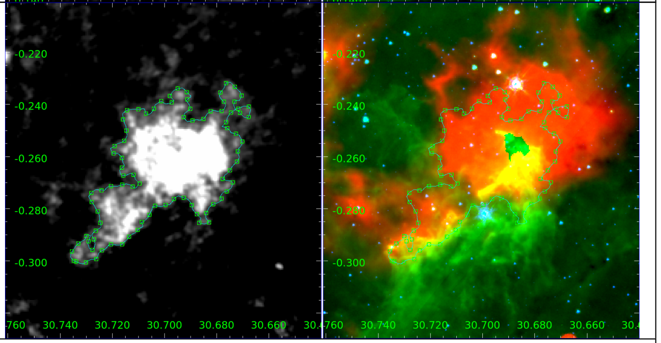
G028.638+00.194



G030.311−00.215



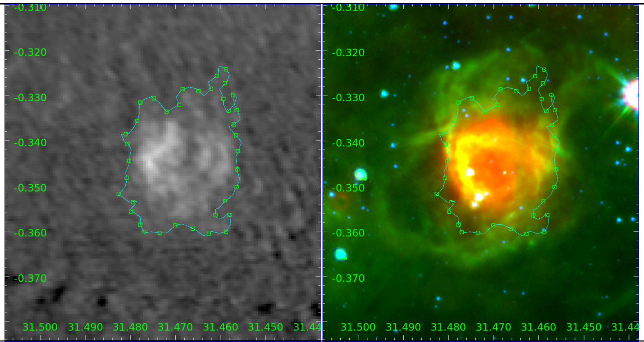
G030.690−00.258



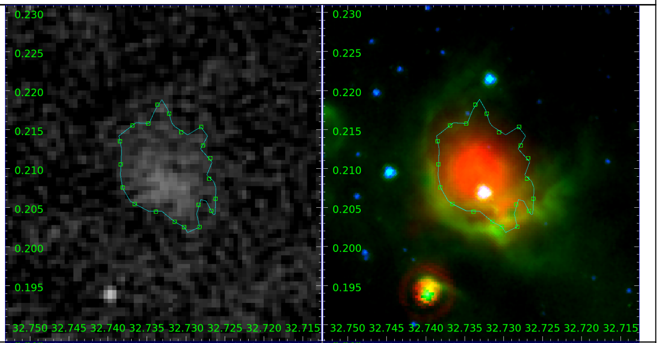
Continued on next page

Group 4

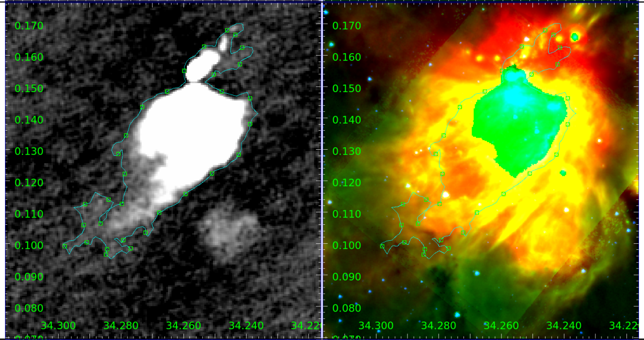
G031.470−00.344



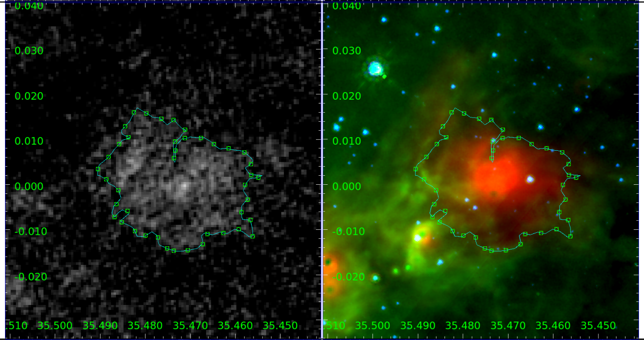
G032.733+00.209



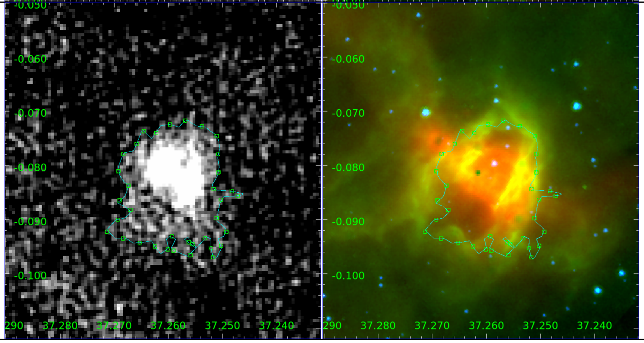
G034.256+00.136



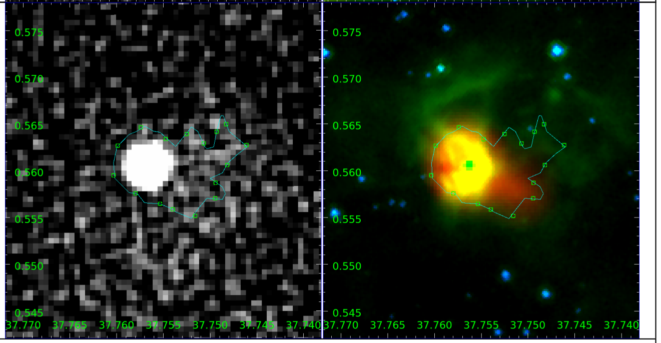
G035.467+00.004



G037.259−00.083

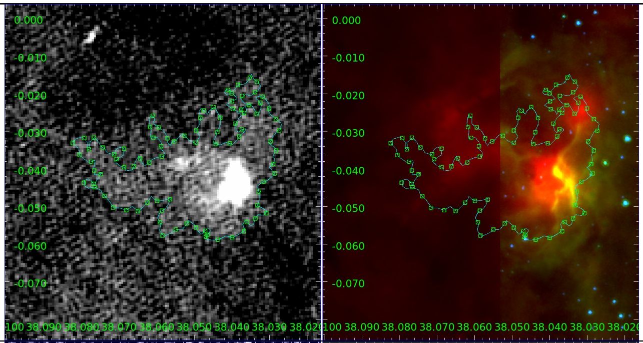
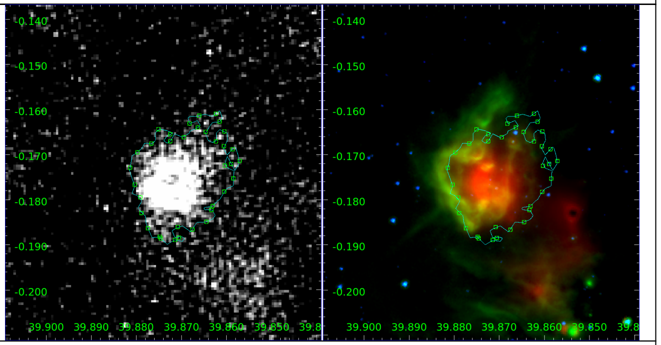
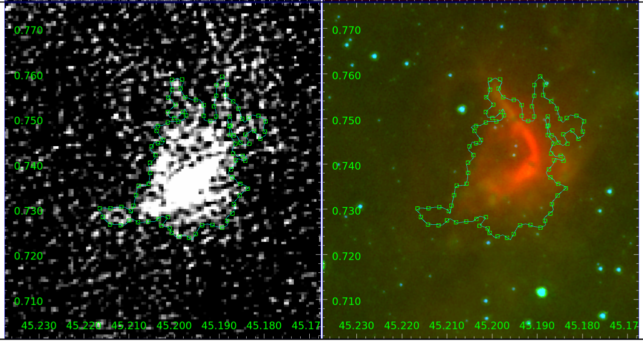


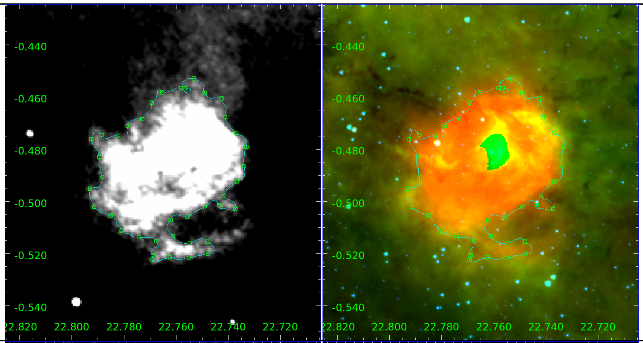
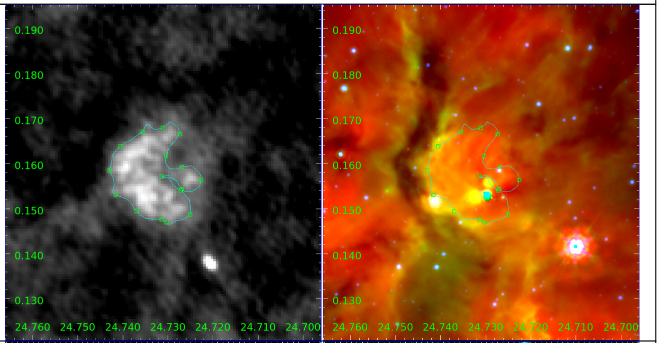
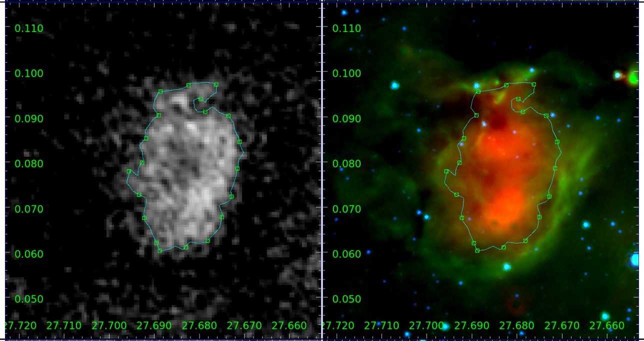
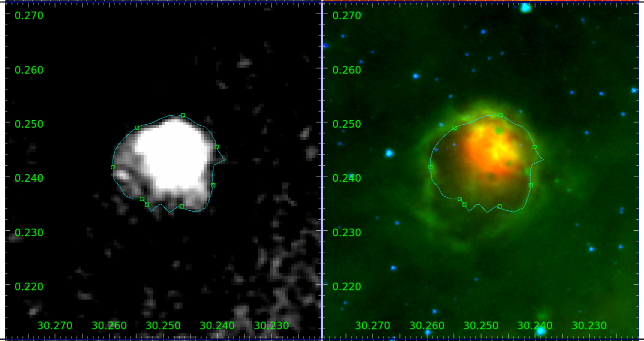
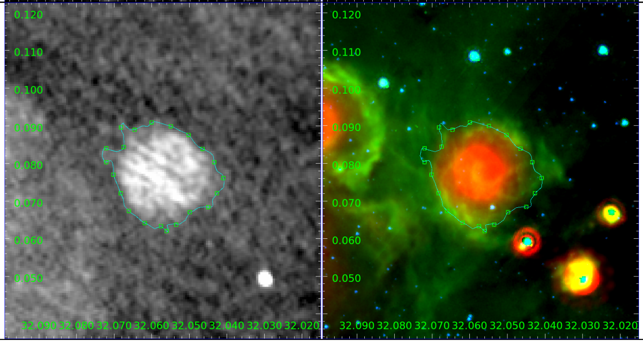
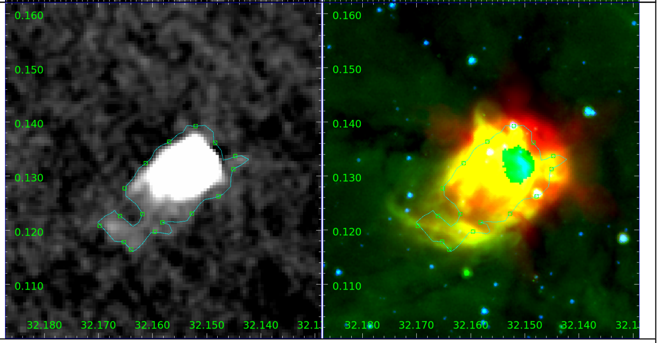
G037.754+00.560



Continued on next page

Group 4

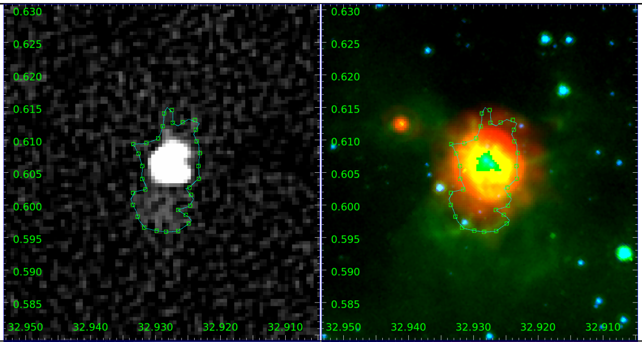
G038.045−00.034	 0.000 -0.010 -0.020 -0.030 -0.040 -0.050 -0.060 -0.070 38.100 38.090 38.080 38.070 38.060 38.050 38.040 38.030 38.020	G039.873−00.177	 -0.140 -0.150 -0.160 -0.170 -0.180 -0.190 -0.200 39.900 39.890 39.880 39.870 39.860 39.850 39.8
G045.197+00.740	 0.770 0.760 0.750 0.740 0.730 0.720 0.710 45.230 45.220 45.210 45.200 45.190 45.180 45.17		

Group 5			
G022.761−00.492	 -0.440 -0.460 -0.480 -0.500 -0.520 -0.540 22.820 22.800 22.780 22.760 22.740 22.720	G024.728+00.159	 0.190 0.180 0.170 0.160 0.150 0.140 0.130 24.760 24.750 24.740 24.730 24.720 24.710 24.700
G027.682+00.076	 0.110 0.100 0.090 0.080 0.070 0.060 0.050 27.720 27.710 27.700 27.690 27.680 27.670 27.660	G030.252+00.053	 0.270 0.260 0.250 0.240 0.230 0.220 30.270 30.260 30.250 30.240 30.230
G032.057+00.077	 0.120 0.110 0.100 0.090 0.080 0.070 0.060 0.050 32.090 32.080 32.070 32.060 32.050 32.040 32.030 32.020	G032.152+00.131	 0.160 0.150 0.140 0.130 0.120 0.110 32.180 32.170 32.160 32.150 32.140

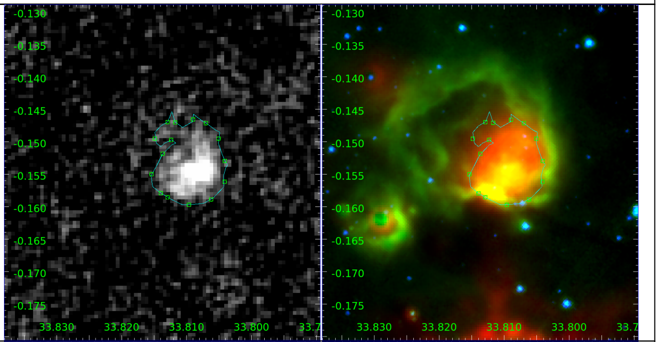
Continued on next page

Group 5

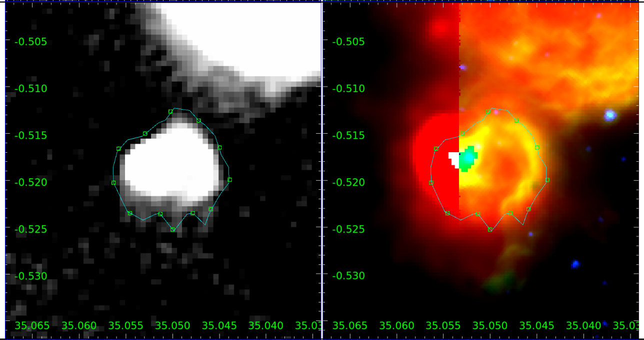
G032.928+00.606



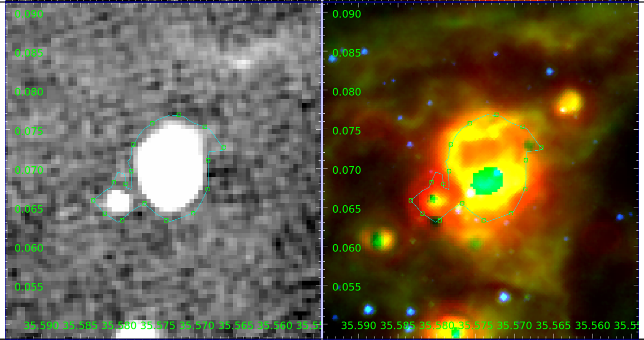
G033.813−00.150



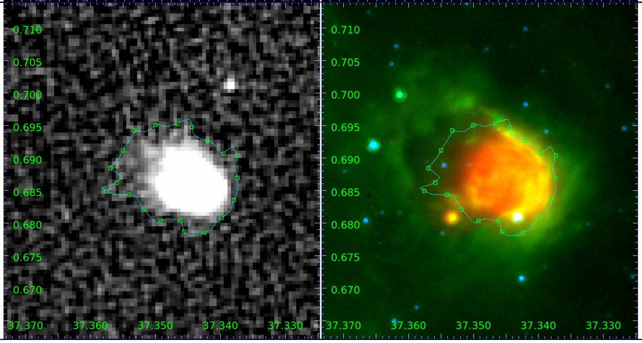
G035.051−00.520



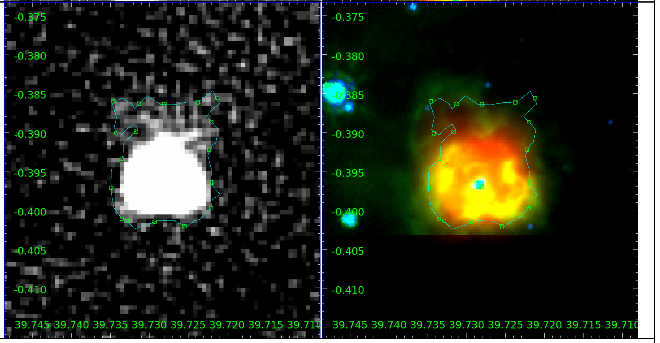
G035.571+00.071



G037.344+00.684



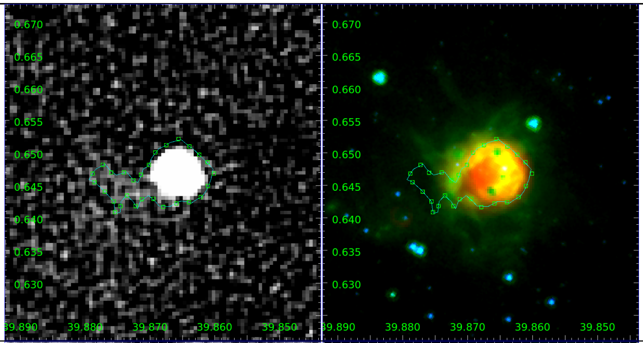
G039.728−00.396



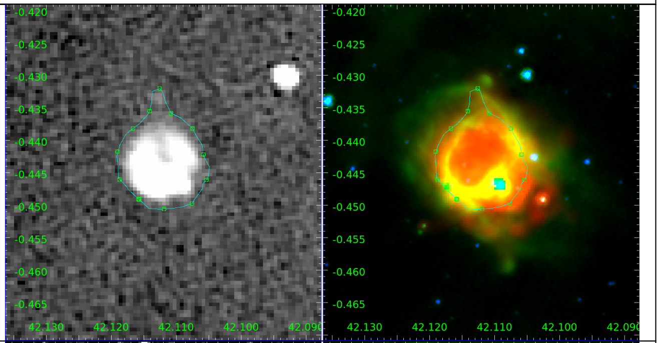
Continued on next page

Group 5

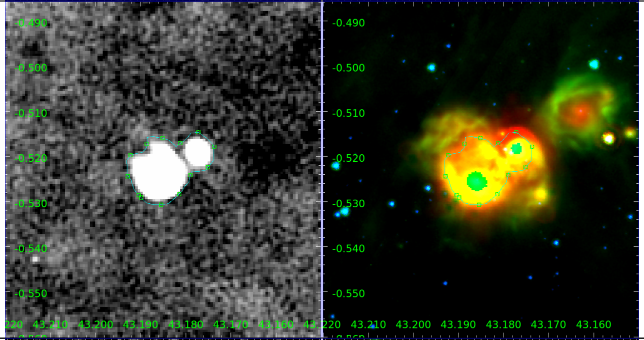
G039.864+00.645



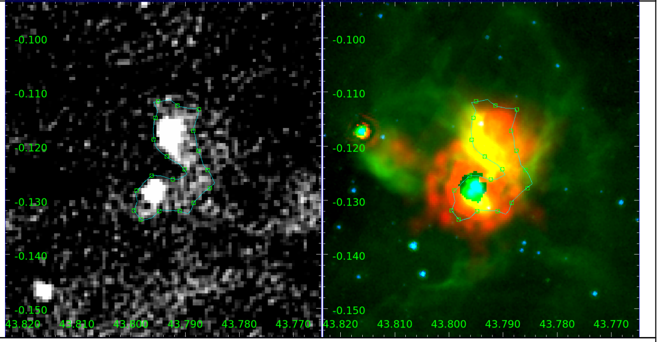
G042.111−00.444



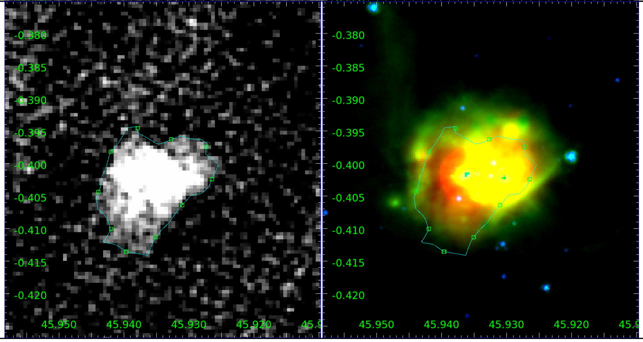
G043.185−00.525



G043.792−00.122

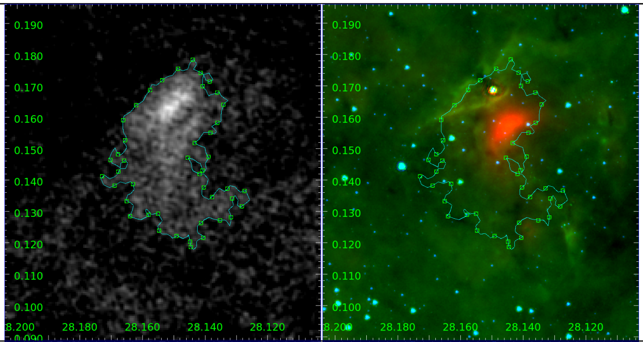


G045.933−00.403

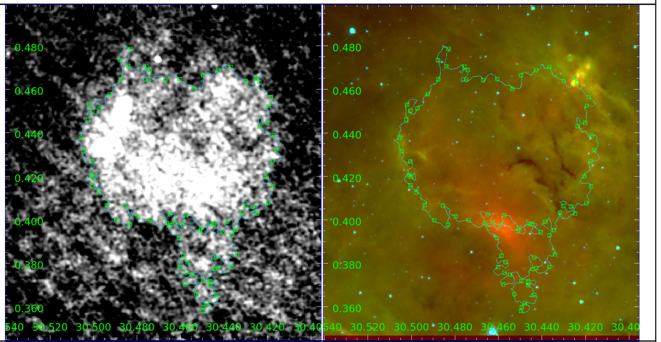


Group 6

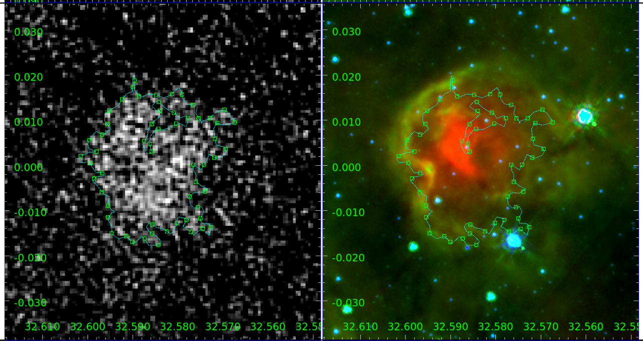
G028.146+00.146



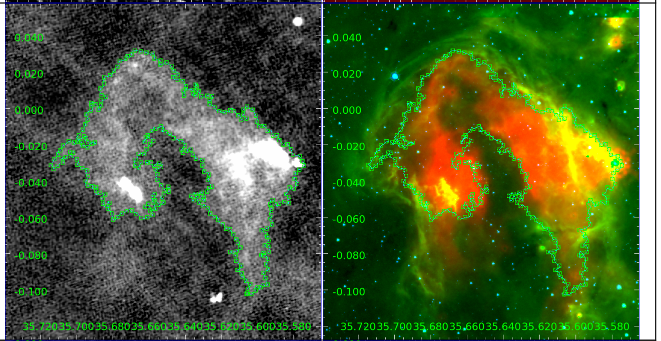
G030.468+00.394 *



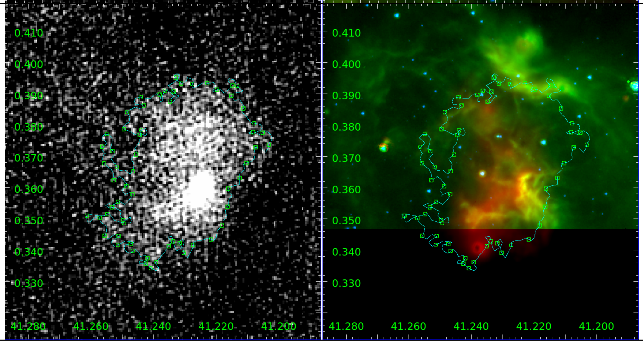
G032.582+00.001



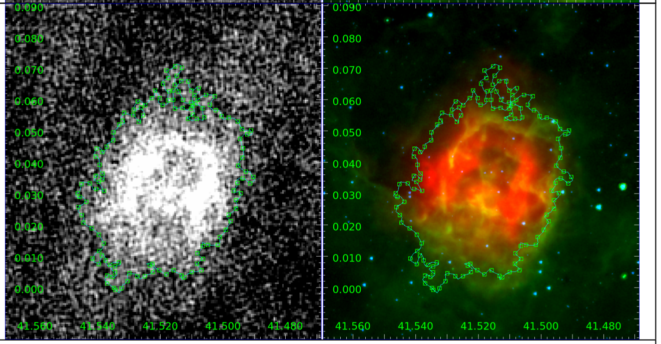
G035.649−00.053



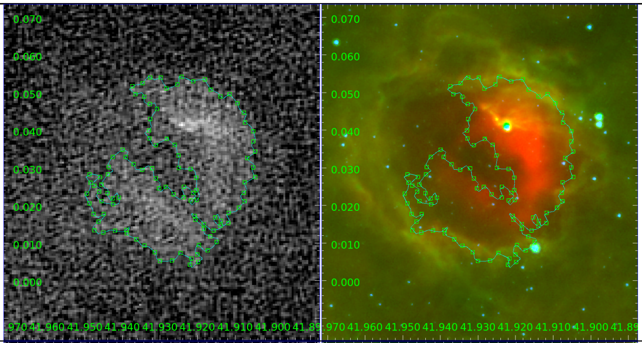
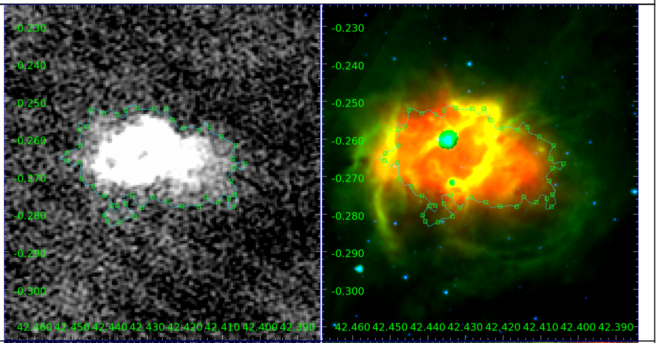
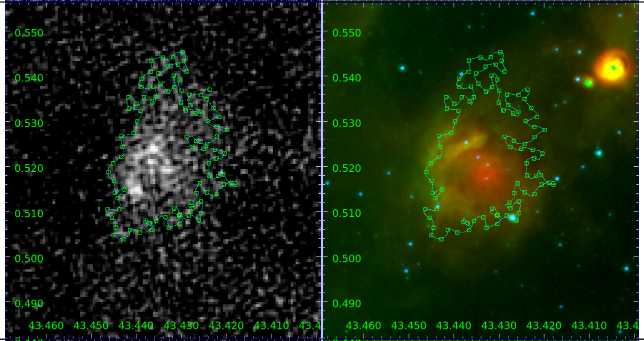
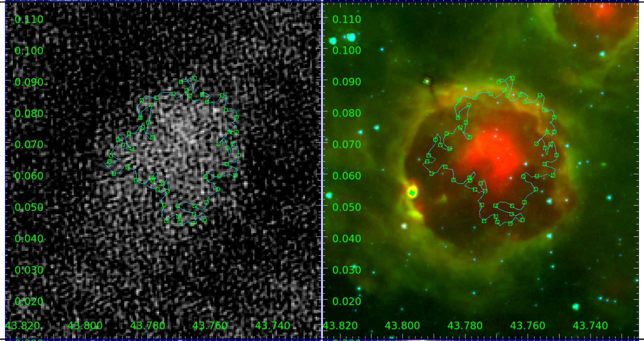
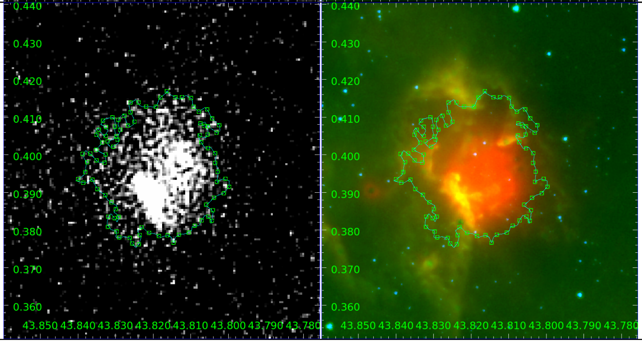
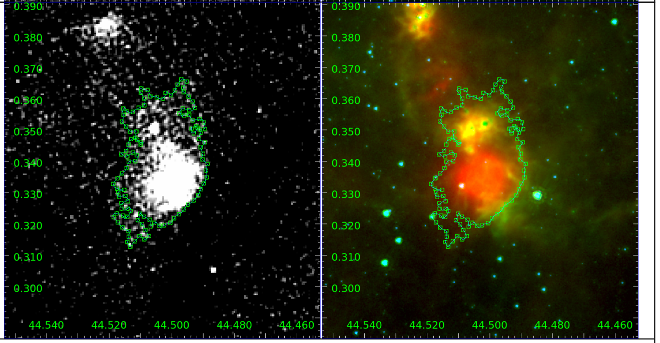
G041.235+00.367



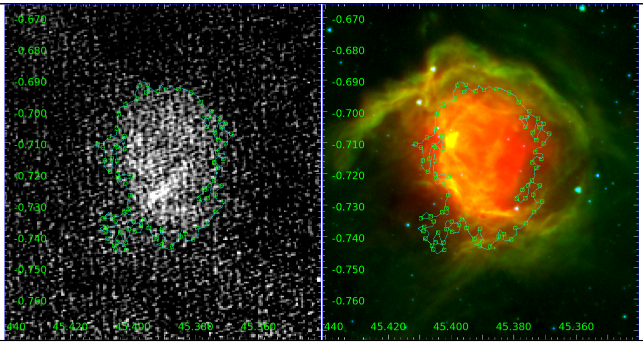
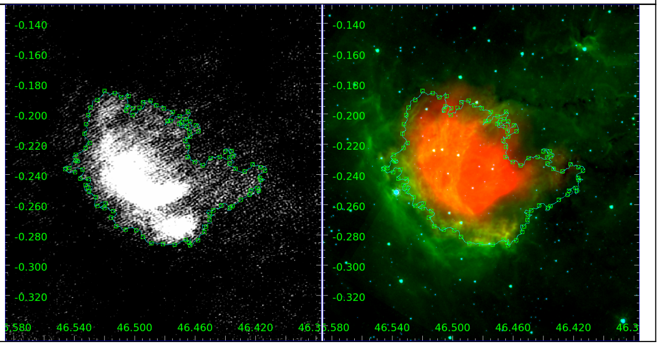
G041.512+00.021



Continued on next page

Group 6			
G041.929+00.030		G042.434−00.275	
G043.432+00.516		G043.774+00.057 *	
G043.818+00.395		G044.501+00.332	

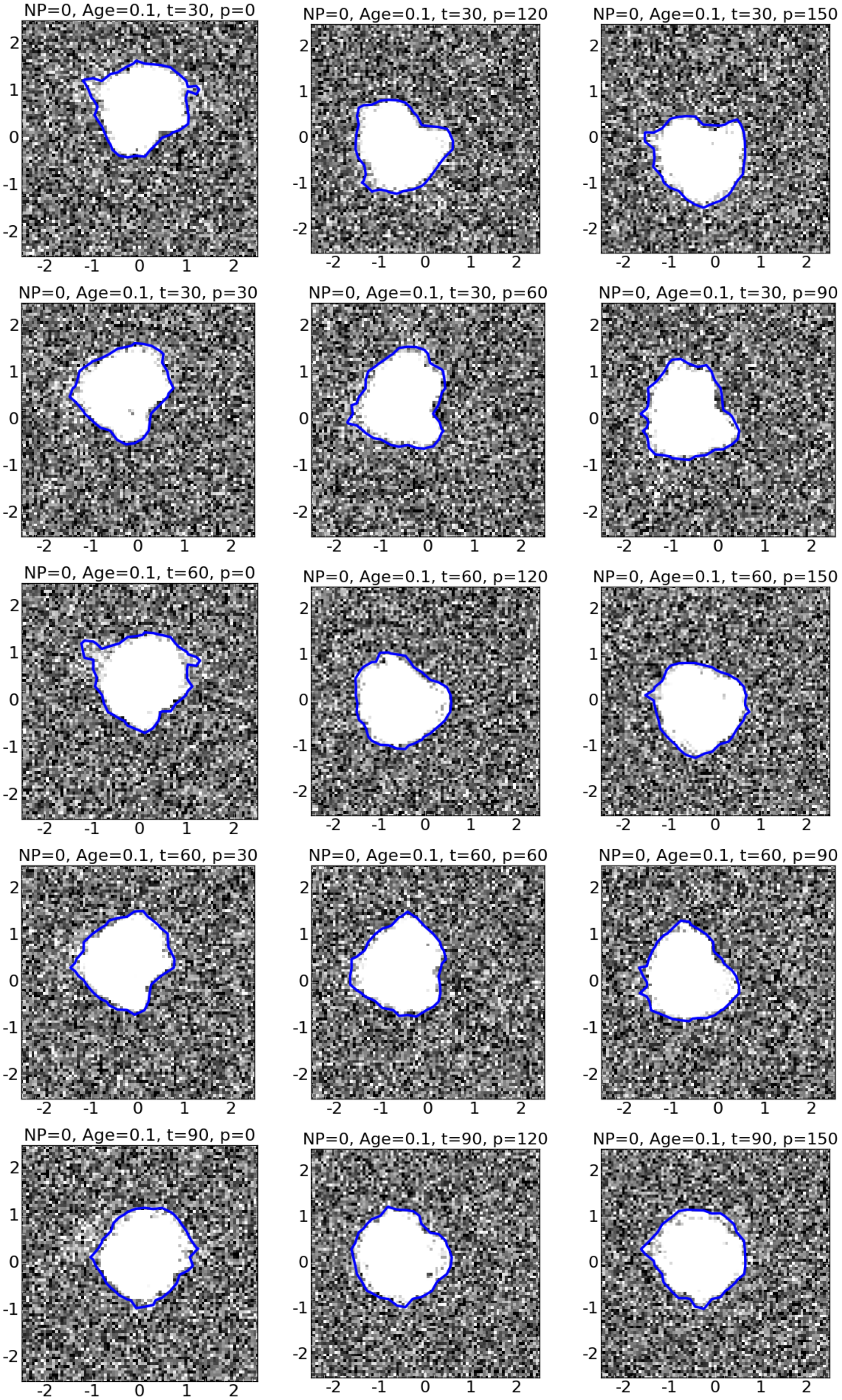
Continued on next page

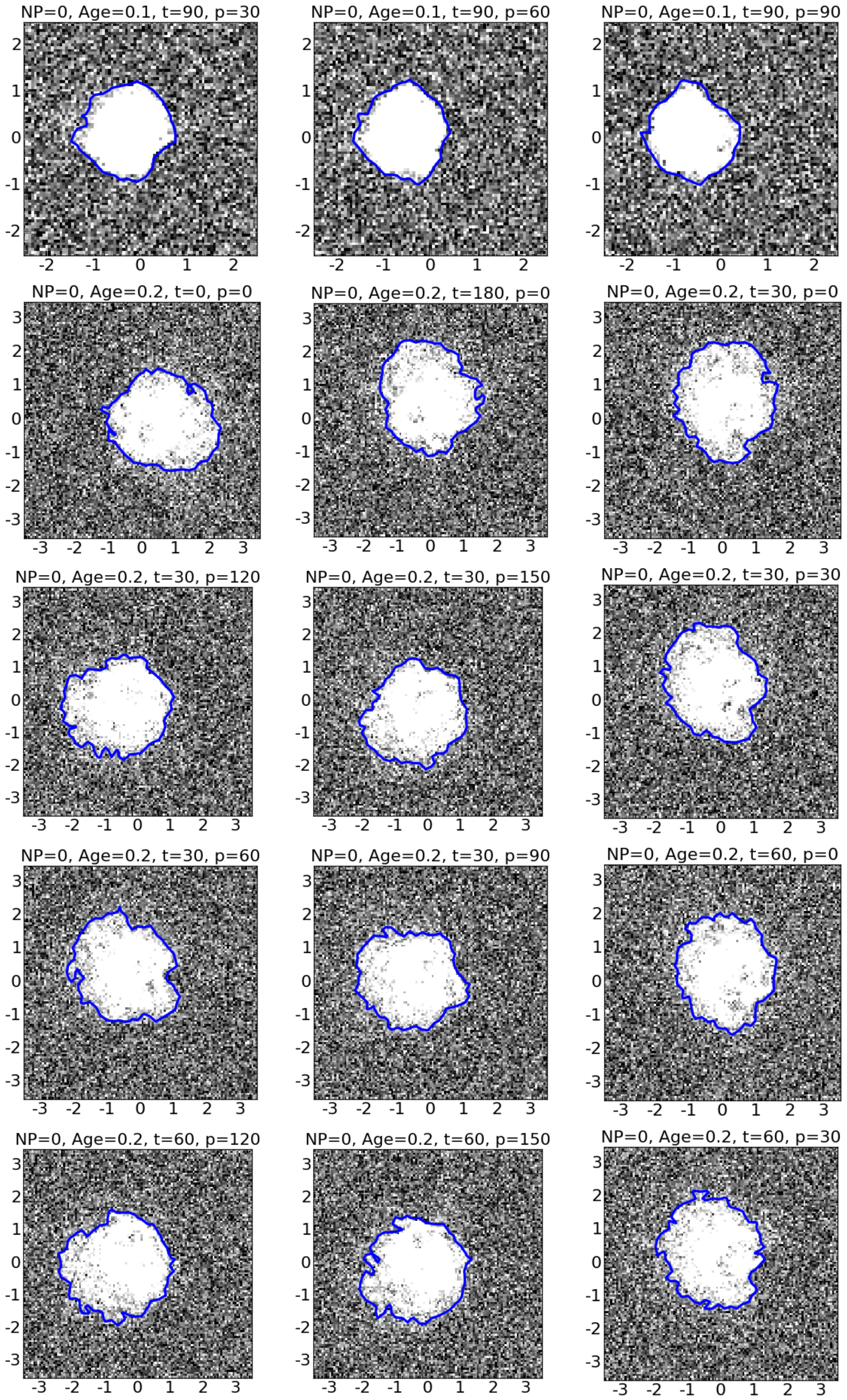
Group 6			
G045.391−00.725	 -0.670 -0.680 -0.690 -0.700 -0.710 -0.720 -0.730 -0.740 -0.750 -0.760 45.440 45.420 45.400 45.380 45.360	G046.495−00.241	 -0.140 -0.160 -0.180 -0.200 -0.220 -0.240 -0.260 -0.280 -0.300 -0.320 46.580 46.540 46.500 46.460 46.420 46.3

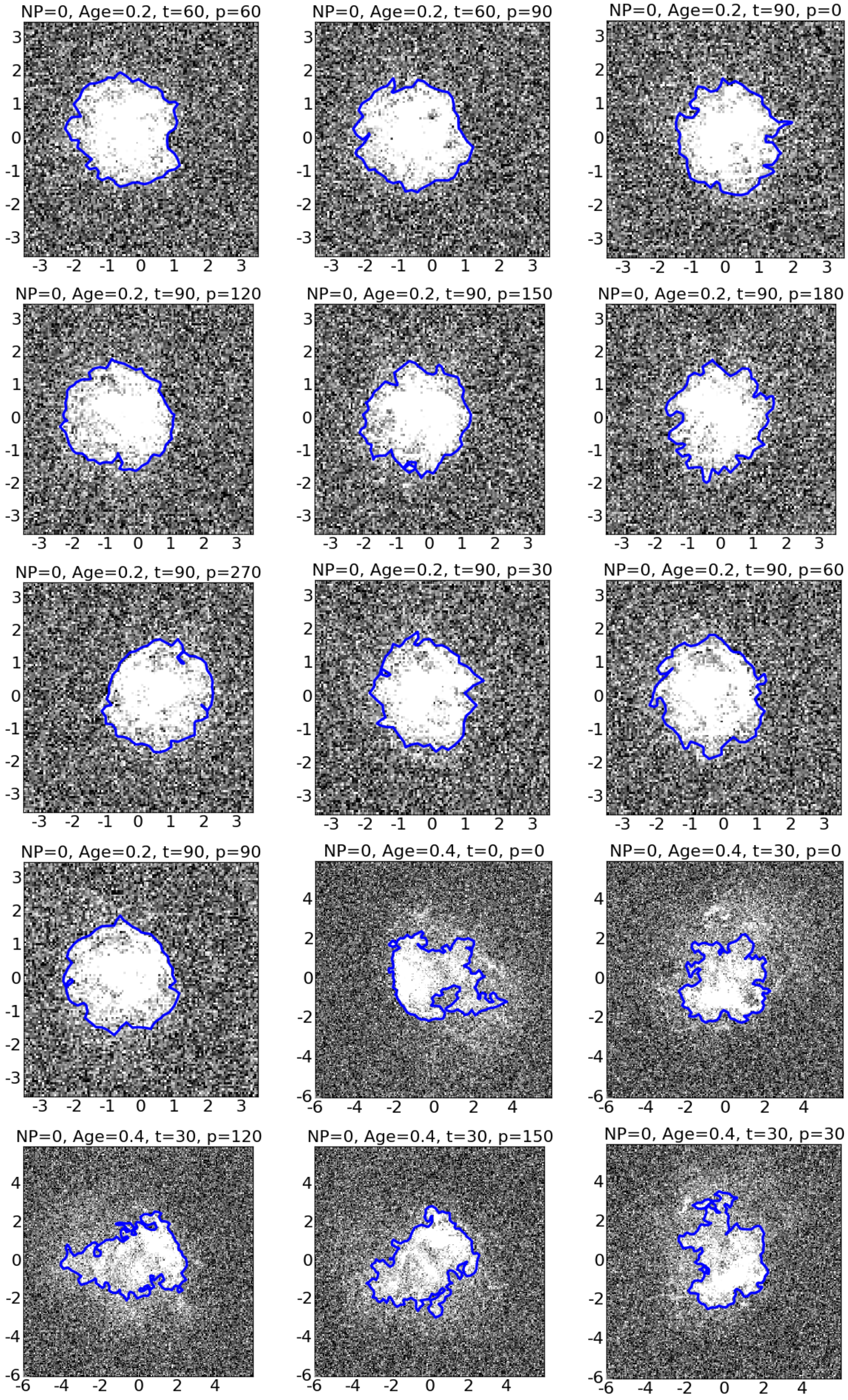
Appendix C

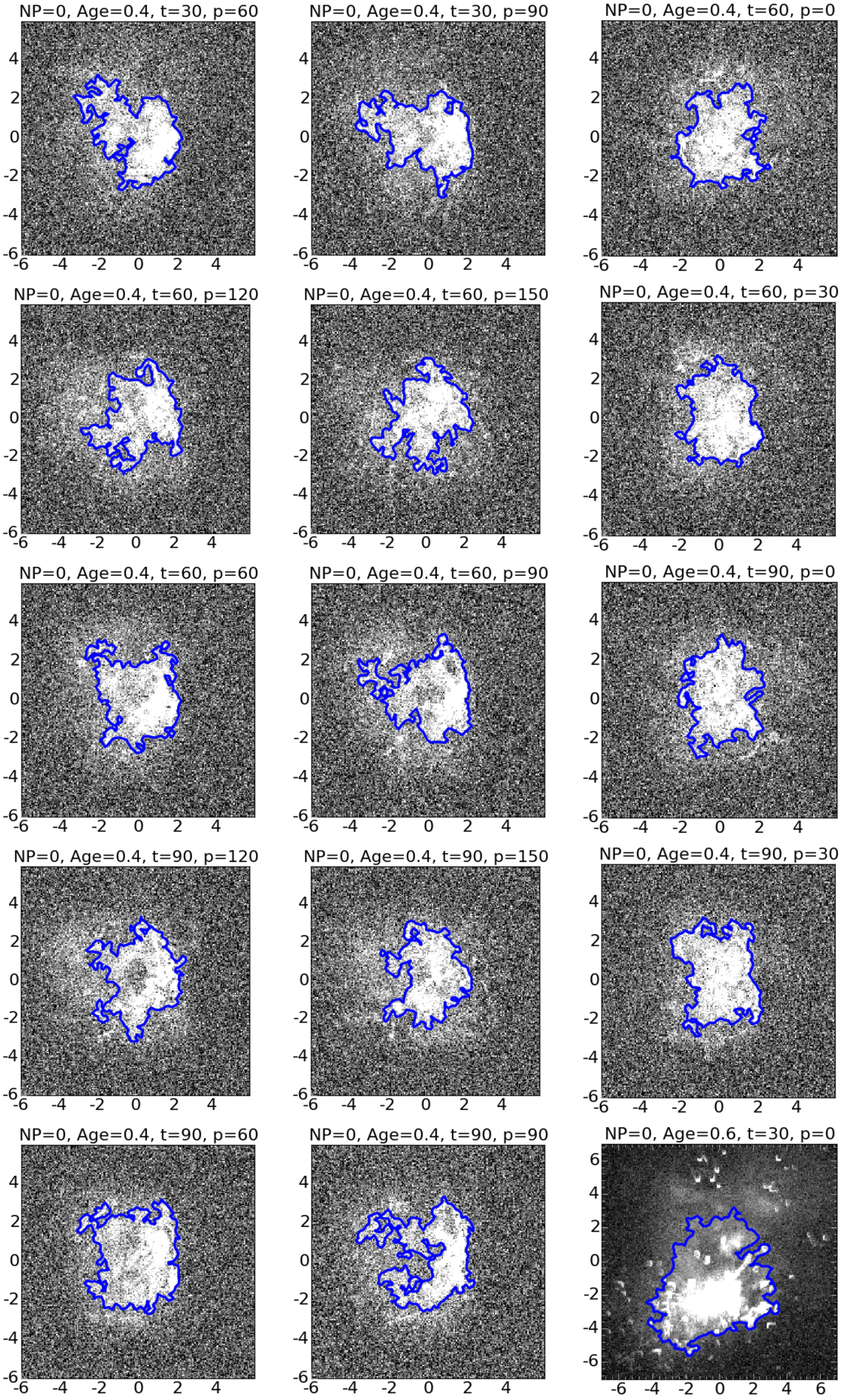
SO Images with Gaussian Noise

Synthetic Observations (SOs) of the numerical simulation in (Ali et al., 2018). The 77 images are shown with coordinates in parsecs. Headings identify the snapshot age and projection viewing angle (t $= \theta$, p $= \phi$). In each image, random Gaussian noise has been added to each pixel value, following the distribution from an example MAGPIS 1.4 GHz image tile. The boundary shown is that of the 1σ above the mean noise level contour.


NP=0, Age=0.1, t=30, p=0
NP=0, Age=0.1, t=30, p=120
NP=0, Age=0.1, t=30, p=150
NP=0, Age=0.1, t=30, p=30
NP=0, Age=0.1, t=30, p=60
NP=0, Age=0.1, t=30, p=90
NP=0, Age=0.1, t=60, p=0
NP=0, Age=0.1, t=60, p=120
NP=0, Age=0.1, t=60, p=150
NP=0, Age=0.1, t=60, p=30
NP=0, Age=0.1, t=60, p=60
NP=0, Age=0.1, t=60, p=90
NP=0, Age=0.1, t=90, p=0
NP=0, Age=0.1, t=90, p=120
NP=0, Age=0.1, t=90, p=150
2
1
0
-1
-2
-2
-1
0
1
2


NP=0, Age=0.1, t=90, p=30
NP=0, Age=0.1, t=90, p=60
NP=0, Age=0.1, t=90, p=90
NP=0, Age=0.2, t=0, p=0
NP=0, Age=0.2, t=180, p=0
NP=0, Age=0.2, t=30, p=0
NP=0, Age=0.2, t=30, p=120
NP=0, Age=0.2, t=30, p=150
NP=0, Age=0.2, t=30, p=30
NP=0, Age=0.2, t=30, p=60
NP=0, Age=0.2, t=30, p=90
NP=0, Age=0.2, t=60, p=0
NP=0, Age=0.2, t=60, p=120
NP=0, Age=0.2, t=60, p=150
NP=0, Age=0.2, t=60, p=30
-3
-2
-1
0
1
2
3


NP=0, Age=0.2, t=60, p=60
NP=0, Age=0.2, t=60, p=90
NP=0, Age=0.2, t=90, p=0
NP=0, Age=0.2, t=90, p=120
NP=0, Age=0.2, t=90, p=150
NP=0, Age=0.2, t=90, p=180
NP=0, Age=0.2, t=90, p=270
NP=0, Age=0.2, t=90, p=30
NP=0, Age=0.2, t=90, p=60
NP=0, Age=0.2, t=90, p=90
NP=0, Age=0.4, t=0, p=0
NP=0, Age=0.4, t=30, p=0
NP=0, Age=0.4, t=30, p=120
NP=0, Age=0.4, t=30, p=150
NP=0, Age=0.4, t=30, p=30
-3
-2
-1
0
1
2
3
-6
-4
-2
0
2
4



NP=0, Age=0.4, t=30, p=60

NP=0, Age=0.4, t=30, p=90

NP=0, Age=0.4, t=60, p=0

NP=0, Age=0.4, t=60, p=120

NP=0, Age=0.4, t=60, p=150

NP=0, Age=0.4, t=60, p=30

NP=0, Age=0.4, t=60, p=60

NP=0, Age=0.4, t=60, p=90

NP=0, Age=0.4, t=90, p=0

NP=0, Age=0.4, t=90, p=120

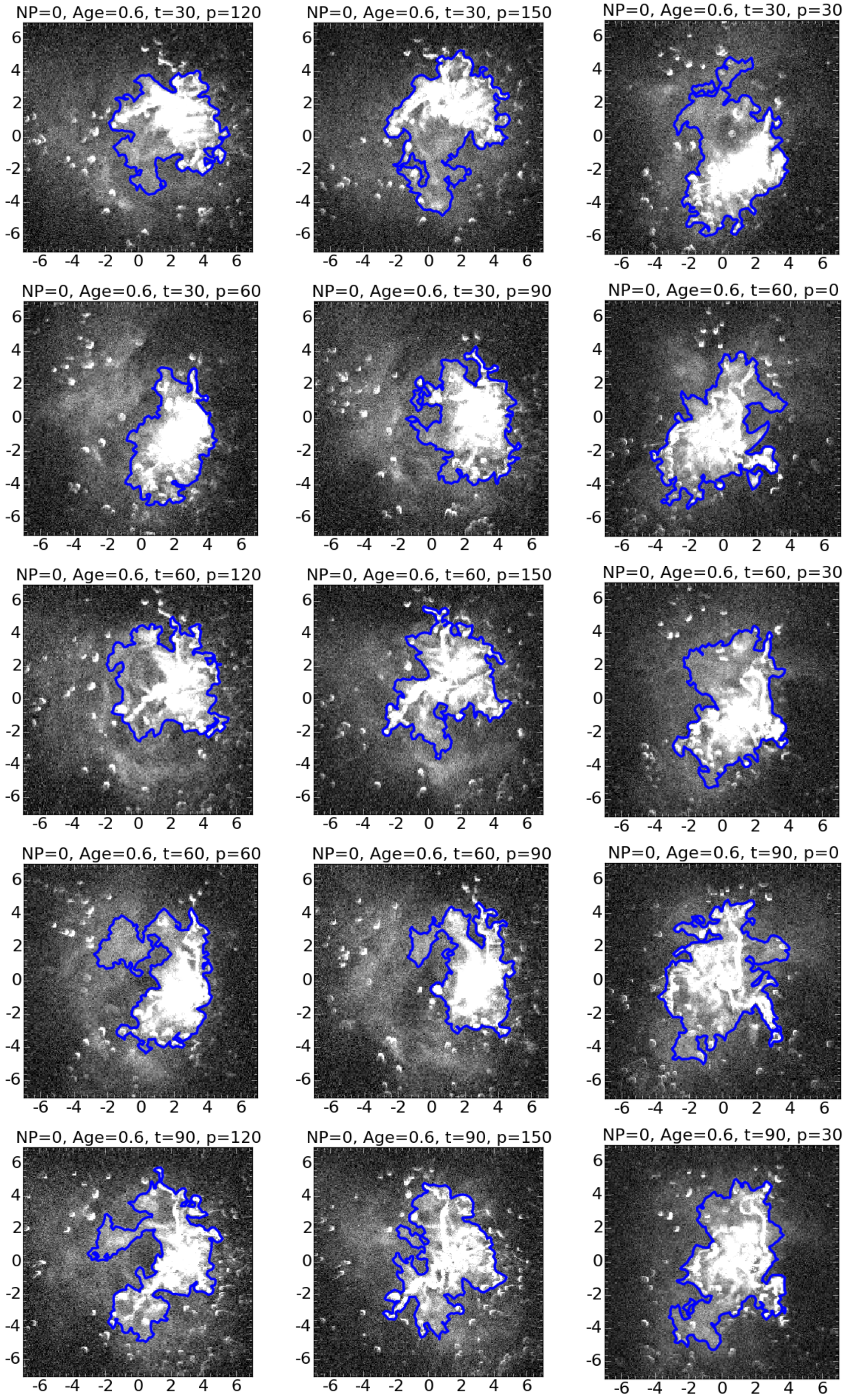
NP=0, Age=0.4, t=90, p=150

NP=0, Age=0.4, t=90, p=30

NP=0, Age=0.4, t=90, p=60

NP=0, Age=0.4, t=90, p=90

NP=0, Age=0.6, t=30, p=0


NP=0, Age=0.6, t=30, p=120

NP=0, Age=0.6, t=30, p=150

NP=0, Age=0.6, t=30, p=30

NP=0, Age=0.6, t=30, p=60

NP=0, Age=0.6, t=30, p=90

NP=0, Age=0.6, t=60, p=0

NP=0, Age=0.6, t=60, p=120

NP=0, Age=0.6, t=60, p=150

NP=0, Age=0.6, t=60, p=30

NP=0, Age=0.6, t=60, p=60

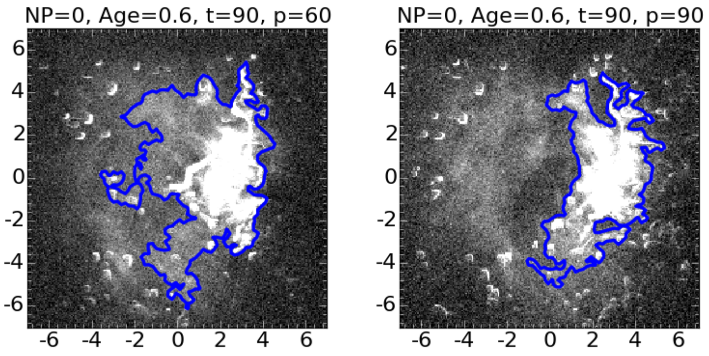
NP=0, Age=0.6, t=60, p=90

NP=0, Age=0.6, t=90, p=0

NP=0, Age=0.6, t=90, p=120

NP=0, Age=0.6, t=90, p=150

NP=0, Age=0.6, t=90, p=30


NP=0, Age=0.6, t=90, p=60
NP=0, Age=0.6, t=90, p=90
6
4
2
0
-2
-4
-6
-6
-4
-2
0
2
4
6

Bibliography

F. C. Adams, C. J. Lada, and F. H. Shu. Spectral Evolution of Young Stellar Objects. *ApJ*, 312:788, Jan 1987. doi:10.1086/164924.

A. Ali, T. J. Harries, and T. A. Douglas. Modelling massive star feedback with Monte Carlo radiation hydrodynamics: photoionization and radiation pressure in a turbulent cloud. *MNRAS*, 477:5422–5436, Jul 2018. doi:10.1093/mnras/sty1001.

L. D. Anderson and T. M. Bania. Resolution of the Distance Ambiguity for Galactic H II Regions. *ApJ*, 690:706–719, Jan. 2009. doi:10.1088/0004-637X/690/1/706.

L. D. Anderson, T. M. Bania, D. S. Balser, and R. T. Rood. The Green Bank Telescope H II Region Discovery Survey. II. The Source Catalog. *ApJS*, 194:32, June 2011. doi:10.1088/0067-0049/194/2/32.

L. D. Anderson, T. M. Bania, D. S. Balser, and R. T. Rood. The Green Bank Telescope H II Region Discovery Survey. III. Kinematic Distances. *ApJ*, 754:62, July 2012. doi:10.1088/0004-637X/754/1/62.

L. D. Anderson, T. M. Bania, D. S. Balser, V. Cunningham, T. V. Wenger, B. M. Johnstone, and W. P. Armentrout. The WISE Catalog of Galactic H II Regions. *ApJS*, 212:1, May 2014. doi:10.1088/0067-0049/212/1/1.

L. D. Anderson, T. V. Wenger, W. P. Armentrout, D. S. Balser, and T. M. Bania. A Galactic Plane Defined by the Milky Way H II Region Distribution. *ApJ*, 871:145, Feb 2019. doi:10.3847/1538-4357/aaf571.

T. W. Anderson and D. A. Darling. Asymptotic theory of certain criteria based on stochastic processes. *Ann. Math. Statist.*, 23(2):193–212, 06 1952. doi:10.1214/aoms/1177729437. URL `http://dx.doi.org/10.1214/aoms/1177729437`.

P. André, J. Di Francesco, D. Ward-Thompson, S.-I. Inutsuka, R. E. Pudritz, and J. E. Pineda. From Filamentary Networks to Dense Cores in Molecular Clouds: Toward a New Paradigm for Star Formation. *Protostars and Planets VI*, pages 27–51, 2014. doi:10.2458/azu_uapress_9780816531240-ch002.

H. G. Arce, M. A. Borkin, A. A. Goodman, J. E. Pineda, and C. N. Beaumont. A Bubbling Nearby Molecular Cloud: COMPLETE Shells in Perseus. *ApJ*, 742:105, Dec. 2011. doi:10.1088/0004-637X/742/2/105.

G. J. Babu and E. D. Feigelson. Astrostatistics: Goodness-of-Fit and All That! In C. Gabriel, C. Arviset, D. Ponz, and S. Enrique, editors, *Astronomical Data Analysis Software and Systems XV*, volume 351 of *Astronomical Society of the Pacific Conference Series*, page 127, July 2006.

I. K. Baldry. Hubble's galaxy nomenclature. *Astronomy and Geophysics*, 49(5): 5.25–5.26, Oct. 2008. doi:10.1111/j.1468-4004.2008.49525.x.

T. M. Bania, L. D. Anderson, and D. S. Balser. The Arecibo H II Region Discovery Survey. *ApJ*, 759:96, Nov. 2012. doi:10.1088/0004-637X/759/2/96.

M. R. Bate. Stellar, brown dwarf and multiple star properties from hydrodynamical simulations of star cluster formation. *MNRAS*, 392:590–616, Jan 2009a. doi:10.1111/j.1365-2966.2008.14106.x.

M. R. Bate. The dependence of star formation on initial conditions and molecular cloud structure. *MNRAS*, 397:232–248, Jul 2009b. doi:10.1111/j.1365-2966.2009.14970.x.

C. N. Beaumont and J. P. Williams. Molecular Rings Around Interstellar Bubbles and the Thickness of Star-Forming Clouds. *ApJ*, 709:791–800, Feb. 2010. doi:10.1088/0004-637X/709/2/791.

R. H. Becker, R. L. White, D. J. Helfand, and S. Zoonematkermani. A 5 GHz VLA Survey of the Galactic Plane. *The Astrophysical Journal Supplement Series*, 91: 347, Mar 1994. doi:10.1086/191941.

E. F. Bell, C. Wolf, K. Meisenheimer, H.-W. Rix, A. Borch, S. Dye, M. Kleinheinrich, L. Wisotzki, and D. H. McIntosh. Nearly 5000 Distant Early-Type Galaxies in COMBO-17: A Red Sequence and Its Evolution since z˜1. *ApJ*, 608:752–767, June 2004. doi:10.1086/420778.

M. Beltrametti, G. Tenorio-Tagle, and H. W. Yorke. The gas dynamics around OB associations. I - Recombining H II regions and the formation of expanding neutral shells. *A&A*, 112:1–10, Aug. 1982.

R. A. Benjamin, E. Churchwell, B. L. Babler, T. M. Bania, D. P. Clemens, M. Cohen, J. M. Dickey, R. Indebetouw, J. M. Jackson, H. A. Kobulnicky, A. Lazarian, A. P. Marston, J. S. Mathis, M. R. Meade, S. Seager, S. R. Stolovy, C. Watson, B. A. Whitney, M. J. Wolff, and M. G. Wolfire. GLIMPSE. I. An SIRTF Legacy Project to Map the Inner Galaxy. *PASP*, 115:953–964, Aug. 2003. doi:10.1086/376696.

E. A. Bergin and M. Tafalla. Cold Dark Clouds: The Initial Conditions for Star Formation. *Annual Review of Astronomy and Astrophysics*, 45:339–396, Sep 2007. doi:10.1146/annurev.astro.45.071206.100404.

O. Berné, A. Fuente, J. R. Goicoechea, P. Pilleri, M. González-García, and C. Joblin. Mid-Infrared Polycyclic Aromatic Hydrocarbon and H_2 Emission as a Probe of Physical Conditions in Extreme Photodissociation Regions. *ApJ*, 706:L160–L163, Nov. 2009. doi:10.1088/0004-637X/706/1/L160.

F. Bertoldi. The Photoevaporation of Interstellar Clouds. I. Radiation-driven Implosion. *ApJ*, 346:735, Nov 1989. doi:10.1086/168055.

H. Beuther, S. Bihr, M. Rugel, K. Johnston, Y. Wang, F. Walter, A. Brunthaler, A. J. Walsh, J. Ott, J. Stil, T. Henning, T. Schierhuber, J. Kainulainen, M. Heyer, P. F. Goldsmith, L. D. Anderson, S. N. Longmore, R. S. Klessen, S. C. O. Glover, J. S. Urquhart, R. Plume, S. E. Ragan, N. Schneider, N. M. McClure-Griffiths, K. M. Menten, R. Smith, N. Roy, R. Shanahan, Q. Nguyen-Luong, and F. Bigiel. The HI/OH/Recombination line survey of the inner Milky Way (THOR). Survey overview and data release 1. *A&A*, 595:A32, Oct 2016. doi:10.1051/0004-6361/201629143.

T. G. Bisbas, R. Wünsch, A. P. Whitworth, and D. A. Hubber. Smoothed particle hydrodynamics simulations of expanding H II regions. I. Numerical method and applications. *A&A*, 497:649–659, Apr. 2009. doi:10.1051/0004-6361/200811522.

L. Blitz. Giant Molecular Clouds. In E. H. Levy and J. I. Lunine, editors, *Protostars and Planets III*, page 125, Jan 1993.

L. Blitz and F. H. Shu. The origin and lifetime of giant molecular cloud complexes. *ApJ*, 238:148–157, May 1980. doi:10.1086/157968.

I. A. Bonnell, M. R. Bate, and H. Zinnecker. On the formation of massive stars. *MNRAS*, 298:93–102, Jul 1998. doi:10.1046/j.1365-8711.1998.01590.x.

F. L. Bookstein. Size and shape spaces for landmark data in two dimensions. *Statist. Sci.*, 1(2):181–222, 05 1986. doi:10.1214/ss/1177013696. URL `https://doi.org/10.1214/ss/1177013696`.

J. Brand, L. Blitz, and J. G. A. Wouterloot. The velocity field of the outer Galaxy in the southern hemisphere. I.Catalogue of nebulous objects. *Astronomy and Astrophysics Supplement Series*, 65:537–550, Sep 1986.

A. G. W. Cameron and J. W. Truran. The Supernova Trigger for Formation of the Solar System. *Icarus*, 30:447–461, Mar 1977. doi:10.1016/0019-1035(77)90101-4.

J. Campbell-White, D. Froebrich, and A. Kume. Shape analysis of H II regions - I. Statistical clustering. *MNRAS*, 477:5486–5500, Jul 2018. doi:10.1093/mnras/sty954.

S. J. Carey, A. Noriega-Crespo, D. R. Mizuno, S. Shenoy, R. Paladini, K. E. Kraemer, S. D. Price, N. Flagey, E. Ryan, J. G. Ingalls, T. A. Kuchar, D. Pinheiro Gonçalves, R. Indebetouw, N. Billot, F. R. Marleau, D. L. Padgett, L. M. Rebull, E. Bressert, B. Ali, S. Molinari, P. G. Martin, G. B. Berriman, F. Boulanger, W. B. Latter, M. A. Miville-Deschenes, R. Shipman, and L. Testi. MIPSGAL: A Survey of the Inner Galactic Plane at 24 and 70 μm. *PASP*, 121:76–97, Jan. 2009. doi:10.1086/596581.

J. Castor, R. McCray, and R. Weaver. Interstellar bubbles. *ApJ*, 200:L107–L110, Sept. 1975. doi:10.1086/181908.

R. Chini, E. Kruegel, and W. Wargau. Dust emission and star formation in compact H II regions. *A&A*, 181:378–382, July 1987.

E. Churchwell, M. S. Povich, D. Allen, M. G. Taylor, M. R. Meade, B. L. Babler, R. Indebetouw, C. Watson, B. A. Whitney, M. G. Wolfire, T. M. Bania, R. A. Benjamin, D. P. Clemens, M. Cohen, C. J. Cyganowski, J. M. Jackson, H. A. Kobulnicky, J. S. Mathis, E. P. Mercer, S. R. Stolovy, B. Uzpen, D. F. Watson, and M. J. Wolff. The Bubbling Galactic Disk. *ApJ*, 649:759–778, Oct. 2006. doi:10.1086/507015.

E. Churchwell, D. F. Watson, M. S. Povich, M. G. Taylor, B. L. Babler, M. R. Meade, R. A. Benjamin, R. Indebetouw, and B. A. Whitney. The Bubbling Galactic Disk. II. The Inner 20deg. *ApJ*, 670:428–441, Nov. 2007. doi:10.1086/521646.

J. J. Condon, W. D. Cotton, E. W. Greisen, Q. F. Yin, R. A. Perley, G. B. Taylor, and J. J. Broderick. The NRAO VLA Sky Survey. *AJ*, 115:1693–1716, May 1998. doi:10.1086/300337.

P. S. Conti and E. M. Leep. Spectroscopic observations of O-type stars. V. The hydrogen lines and lambda 4686 He II. *ApJ*, 193:113–124, Oct. 1974. doi:10.1086/153135.

H. Cramér. On the composition of elementary errors. *Scandinavian Actuarial Journal*, 1928(1):13–74, 1928. doi:10.1080/03461238.1928.10416862. URL `https://doi.org/10.1080/03461238.1928.10416862`.

J. E. Dale, I. A. Bonnell, and A. P. Whitworth. Ionization-induced star formation - I. The collect-and-collapse model. *MNRAS*, 375:1291–1298, Mar 2007a. doi:10.1111/j.1365-2966.2006.11368.x.

J. E. Dale, P. C. Clark, and I. A. Bonnell. Ionization-induced star formation - II. External irradiation of a turbulent molecular cloud. *MNRAS*, 377:535–544, May 2007b. doi:10.1111/j.1365-2966.2007.11515.x.

L. Deharveng, A. Zavagno, and J. Caplan. Triggered massive-star formation on the borders of Galactic H II regions. I. A search for "collect and collapse" candidates. *A&A*, 433:565–577, Apr 2005. doi:10.1051/0004-6361:20041946.

L. Deharveng, F. Schuller, L. D. Anderson, A. Zavagno, F. Wyrowski, K. M. Menten, L. Bronfman, L. Testi, C. M. Walmsley, and M. Wienen. A gallery of bubbles. The nature of the bubbles observed by Spitzer and what ATLASGAL tells us about the surrounding neutral material. *A&A*, 523:A6, Nov. 2010. doi:10.1051/0004-6361/201014422.

L. K. Dewangan, D. K. Ojha, I. Zinchenko, P. Janardhan, and A. Luna. Multiwavelength Study of the Star Formation in the S237 H ii Region. *ApJ*, 834:22, Jan. 2017. doi:10.3847/1538-4357/834/1/22.

Y. Dodge. *Kolmogorov–Smirnov Test*. Springer New York, New York, NY, 2008. ISBN 978-0-387-32833-1. doi:10.1007/978-0-387-32833-1_214. URL `https://doi.org/10.1007/978-0-387-32833-1_214`.

Y. Doi, S. Takita, T. Ootsubo, K. Arimatsu, M. Tanaka, Y. Kitamura, M. Kawada, S. Matsuura, T. Nakagawa, T. Morishima, M. Hattori, S. Komugi, G. J. White, N. Ikeda, D. Kato, Y. Chinone, M. Etxaluze, and E. F. Cypriano. The AKARI far-infrared all-sky survey maps. *PASJ*, 67:50, June 2015. doi:10.1093/pasj/psv022.

I. L. Dryden and K. V. Mardia. *Statistical Shape Analysis*. Wiley, Chichester, 1998.

A. Dumitrescu, A. Pop, and D. Dumitrescu. Structural Properties of Pulsating Star Light Curves Through Fuzzy Divisive Hierarchical Clustering. *Ap&SS*, 250: 205–226, Jan 1997. doi:10.1023/A:1000414722069.

E. Dwek, R. G. Arendt, D. J. Fixsen, T. J. Sodroski, N. Odegard, J. L. Weiland, W. T. Reach, M. G. Hauser, T. Kelsall, S. H. Moseley, R. F. Silverberg, R. A. Shafer, J. Ballester, D. Bazell, and R. Isaacman. Detection and Characterization of Cold Interstellar Dust and Polycyclic Aromatic Hydrocarbon Emission, from COBE Observations. *ApJ*, 475:565–579, Feb 1997. doi:10.1086/303568.

J. E. Dyson and D. A. Williams. *Physics of the interstellar medium*. Manchester University Press Manchester, 1980. ISBN 0719007976.

A. S. Eddington. On the radiative equilibrium of the stars. *MNRAS*, 77:16–35, Nov 1916. doi:10.1093/mnras/77.1.16.

B. G. Elmegreen. Triggering the Formation of Young Clusters. In D. P. Geisler, E. K. Grebel, and D. Minniti, editors, *Extragalactic Star Clusters*, volume 207 of *IAU Symposium*, page 390, Jan 2002.

S. Engmann and D. Cousineau. Comparing distributions: the two-sample anderson-darling test as an alternative to the kolmogorov-smirnoff test. *Journal of Applied Quantitative Methods*, 6(3):1–17, 2011.

B. S. Everitt, G. Dunn, et al. *Applied multivariate data analysis*, volume 2. Arnold London, 2001.

G. G. Fazio, J. L. Hora, L. E. Allen, M. L. N. Ashby, P. Barmby, L. K. Deutsch, J.-S. Huang, S. Kleiner, M. Marengo, S. T. Megeath, G. J. Melnick, M. A. Pahre, B. M. Patten, J. Polizotti, H. A. Smith, R. S. Taylor, Z. Wang, S. P. Willner, W. F. Hoffmann, J. L. Pipher, W. J. Forrest, C. W. McMurty, C. R. McCreight, M. E. McKelvey, R. E. McMurray, D. G. Koch, S. H. Moseley, R. G. Arendt, J. E. Mentzell, C. T. Marx, P. Losch, P. Mayman, W. Eichhorn, D. Krebs, M. Jhabvala, D. Y. Gezari, D. J. Fixsen, J. Flores, K. Shakoorzadeh, R. Jungo, C. Hakun, L. Workman, G. Karpati, R. Kichak, R. Whitley, S. Mann, E. V. Tollestrup, P. Eisenhardt, D. Stern, V. Gorjian, B. Bhattacharya, S. Carey, B. O. Nelson, W. J. Glaccum, M. Lacy, P. J. Lowrance, S. Laine, W. T. Reach, J. A. Stauffer, J. A. Surace, G. Wilson, E. L. Wright, A. Hoffman, G. Domingo, and M. Cohen. The infrared array camera (irac) for the spitzer space telescope. *The Astrophysical Journal Supplement Series*, 154(1):10, 2004. URL `http://stacks.iop.org/0067-0049/154/i=1/a=10`.

E. D. Feigelson and G. J. Babu. *Modern Statistical Methods for Astronomy*. July 2012.

L. Ferreira and D. B. Hitchcock. A comparison of hierarchical methods for clustering functional data. *Communications in Statistics-Simulation and Computation*, 38 (9):1925–1949, 2009.

K. M. Ferrière. The interstellar environment of our galaxy. *Reviews of Modern Physics*, 73:1031–1066, Oct 2001. doi:10.1103/RevModPhys.73.1031.

A. V. Filippenko. Optical Spectra of Supernovae. *Annual Review of Astronomy and Astrophysics*, 35:309–355, Jan 1997. doi:10.1146/annurev.astro.35.1.309.

K. Florek, J. Łukaszewicz, J. Perkal, H. Steinhaus, and S. Zubrzycki. Sur la liaison et la division des points d'un ensemble fini. In *Colloquium Mathematicae*, volume 2, pages 282–285, 1951.

G. E. Forsythe, C. B. Moler, and M. A. Malcolm. Computer methods for mathematical computations. 1977.

P. N. Foster and A. P. Boss. Triggering Star Formation with Stellar Ejecta. *ApJ*, 468:784, Sep 1996. doi:10.1086/177735.

J. Franco, S. E. Kurtz, G. García-Segura, and P. Hofner. The Evolution of HII Regions. *Ap&SS*, 272:169–179, 2000. doi:10.1023/A:1002680025946.

R. Garcia-Dias, C. Allende Prieto, J. Sánchez Almeida, and I. Ordovás-Pascual. Machine learning in APOGEE. Unsupervised spectral classification with K-means. *A&A*, 612:A98, May 2018. doi:10.1051/0004-6361/201732134.

S. Geen, J. D. Soler, and P. Hennebelle. Interpreting the star formation efficiency of nearby molecular clouds with ionizing radiation. *MNRAS*, 471:4844–4855, Nov 2017. doi:10.1093/mnras/stx1765.

Y. M. Georgelin and Y. P. Georgelin. The spiral structure of our Galaxy determined from H II regions. *A&A*, 49:57–79, May 1976.

U. Giveon, R. H. Becker, D. J. Helfand, and R. L. White. A New Catalog of Radio Compact H II Regions in the Milky Way. *AJ*, 129:348–354, Jan. 2005. doi:10.1086/426360.

P. F. Goldsmith. Molecular Clouds: An Overview. In D. J. Hollenbach and J. Thronson, Harley A., editors, *Interstellar Processes*, volume 134, page 51, Jan 1987. doi:10.1007/978-94-009-3861-8_3.

J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. ISSN 1860-0980. doi:10.1007/BF02291478. URL `http://dx.doi.org/10.1007/BF02291478`.

M. Gritschneder, D. N. C. Lin, S. D. Murray, Q. Z. Yin, and M. N. Gong. The Supernova Triggered Formation and Enrichment of Our Solar System. *ApJ*, 745: 22, Jan 2012. doi:10.1088/0004-637X/745/1/22.

L. Haemmerlé, P. Eggenberger, G. Meynet, A. Maeder, and C. Charbonnel. Massive star formation by accretion. I. Disc accretion. *A&A*, 585:A65, Jan 2016. doi:10.1051/0004-6361/201527202.

T. J. Harries. Radiation-hydrodynamical simulations of massive star formation using Monte Carlo radiative transfer - I. Algorithms and numerical methods. *MNRAS*, 448:3156–3166, Apr. 2015. doi:10.1093/mnras/stv158.

T. I. Hasegawa and E. Herbst. New gas-grain chemical models of quiscent dense interstellar clouds :the effects of H2 tunnelling reactions and cosmic ray induced desorption. *MNRAS*, 261:83–102, Mar 1993. doi:10.1093/mnras/261.1.83.

D. J. Helfand, R. H. Becker, R. L. White, A. Fallon, and S. Tuttle. MAGPIS: A Multi-Array Galactic Plane Imaging Survey. *AJ*, 131:2525–2537, May 2006. doi:10.1086/503253.

G. H. Herbig. FU Orionis eruptions. In *European Southern Observatory Conference and Workshop Proceedings*, volume 33, pages 233–246, Sep 1989.

M. G. Hoare, S. L. Lumsden, R. D. Oudmaijer, J. S. Urquhart, A. L. Busfield, T. L. Sheret, A. J. Clarke, T. J. T. Moore, J. Allsopp, M. G. Burton, C. R. Purcell, Z. Jiang, and M. Wang. The RMS survey: Massive young stars throughout the galaxy. In R. Cesaroni, M. Felli, E. Churchwell, and M. Walmsley, editors, *Massive Star Birth: A Crossroads of Astrophysics*, volume 227 of *IAU Symposium*, pages 370–375, Jan 2005. doi:10.1017/S174392130500476X.

M. G. Hoare, C. R. Purcell, E. B. Churchwell, P. Diamond, W. D. Cotton, C. J. Chandler, S. Smethurst, S. E. Kurtz, L. G. Mundy, S. M. Dougherty, R. P. Fender, G. A. Fuller, J. M. Jackson, S. T. Garrington, T. R. Gledhill, P. F. Goldsmith, S. L. Lumsden, J. Martí, T. J. T. Moore, T. W. B. Muxlow, R. D.

Oudmaijer, J. D. Pand ian, J. M. Paredes, D. S. Shepherd, R. E. Spencer, M. A. Thompson, G. Umana, J. S. Urquhart, and A. A. Zijlstra. The Coordinated Radio and Infrared Survey for High-Mass Star Formation (The CORNISH Survey). I. Survey Design. *Publications of the Astronomical Society of the Pacific*, 124:939, Sep 2012. doi:10.1086/668058.

A. Hou, L. C. Parker, W. E. Harris, and D. J. Wilman. Statistical Tools for Classifying Galaxy Group Dynamics. *ApJ*, 702:1199–1210, Sept. 2009. doi:10.1088/0004-637X/702/2/1199.

L. G. Hou and X. Y. Gao. A statistical study of gaseous environment of Spitzer interstellar bubbles. *MNRAS*, 438:426–437, Feb. 2014. doi:10.1093/mnras/stt2212.

E. Hubble. A Relation between Distance and Radial Velocity among Extra-Galactic Nebulae. *Proceedings of the National Academy of Science*, 15:168–173, Mar 1929. doi:10.1073/pnas.15.3.168.

E. Hubble. *The realm of the nebulae*, volume 25. Yale University Press, 1982.

E. P. Hubble. The classification of spiral nebulae. *The Observatory*, 50:276–281, Sept. 1927.

W.-G. Ji, J.-J. Zhou, J. Esimbek, Y.-F. Wu, G. Wu, and X.-D. Tang. The infrared dust bubble N22: an expanding H ii region and the star formation around it. *A&A*, 544:A39, Aug. 2012. doi:10.1051/0004-6361/201218861.

C. Jones and J. M. Dickey. Kinematic Distance Assignments with H I Absorption. *ApJ*, 753:62, July 2012. doi:10.1088/0004-637X/753/1/62.

L. Kaufman and P. J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

D. G. Kendall. The diffusion of shape. *Advances in Applied Probability*, 9(3):428–430, 1977. ISSN 00018678. URL `http://www.jstor.org/stable/1426091`.

S. Kendrew, R. Simpson, E. Bressert, M. S. Povich, R. Sherman, C. J. Lintott, T. P. Robitaille, K. Schawinski, and G. Wolf-Chase. The Milky Way Project: A Statistical Study of Massive Star Formation Associated with Infrared Bubbles. *ApJ*, 755:71, Aug. 2012. doi:10.1088/0004-637X/755/1/71.

F. J. Kerr. The Large-Scale Distribution of Hydrogen in the Galaxy. *Annual Review of Astronomy and Astrophysics*, 7:39, Jan 1969. doi:10.1146/annurev.aa.07.090169.000351.

C. M. Koepferl and T. P. Robitaille. The FluxCompensator: Making Radiative Transfer Models of Hydrodynamical Simulations Directly Comparable to Real Observations. *ApJ*, 849:3, Nov. 2017. doi:10.3847/1538-4357/aa8666.

A. Kolmogorov. Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.*, 4:83–91, 1933.

P. Kroupa. On the variation of the initial mass function. *MNRAS*, 322:231–246, Apr. 2001. doi:10.1046/j.1365-8711.2001.04022.x.

M. R. Krumholz, R. I. Klein, and C. F. McKee. Radiation-Hydrodynamic Simulations of Collapse and Fragmentation in Massive Protostellar Cores. *ApJ*, 656: 959–979, Feb 2007. doi:10.1086/510664.

J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

G. P. Kuiper. The Empirical Mass-Luminosity Relation. *ApJ*, 88:472, Nov. 1938. doi:10.1086/143999.

R. Kuiper, H. Klahr, H. Beuther, and T. Henning. Circumventing the Radiation Pressure Barrier in the Formation of Massive Stars via Disk Accretion. *ApJ*, 722: 1556–1576, Oct 2010. doi:10.1088/0004-637X/722/2/1556.

C. J. Lada and E. A. Lada. Embedded Clusters in Molecular Clouds. *Annual Review of Astronomy and Astrophysics*, 41:57–115, Jan 2003. doi:10.1146/annurev.astro.41.011802.094844.

G. Lange and W. Williams. A general theory of classificatory sorting strategies. i. hierarchical systems. *Computer Journal*, 9:373–380, 1967.

R. B. Larson. The physics of star formation. *Reports on Progress in Physics*, 66: 1651–1697, Oct 2003. doi:10.1088/0034-4885/66/10/R03.

R. B. Larson and S. Starrfield. On the formation of massive stars and the upper limit of stellar masses. *A&A*, 13:190, Jul 1971.

G. Lemaître. Un Univers homogène de masse constante et de rayon croissant rendant compte de la vitesse radiale des nébuleuses extra-galactiques. *Annales de la Société Scientifique de Bruxelles*, 47:49–59, 1927.

C. J. Lintott, K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas, M. J. Raddick, R. C. Nichol, A. Szalay, D. Andreescu, P. Murray, and J. Vandenberg. Galaxy

Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *MNRAS*, 389:1179–1189, Sept. 2008. doi:10.1111/j.1365-2966.2008.13689.x.

F. J. Lockman. A survey of radio H II regions in the northern sky. *ApJS*, 71: 469–479, Nov. 1989. doi:10.1086/191383.

L. B. Lucy. Computing radiative equilibria with Monte Carlo techniques. *A&A*, 344:282–288, Apr 1999.

J. S. Mathis and G. Whiffen. Composite Interstellar Grains. *ApJ*, 341:808, Jun 1989. doi:10.1086/167538.

D. N. Matsakis, N. J. Evans, II, T. Sato, and B. Zuckerman. Radio continuum measurements of compact H II regions and other sources. *AJ*, 81:172–177, Mar. 1976. doi:10.1086/111871.

N. M. McClure-Griffiths, J. M. Dickey, B. M. Gaensler, A. J. Green, M. Haverkorn, and S. Strasser. The Southern Galactic Plane Survey: H I Observations and Analysis. *ApJS*, 158:178–187, June 2005. doi:10.1086/430114.

C. F. McKee and E. C. Ostriker. Theory of Star Formation. *Annual Review of Astronomy and Astrophysics*, 45:565–687, Sep 2007. doi:10.1146/annurev.astro.45.051806.110602.

C. F. McKee and J. C. Tan. Massive star formation in 100,000 years from turbulent and pressurized molecular clouds. *Nature*, 416:59–61, Mar 2002. doi:10.1038/416059a.

P. G. Mezger and A. P. Henderson. Galactic H II Regions. I. Observations of Their Continuum Radiation at the Frequency 5 GHz. *ApJ*, 147:471, Feb. 1967. doi:10.1086/149030.

D. Mihalas and J. Binney. *Galactic astronomy. Structure and kinematics.* 1981.

S. Molinari, B. Swinyard, J. Bally, M. Barlow, J. P. Bernard, P. Martin, T. Moore, A. Noriega-Crespo, R. Plume, L. Testi, A. Zavagno, A. Abergel, B. Ali, P. André, J. P. Baluteau, M. Benedettini, O. Berné, N. P. Billot, J. Blommaert, S. Bontemps, F. Boulanger, J. Brand, C. Brunt, M. Burton, L. Campeggio, S. Carey, P. Caselli, R. Cesaroni, J. Cernicharo, S. Chakrabarti, A. Chrysostomou, C. Codella, M. Cohen, M. Compiegne, C. J. Davis, P. de Bernardis, G. de Gasperis, J. Di Francesco, A. M. di Giorgio, D. Elia, F. Faustini, J. F. Fischera, Y. Fukui, G. A. Fuller,

K. Ganga, P. Garcia-Lario, M. Giard, G. Giardino, J. Glenn, P. Goldsmith, M. Griffin, M. Hoare, M. Huang, B. Jiang, C. Joblin, G. Joncas, M. Juvela, J. Kirk, G. Lagache, J. Z. Li, T. L. Lim, S. D. Lord, P. W. Lucas, B. Maiolo, M. Marengo, D. Marshall, S. Masi, F. Massi, M. Matsuura, C. Meny, V. Minier, M. A. Miville-Deschênes, L. Montier, F. Motte, T. G. Müller, P. Natoli, J. Neves, L. Olmi, R. Paladini, D. Paradis, M. Pestalozzi, S. Pezzuto, F. Piacentini, M. Pomarès, C. C. Popescu, W. T. Reach, J. Richer, I. Ristorcelli, A. Roy, P. Royer, D. Russeil, P. Saraceno, M. Sauvage, P. Schilke, N. Schneider-Bontemps, F. Schuller, B. Schultz, D. S. Shepherd, B. Sibthorpe, H. A. Smith, M. D. Smith, L. Spinoglio, D. Stamatellos, F. Strafella, G. Stringfellow, E. Sturm, R. Taylor, M. A. Thompson, R. J. Tuffs, G. Umana, L. Valenziano, R. Vavrek, S. Viti, C. Waelkens, D. Ward-Thompson, G. White, F. Wyrowski, H. W. Yorke, and Q. Zhang. Hi-GAL: The Herschel Infrared Galactic Plane Survey. *Publications of the Astronomical Society of the Pacific*, 122:314, Mar 2010. doi:10.1086/651314.

D. C. Morton. Mass Loss from Three OB Supergiants in Orion. *ApJ*, 150:535, Nov. 1967. doi:10.1086/149356.

T. Murphy, T. Mauch, A. Green, R. W. Hunstead, B. Piestrzynska, A. P. Kels, and P. Sztajer. The second epoch Molonglo Galactic Plane Survey: compact source catalogue. *MNRAS*, 382:382–392, Nov 2007. doi:10.1111/j.1365-2966.2007.12379.x.

F. Murtagh and P. Legendre. Ward's hierarchical agglomerative clustering method: which algorithms implement ward???s criterion? *Journal of Classification*, 31(3): 274–295, 2014.

G. Neugebauer, H. J. Habing, R. van Duinen, H. H. Aumann, B. Baud, C. A. Beichman, D. A. Beintema, N. Boggess, P. E. Clegg, T. de Jong, J. P. Emerson, T. N. Gautier, F. C. Gillett, S. Harris, M. G. Hauser, J. R. Houck, R. E. Jennings, F. J. Low, P. L. Marsden, G. Miley, F. M. Olnon, S. R. Pottasch, E. Raimond, M. Rowan-Robinson, B. T. Soifer, R. G. Walker, P. R. Wesselius, and E. Young. The Infrared Astronomical Satellite (IRAS) mission. *ApJ*, 278:L1–L6, Mar. 1984. doi:10.1086/184209.

D. E. Osterbrock and G. J. Ferland. *Astrophysics of gaseous nebulae and active galactic nuclei*. 2006.

R. Paladini, C. Burigana, R. D. Davies, D. Maino, M. Bersanelli, B. Cappellini, P. Platania, and G. Smoot. A radio catalog of Galactic HII regions for applications from decimeter to millimeter wavelengths. *A&A*, 397:213–226, Jan. 2003. doi:10.1051/0004-6361:20021466.

P. Palmeirim, A. Zavagno, D. Elia, T. J. T. Moore, A. Whitworth, P. Tremblin, A. Traficante, M. Merello, D. Russeil, S. Pezzuto, L. Cambrésy, A. Baldeschi, M. Bandieramonte, U. Becciani, M. Benedettini, C. Buemi, F. Bufano, A. Bulpitt, R. Butora, D. Carey, A. Costa, L. Deharveng, A. Di Giorgio, D. Eden, A. Hajnal, M. Hoare, P. Kacsuk, P. Leto, K. Marsh, P. Mège, S. Molinari, M. Molinaro, A. Noriega-Crespo, E. Schisano, E. Sciacca, C. Trigilio, G. Umana, and F. Vitello. Spatial distribution of star formation related to ionized regions throughout the inner Galactic plane. *A&A*, 605:A35, Sept. 2017. doi:10.1051/0004-6361/201629963.

N. Panagia. Some Physical parameters of early-type stars. *AJ*, 78:929–934, Nov. 1973. doi:10.1086/111498.

M. Pasquato and C. Chung. Clustering clusters: unsupervised machine learning on globular cluster structural parameters. *arXiv e-prints*, art. arXiv:1901.05354, Jan 2019.

A. N. Pettitt. A two-sample anderson–darling rank statistic. *Biometrika*, pages 161–168, 1976.

M. Pomarès, A. Zavagno, L. Deharveng, M. Cunningham, P. Jones, S. Kurtz, D. Russeil, J. Caplan, and F. Comerón. Triggered star formation on the borders of the Galactic Hii region RCW 82. *A&A*, 494:987–1003, Feb 2009. doi:10.1051/0004-6361:200811050.

S. D. Price, M. P. Egan, S. J. Carey, D. R. Mizuno, and T. A. Kuchar. Midcourse Space Experiment Survey of the Galactic Plane. *AJ*, 121:2819–2842, May 2001. doi:10.1086/320404.

C. Quireza, R. T. Rood, D. S. Balser, and T. M. Bania. Radio Recombination Lines in Galactic H II Regions. *ApJS*, 165:338–359, July 2006. doi:10.1086/503901.

A. C. Raga, J. Cantó, and L. F. Rodríguez. Analytic and numerical models for the expansion of a compact H II region. *MNRAS*, 419:L39–L43, Jan. 2012a. doi:10.1111/j.1745-3933.2011.01173.x.

A. C. Raga, J. Cantó, and L. F. Rodríguez. The universal time-evolution of an expanding HII region. *Rev. Mexicana Astron. Astrofis.*, 48:149–157, Apr. 2012b.

W. Reich, P. Reich, and E. Fuerst. The Effelsberg 21 CM radio continuum survey of theGalacticplane between L = 357 deg. and L = 95.5 deg. *Astronomy and Astrophysics Supplement Series*, 83:539, Jun 1990.

M. J. Reid, K. M. Menten, X. W. Zheng, A. Brunthaler, L. Moscadelli, Y. Xu, B. Zhang, M. Sato, M. Honma, T. Hirota, K. Hachisuka, Y. K. Choi, G. A. Moellenbrock, and A. Bartkiewicz. Trigonometric Parallaxes of Massive Star-Forming Regions. VI. Galactic Structure, Fundamental Parameters, and Noncircular Motions. *ApJ*, 700:137–148, Jul 2009. doi:10.1088/0004-637X/700/1/137.

M. J. Reid, K. M. Menten, A. Brunthaler, X. W. Zheng, T. M. Dame, Y. Xu, Y. Wu, B. Zhang, A. Sanna, M. Sato, K. Hachisuka, Y. K. Choi, K. Immer, L. Moscadelli, K. L. J. Rygl, and A. Bartkiewicz. Trigonometric Parallaxes of High Mass Star Forming Regions: The Structure and Kinematics of the Milky Way. *ApJ*, 783:130, Mar 2014. doi:10.1088/0004-637X/783/2/130.

B. Reipurth. Star formation in BOK globules and low-mass clouds. *A&A*, 117:183–198, Jan 1983.

J. Roman-Duval, J. M. Jackson, M. Heyer, J. Rathborne, and R. Simon. Physical Properties and Galactic Distribution of Molecular Clouds Identified in the Galactic Ring Survey. *ApJ*, 723:492–507, Nov 2010. doi:10.1088/0004-637X/723/1/492.

R. H. Rubin. A Discussion of the Sizes and Excitation of H II Regions. *ApJ*, 154:391, Oct. 1968. doi:10.1086/149766.

D. Russeil, A. Zavagno, P. Mège, Y. Poulin, S. Molinari, and L. Cambresy. The Milky Way rotation curve revisited. *A&A*, 601:L5, May 2017. doi:10.1051/0004-6361/201730540.

E. E. Salpeter. The Luminosity Function and Stellar Evolution. *ApJ*, 121:161, Jan. 1955. doi:10.1086/145971.

J. Sánchez Almeida and C. Allende Prieto. Automated Unsupervised Classification of the Sloan Digital Sky Survey Stellar Spectra using k-means Clustering. *ApJ*, 763:50, Jan 2013. doi:10.1088/0004-637X/763/1/50.

B. D. Savage, R. C. Bohlin, J. F. Drake, and W. Budich. A survey of interstellar molecular hydrogen. I. *ApJ*, 216:291–307, Aug 1977. doi:10.1086/155471.

D. J. Schlegel, D. P. Finkbeiner, and M. Davis. Maps of Dust Infrared Emission for Use in Estimation of Reddening and Cosmic Microwave Background Radiation Foregrounds. *ApJ*, 500:525–553, Jun 1998. doi:10.1086/305772.

M. Sewilo, C. Watson, E. Araya, E. Churchwell, P. Hofner, and S. Kurtz. Resolution of Distance Ambiguities of Inner Galaxy Massive Star Formation Regions. II. *ApJS*, 154:553–578, Oct. 2004. doi:10.1086/423247.

H. Shapley. On the changes in the spectrum, period, and lightcurve of the Cepheid variable RR Lyrae. *ApJ*, 43:217–233, Apr 1916. doi:10.1086/142246.

J. Shi, A. Lapi, C. Mancuso, H. Wang, and L. Danese. Angular momentum of early- and late-type galaxies: Nature or nurture? *The Astrophysical Journal*, 843(2): 105, jul 2017. doi:10.3847/1538-4357/aa7893. URL `https://doi.org/10.3847%2F1538-4357%2Faa7893`.

H. Shimodaira. Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling. *The Annals of Statistics*, 32(6):2616–2641, 2004.

F. H. Shu. Self-similar collapse of isothermal spheres and star formation. *ApJ*, 214: 488–497, Jun 1977. doi:10.1086/155274.

F. H. Shu, F. C. Adams, and S. Lizano. Star formation in molecular clouds: observation and theory. *Annual Review of Astronomy and Astrophysics*, 25:23–81, Jan 1987. doi:10.1146/annurev.aa.25.090187.000323.

R. J. Simpson, M. S. Povich, S. Kendrew, C. J. Lintott, E. Bressert, K. Arvidsson, C. Cyganowski, S. Maddison, K. Schawinski, R. Sherman, A. M. Smith, and G. Wolf-Chase. The Milky Way Project First Data Release: a bubblier Galactic disc. *MNRAS*, 424:2442–2460, Aug. 2012. doi:10.1111/j.1365-2966.2012.20770.x.

N. V. Smirnov. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bull. Math. Univ. Moscou*, 2(2):3–14, 1939.

L. Spitzer. *Physical processes in the interstellar medium.* 1978. doi:10.1002/9783527617722.

L. Spitzer, Jr. Diffuse matter in space. *Interscience Tracts on Physics and Astronomy*, 28, 1968.

J. Steinacker, M. Baes, and K. D. Gordon. Three-dimensional dust radiative transfer. *Annual Review of Astronomy and Astrophysics*, 51:63–104, 2013.

J. M. Stil, A. R. Taylor, J. M. Dickey, D. W. Kavars, P. G. Martin, T. A. Rothwell, A. I. Boothroyd, F. J. Lockman, and N. M. McClure-Griffiths. The VLA Galactic Plane Survey. *AJ*, 132:1158–1176, Sep 2006. doi:10.1086/505940.

B. Strömgren. The Physical State of Interstellar Hydrogen. *ApJ*, 89:526, May 1939. doi:10.1086/144074.

O. Struve and C. T. Elvey. Emission Nebulosities in Cygnus and Cepheus. *ApJ*, 88: 364, Oct. 1938. doi:10.1086/143992.

J. C. Tan, M. T. Beltrán, P. Caselli, F. Fontani, A. Fuente, M. R. Krumholz, C. F. McKee, and A. Stolte. Massive Star Formation. *Protostars and Planets VI*, pages 149–172, 2014. doi:10.2458/azu_uapress_9780816531240-ch007.

A. R. Taylor, S. J. Gibson, M. Peracaula, P. G. Martin, T. L. Landecker, C. M. Brunt, P. E. Dewdney, S. M. Dougherty, A. D. Gray, L. A. Higgs, C. R. Kerton, L. B. G. Knee, R. Kothes, C. R. Purton, B. Uyaniker, B. J. Wallace, A. G. Willis, and D. Durand. The Canadian Galactic Plane Survey. *AJ*, 125:3145–3164, Jun 2003. doi:10.1086/375301.

M. A. Thompson, J. S. Urquhart, T. J. T. Moore, and L. K. Morgan. The statistics of triggered star formation: an overdensity of massive young stellar objects around Spitzer bubbles. *MNRAS*, 421:408–418, Mar. 2012. doi:10.1111/j.1365-2966.2011.20315.x.

A. G. G. M. Tielens and D. Hollenbach. Photodissociation regions. I. Basic model. *ApJ*, 291:722–746, Apr 1985. doi:10.1086/163111.

A. P. Topchieva, D. S. Wiebe, M. S. Kirsanova, and V. V. Krushinskii. Infrared Morphology of Regions of Ionized Hydrogen. *ArXiv e-prints*, Jan. 2018.

W. S. Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419, 1952.

P. Tremblin, L. D. Anderson, P. Didelon, A. C. Raga, V. Minier, E. Ntormousi, A. Pettitt, C. Pinto, M. R. Samal, N. Schneider, and A. Zavagno. Age, size, and position of H ii regions in the Galaxy. Expansion of ionized gas in turbulent molecular clouds. *A&A*, 568:A4, Aug. 2014. doi:10.1051/0004-6361/201423959.

J. S. Urquhart, T. J. T. Moore, M. G. Hoare, S. L. Lumsden, R. D. Oudmaijer, J. M. Rathborne, J. C. Mottram, B. Davies, and J. J. Stead. The Red MSX Source survey: distribution and properties of a sample of massive young stars. *MNRAS*, 410:1237–1250, Jan 2011a. doi:10.1111/j.1365-2966.2010.17514.x.

J. S. Urquhart, L. K. Morgan, C. C. Figura, T. J. T. Moore, S. L. Lumsden, M. G. Hoare, R. D. Oudmaijer, J. C. Mottram, B. Davies, and M. K. Dunham.

The Red MSX Source survey: ammonia and water maser analysis of massive star-forming regions. *MNRAS*, 418:1689–1706, Dec 2011b. doi:10.1111/j.1365-2966.2011.19594.x.

J. S. Urquhart, C. C. Figura, T. J. T. Moore, M. G. Hoare, S. L. Lumsden, J. C. Mottram, M. A. Thompson, and R. D. Oudmaijer. The RMS survey: galactic distribution of massive star formation. *MNRAS*, 437:1791–1807, Jan. 2014. doi:10.1093/mnras/stt2006.

C. M. Urry and P. Padovani. Unified Schemes for Radio-Loud Active Galactic Nuclei. *Publications of the Astronomical Society of the Pacific*, 107:803, Sep 1995. doi:10.1086/133630.

D. van Buren and R. McCray. Bow shocks and bubbles are seen around hot stars by IRAS. *ApJ*, 329:L93–L96, June 1988. doi:10.1086/185184.

E. A. Vitrichenko, D. K. Nadyozhin, and T. L. Razinkova. Mass-luminosity relation for massive stars. *Astronomy Letters*, 33:251–258, Apr. 2007. doi:10.1134/S1063773707040044.

S. Walch, A. P. Whitworth, T. G. Bisbas, R. Wünsch, and D. A. Hubber. Clumps and triggered star formation in ionized molecular clouds. *MNRAS*, 435:917–927, Oct 2013. doi:10.1093/mnras/stt1115.

S. K. Walch, A. P. Whitworth, T. Bisbas, R. Wünsch, and D. Hubber. Dispersal of molecular clouds by ionizing radiation. *MNRAS*, 427:625–636, Nov 2012. doi:10.1111/j.1365-2966.2012.21767.x.

A. J. Walsh, M. G. Burton, A. R. Hyland, and G. Robinson. Studies of ultracompact HII regions - II. High-resolution radio continuum and methanol maser survey. *MNRAS*, 301:640–698, Dec 1998. doi:10.1046/j.1365-8711.1998.02014.x.

J. Ward and H. Joe. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

A. Watanabe. How many landmarks are enough to characterize shape and size variation? *PLOS ONE*, 13(6):1–17, 06 2018. doi:10.1371/journal.pone.0198341. URL `https://doi.org/10.1371/journal.pone.0198341`.

C. Watson, M. S. Povich, E. B. Churchwell, B. L. Babler, G. Chunev, M. Hoare, R. Indebetouw, M. R. Meade, T. P. Robitaille, and B. A. Whitney. Infrared Dust Bubbles: Probing the Detailed Structure and Young Massive Stellar Populations of Galactic H II Regions. *ApJ*, 681:1341–1355, July 2008. doi:10.1086/588005.

C. Watson, T. Corn, E. B. Churchwell, B. L. Babler, M. S. Povich, M. R. Meade, and B. A. Whitney. IR Dust Bubbles. II. Probing the Detailed Structure and Young Massive Stellar Populations of Galactic H II Regions. *ApJ*, 694:546–555, Mar. 2009. doi:10.1088/0004-637X/694/1/546.

R. Weaver, R. McCray, J. Castor, P. Shapiro, and R. Moore. Interstellar bubbles. II - Structure and evolution. *ApJ*, 218:377–395, Dec. 1977. doi:10.1086/155692.

C. Weidner and J. S. Vink. The masses, and the mass discrepancy of O-type stars. *A&A*, 524:A98, Dec. 2010. doi:10.1051/0004-6361/201014491.

C. Weidner, P. Kroupa, and I. A. D. Bonnell. The relation between the most-massive star and its parental star cluster mass. *MNRAS*, 401:275–293, Jan. 2010. doi:10.1111/j.1365-2966.2009.15633.x.

C. Wolf, K. Meisenheimer, H.-W. Rix, A. Borch, S. Dye, and M. Kleinheinrich. The COMBO-17 survey: Evolution of the galaxy luminosity function from 25 000 galaxies with 0.2 - z - 1.2. *A&A*, 401:73–98, Apr. 2003. doi:10.1051/0004-6361:20021513.

D. O. S. Wood and E. Churchwell. The morphologies and physical properties of ultracompact H II regions. *ApJS*, 69:831–895, Apr. 1989. doi:10.1086/191329.

S. E. Woosley, A. Heger, and T. A. Weaver. The evolution and explosion of massive stars. *Reviews of Modern Physics*, 74:1015–1071, Nov 2002. doi:10.1103/RevModPhys.74.1015.

E. L. Wright, P. R. M. Eisenhardt, A. K. Mainzer, M. E. Ressler, R. M. Cutri, T. Jarrett, J. D. Kirkpatrick, D. Padgett, R. S. McMillan, M. Skrutskie, S. A. Stanford, M. Cohen, R. G. Walker, J. C. Mather, D. Leisawitz, T. N. Gautier, III, I. McLean, D. Benford, C. J. Lonsdale, A. Blain, B. Mendez, W. R. Irace, V. Duval, F. Liu, D. Royer, I. Heinrichsen, J. Howard, M. Shannon, M. Kendall, A. L. Walsh, M. Larsen, J. G. Cardon, S. Schick, M. Schwalm, M. Abid, B. Fabinsky, L. Naes, and C.-W. Tsai. The Wide-field Infrared Survey Explorer (WISE): Mission Description and Initial On-orbit Performance. *AJ*, 140:1868–1881, Dec. 2010. doi:10.1088/0004-6256/140/6/1868.

C. G. Wynn-Williams. The search for infrared protostars. *Annual Review of Astronomy and Astrophysics*, 20:587–618, Jan 1982. doi:10.1146/annurev.aa.20.090182.003103.

J.-L. Xu and B.-G. Ju. Star formation associated with a large-scale infrared bubble. *A&A*, 569:A36, Sept. 2014. doi:10.1051/0004-6361/201423952.

A. Zavagno, L. Deharveng, F. Comerón, J. Brand, F. Massi, J. Caplan, and D. Russeil. Triggered massive-star formation on the borders of Galactic H II regions. II. Evidence for the collect and collapse process around RCW 79. *A&A*, 446:171–184, Jan 2006. doi:10.1051/0004-6361:20053952.

H. Zinnecker and H. W. Yorke. Toward Understanding Massive Star Formation. *Annual Review of Astronomy and Astrophysics*, 45:481–563, Sep 2007. doi:10.1146/annurev.astro.44.051905.092549.

Author's Publications

Justyn Campbell-White, Ali Ahmad, Dirk Froebrich, and Alfred Kume. **Shape analysis of H II regions - II. Synthetic Observations**. *MNRAS*. In Prep.

R. A. Street, E. Bachelet, Y. Tsapras, M. P. G. Hundertmark, V. Bozza, M. Dominik, D. M. Bramich, A. Cassan, K. Horne, S. Mao, A. Saha, J. Wambsganss, Weicheng Zang, U. G. Jorgensen, P. Longa-Pena, N. Peixinho, S. Sajadian, M. J. Burgdorf, **Justyn Campbell-White**, S. Dib, D. F. Evans, Y. I. Fujii, T. C. Hinse, E. Khalouei, S. Lowry, S. Rahvar, M. Rabus, J. Skottfelt, C. Snodgrass, J. Southworth, and J. Tregloan-Reed. **OGLE-2018-BLG-0022: A Nearby M-dwarf Binary**. *arXiv e-prints*, art. arXiv:1903.08733, Mar 2019. https://ui.adsabs.harvard.edu/#abs/2019arXiv190308733S.

Yossi Shvartzvald, Jennifer C. Yee, Jan Skowron, Chung-Uk Lee, Andrzej Udalski, Sebastiano Calchi Novati, Valerio Bozza, Charles A. Beichman, Geoffery Bryden, Sean Carey, B. Scott Gaudi, Calen B. Henderson, Wei Zhu, Spitzer team, Etienne Bachelet, Greg Bolt, Grant Christie, Dan Maoz, Tim Natusch, Richard W. Pogge, Rachel A. Street, Thiam-Guan Tan, Yiannis Tsapras, LCO, μFUN Follow-up Teams, Paweł Pietrukowicz, Igor Soszyński, Michał K. Szymański, Przemek Mróz, Radoslaw Poleski, Szymon Kozłowski, Krzysztof Ulaczyk, Michał Pawlak, Krzysztof A. Rybicki, Patryk Iwanek, OGLE Collaboration, Michael D. Albrow, Sang-Mok Cha, Sun-Ju Chung, Andrew Gould, Cheongho Han, Kyu-Ha Hwang, Youn Kil Jung, Dong-Jin Kim, Hyoun-Woo Kim, Seung-Lee Kim, Dong-Joo Lee, Yongseok Lee, Byeong-Gon Park, Yoon-Hyun Ryu, In-Gu Shin, Weicheng Zang, KMTNet Collaboration, Martin Dominik, Christiane Helling, Markus Hundertmark, Uffe G. Jørgensen, Penelope Longa-Peña, Stephen Lowry, Sedighe Sajadian, Martin J. Burgdorf, **Justyn Campbell-White**, Simona Ciceri, Daniel F. Evans, Yuri I. Fujii, Tobias C. Hinse, Sohrab Rahvar, Markus Rabus, Jesper Skot-

tfelt, Colin Snodgrass, John Southworth, and MiNDSTEp Collaboration. **Spitzer Microlensing Parallax for OGLE-2017-BLG-0896 Reveals a Counter-rotating Low-mass Brown Dwarf**. *AJ*, 157:106, Mar 2019. doi:10.3847/1538-3881/aafe12. https://ui.adsabs.harvard.edu/#abs/2019AJ....157..106S.

P. Bacci, M. Maestripieri, S. Sisi, L. Tesi, G. Fagioli, W. Hasubick, P. Camilleri, K. Kadota, A. Baransky, O. Maznychenko, A. Kasianchuk, **J. Campbell-White**, S. Rahvar, M. Rabus, S. Dib, A. Andrews, M. Burgdorf, Y. Fujii, T. L. Farnham, M. M. Knight, M. Jaeger, E. Prosperi, W. Vollmann, J. Nicolas, C. Rinner, F. Kugel, E. Bryssinck, J. F. Soulier, M. T. Hui, C. H. Hsia, Z. Y. Lin, H. Y. Hsiao, C. S. Lin, H. C. Lin, Y. Liao, W. H. Ip, M. Dahlhaus, D. Bloom, I. Welzel, L. Belli, P. Breitenstein, J. Drummond, J. Bulger, T. Dukes, T. Lowe, A. Schultz, M. Willman, K. Chambers, S. Chastel, L. Denneau, H. Flewelling, M. Huber, E. Magnier, Y. Ramanjooloo, R. Wainscoat, C. Waters, R. Weryk, W. F. Cashwell, B. Lutkenhoner, C. Bell, J. Gonzalez, K. Hills, P. Carson, C. Gerhard, R. Fichtl, K. Korlevic, D. Vida, F. Hrzenjak, E. Pettarin, G. Ventre, P. Sicoli, T. Ikemura, H. Sato, M. Mattiazzo, H. Williams, A. Heinze, H. Weiland, J. Tonry, A. Fitzsimmons, D. Young, N. Paul, A. Maury, J. B. de Vanssay, and J. G. Bosch. **COMET 46P/Wirtanen**. *Minor Planet Electronic Circulars*, 2018-R14, Sep 2018. https://minorplanetcenter.net//iau/mpec/K18/K18R14.html.

T. M. Gledhill, D. Froebrich, **J. Campbell-White**, and A. M. Jones. **Planetary nebulae in the UWISH2 survey**. *MNRAS*, 479:3759–3777, September 2018. doi:10.1093/mnras/sty1580. https://ui.adsabs.harvard.edu/#abs/2018MNRAS.479.3759G.

D. Froebrich, **J. Campbell-White**, A. Scholz, J. Eislöffel, T. Zegmott, S. J. Billington, J. Donohoe, S. V. Makin, R. Hibbert, R. J. Newport, R. Pickard, N. Quinn, T. Rodda, G. Piehler, M. Shelley, S. Parkinson, K. Wiersema, and I. Walton. **A survey for variable young stars with small telescopes: First results from HOYS-CAPS**. *MNRAS*, 478:5091–5103, August 2018a. doi:10.1093/mnras/sty1350. https://ui.adsabs.harvard.edu/#abs/2018MNRAS.478.5091F.

Justyn Campbell-White, Dirk Froebrich, and Alfred Kume. **Shape analysis of H II regions - I. Statistical clustering**. *MNRAS*, 477:5486–5500, July 2018. doi:10.1093/mnras/sty954. https://ui.adsabs.harvard.edu/#abs/2018MNRAS.477.5486C.

Dirk Froebrich, Aleks Scholz, **Justyn. Campbell-White**, James Crumpton, Emma

D'Arcy, Sally V. Makin, Tarik Zegmott, Samuel J. Billington, Ricky Hibbert, Robert J. Newport, and Callum R. Fisher. **Variability in IC5070: Two Young Stars with Deep Recurring Eclipses**. *Research Notes of the American Astronomical Society*, 2:61, June 2018b. doi:10.3847/2515-5172/aacd48. https://ui.adsabs.harvard.edu/#abs/2018RNAAS...2b..61F.

C. Han, S. Calchi Novati, A. Udalski, C. U. Lee, A. Gould, V. Bozza, P. Mróz, P. Pietrukowicz, J. Skowron, M. K. Szymański, R. Poleski, I. Soszyński, S. Kozłowski, K. Ulaczyk, M. Pawlak, K. Rybicki, P. Iwanek, OGLE Collaboration, M. D. Albrow, S. J. Chung, K. H. Hwang, Y. K. Jung, Y. H. Ryu, I. G. Shin, Y. Shvartzvald, J. C. Yee, W. Zang, W. Zhu, S. M. Cha, D. J. Kim, H. W. Kim, S. L. Kim, D. J. Lee, Y. Lee, B. G. Park, R. W. Pogge, W. T. Kim, KMTNet Collaboration, C. Beichman, G. Bryden, S. Carey, B. S. Gaudi, C. B. Henderson, Spitzer Team, M. Dominik, C. Helling, M. Hundertmark, U. G. Jørgensen, P. Longa- Peña, S. Lowry, S. Sajadian, M. J. Burgdorf, **J. Campbell-White**, S. Ciceri, D. F. Evans, L. K. Haikala, T. C. Hinse, S. Rahvar, M. Rabus, C. Snodgrass, and MiNDSTEp Collaboration. **OGLE-2017-BLG-0329L: A Microlensing Binary Characterized with Dramatically Enhanced Precision Using Data from Space-based Observations**. *ApJ*, 859:82, June 2018. doi:10.3847/1538-4357/aabd87. https://ui.adsabs.harvard.edu/#abs/2018ApJ...859...82H.

A. Udalski, Y. H. Ryu, S. Sajadian, A. Gould, P. Mróz, R. Poleski, M. K. Szymański, J. Skowron, I. Soszyński, S. Kozłowski, P. Pietrukowicz, K. Ulaczyk, M. Pawlak, K. Rybicki, P. Iwanek, M. D. Albrow, S. J. Chung, C. Han, K. H. Hwang, K. Jung, Y., I. G. Shin, Y. Shvartzvald, J. C. Yee, W. Zang, W. Zhu, S. M. Cha, D. J. Kim, H. W. Kim, S. L. Kim, C. U. Lee, D. J. Lee, Y. Lee, B. G. Park, R. W. Pogge, V. Bozza, M. Dominik, C. Helling, M. Hundertmark, U. G. Jørgensen, P. Longa- Peña, S. Lowry, M. Burgdorf, **J. Campbell-White**, S. Ciceri, D. Evans, R. Figuera Jaimes, Y. I. Fujii, L. K. Haikala, T. Henning, T. C. Hinse, L. Mancini, N. Peixinho, S. Rahvar, M. Rabus, J. Skottfelt, C. Snodgrass, J. Southworth, and C. von Essen. **OGLE-2017-BLG-1434Lb: Eighth $q < 1X10^{-4}$ Mass-Ratio Microlens Planet Confirms Turnover in Planet Mass-Ratio Function**. *Acta Astron.*, 68:1–42, March 2018. https://ui.adsabs.harvard.edu/#abs/2018AcA....68....1U.

A. Sicilia-Aguilar, A. Oprandi, D. Froebrich, M. Fang, J. L. Prieto, K. Stanek, A. Scholz, C. S. Kochanek, Th. Henning, R. Gredel, T. W. S. Holoien, M. Rabus, B. J. Shappee, S. J. Billington, **J. Campbell-White**, and T. J. Zegmott. **The**

2014-2017 outburst of the young star ASASSN-13db. A time-resolved picture of a very-low-mass star between EXors and FUors. *A&A*, 607: A127, November 2017. doi:10.1051/0004-6361/201731263. https://ui.adsabs.harvard.edu/#abs/2017A&A...607A.127S.

Dirk Froebrich, **Justyn Campbell-White**, Tarik Zegmott, Samuel J. Billington, Sally V. Makin, and Justin Donohoe. **Optical brightness and colours of V2492Cyg before, during and after the recent record peak in brightness**. *The Astronomer's Telegram*, 10259:1, April 2017. https://ui.adsabs.harvard.edu/#abs/2017ATel10259....1F.