A Search for Star Clusters Across the Galactic Plane

Author: **David Bush**

A thesis submitted in partial fulfilment of the

requirements for the degree of

Doctor of Philosophy

Supervised By:

Prof. Phil Lucas,

Prof. Jan Forbrich

(University of Hertfordshire)

October 2026

Declaration

I declare that no part of this work is being submitted concurrently for another award of the University or any other awarding body or institution. This thesis contains a substantial body of work that has not previously been submitted successfully for an award of the University or any other awarding body or institution. Except where indicated otherwise in the submission, the submission is my own work and has not previously been submitted successfully for any award.

Abstract

This work exploits the resolution and sensitivity of the UKIRT-WFCAM Galactic Plane Survey (UGPS) to extend the census of young, compact Galactic clusters beyond the reach of infrared surveys such as 2MASS and Spitzer. An initial sample of 403 candidates was selected by visual inspection of UGPS K-band images, a strategy that finds compact or heavily extincted systems often missed by automated detection algorithms.

Each candidate underwent a uniform analysis. Statistical field-star decontamination removed likely non-members and enabled construction of intrinsic colour-magnitude and colour-colour diagrams, from which I derived central coordinates, core radii and angular diameters. To estimate ages and distances I developed an automated MIST isochrone-fitting pipeline incorporating Bayestar19 dust maps, validated against six known clusters spanning 1 to over 300 Myr, with the reddening-free Q parameter used to identify infrared excess from young stellar objects.

Multi-wavelength and kinematic data provided independent corroboration. WISE imaging characterised local environments, identified star-formation signatures such as HII regions and yielded kinematic distances. Gaia DR3 cross-matching gave proper motions and photogeometric distances for less-extincted stars, and applying the clustering algorithm HDBSCAN to this astrometric space corroborated a subset of candidates while exposing the limits of optical astrometry for embedded systems.

This integrated analysis identified 178 previously unrecognised clusters with high confidence. Combining the isochrone fits with the reddening-free Q diagnostic and mid-infrared star-formation markers, 154 are classified as young or emerging systems and 24 as more evolved open clusters; 76 were recovered from candidates initially

screened as containing no cluster, and would have been lost to visual inspection alone. The analysis also confirmed 74 previously reported candidates and rejected 20 spurious over-densities as artefacts. The discoveries are on average 13% more compact than catalogued clusters, with a median visual extinction of $A_V = 5.11$ mag and twelve systems above $A_V = 10$, extending the known population into the obscured regime. Eight clusters have independent WISE HII kinematic distances of 7.7 to 12.6 kpc, placing them in the outer spiral arms. The catalogue substantially enlarges the inventory of young Galactic clusters and provides a well-characterised sample for follow-up.

Keywords: Star Formation, Galactic Plane, Star Clusters

Acknowledgments

Firstly I would like to thank my Grandparents Margery and Reginald Stokes who, in gifting me 'The Kingfisher Astronomy Handbook' for my 9th Birthday, and later a nice Newtonian telescope, started a lifetime's interest in Space, Science and Astronomy.

I would also like to thank my parents Christine and Bernard, my partner Kelly and our children Lily, Archie and Monty for their long suffering support over the years, especially the last few months of 14 hr days writing up and working long hours at the same time!

Next, I would like to thank my supervisors Prof. Phil Lucas and Prof. Jan Forbrich for their support, especially for the many times I was ready to quit and Phil talked me round!

I would also like to thank the staff at the University of Hertfordshire Doctoral College for all their help and support over the years.

This work has made use of data from the European Space Agency (ESA) mission *Gaia* (`https://www.cosmos.esa.int/gaia`), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, `https://www.cosmos.esa.int/web/gaia/dpac/consortium`). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement.

This research has also made use of NASA's Astrophysics Data System Abstract Service; the NASA / IPAC Infrared Science Archive (which is operated by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration)

Table of Contents

Declaration i

Abstract ii

Acknowledgments iv

Table of Contents v

List of Tables ix

List of Figures x

1 Introduction 1

1.1 Gravitational Instability and Mechanisms of Cloud Collapse 5

1.1.1 Isothermal Collapse 5

1.1.2 Formation of the First Hydrostatic Core 9

1.1.3 Second Collapse, Protostar Formation and Evolution 9

1.1.4 Young Stellar Object Evolution onto the Main Sequence 13

1.2 Cluster Formation 20

1.2.1 Cluster Taxonomy 26

1.3 Outstanding Questions 28

1.4 Project Aims 30

2 Data Sources and Observations 32

2.1 The United Kingdom Infrared Deep Sky Survey's Galactic Plane Survey (UKIDSS GPS) . . . 33
2.1.1 The Wide Field Camera (WFCAM) . . . 34
2.2 The Two-Micron All-Sky Survey (2MASS) . . . 38
2.2.1 Managing saturation limits of the GPS with 2MASS . . . 40
2.3 The Wide-field Infrared Survey Explorer (WISE) . . . 48
2.4 The Gaia mission . . . 51
2.4.1 Using Gaia data in the confirmation and basic characterisation of cluster candidates. . . . 55
2.4.2 Strasbourg astronomical Data Centre (CDS) and Online Astronomical Databases . . . 57

3 Initial Search, Classification and Field Star Decontamination 60
3.1 Initial Coordinates and Apparent Radius . . . 63
3.2 Field-star decontamination . . . 64
3.2.1 Colour-Colour-Magnitude Method (CCM) . . . 65
3.2.2 Bonatto Method . . . 66
3.2.3 Decontamination Examples . . . 69
3.2.4 Bulk Decontamination Process . . . 72
3.2.5 Summary of the Decontamination Process . . . 73

4 Isochrone Analysis 74
4.1 Introduction . . . 74
4.2 Methodological Framework . . . 75
4.2.1 Theoretical Models and Metallicity . . . 75
4.2.2 Reddening and Extinction Law . . . 75
4.2.3 Population Diagnostics: The Q-Parameter and Multi-Wavelength Cross-Checks . . . 76
4.2.4 2D Euclidean Fitting Metric . . . 77

4.2.5 3D Dust Map Integration and Reproducibility of the '3D Anchored Fit' . . . 77
4.2.6 Systematic Uncertainties: Binary Fraction and Worst-Case Quantitative Envelope . . . 78
4.2.7 Limitations at Extreme Extinction: Differential Reddening . . . 79
4.2.8 Limitations of Photometric Decontamination . . . 81
4.3 Validation Against Known Clusters . . . 83
4.4 Results: Parameters of New Discovery Candidates . . . 86
4.4.1 Quality Control and Validity Grading . . . 86
4.4.2 Standout Discovery Populations . . . 89
4.5 Summary . . . 93

5 Basic Characterisation 95
5.1 Algorithmically Determined Central Coordinates . . . 96
5.1.1 Kernel Density Estimation (KDE) . . . 97
5.1.2 Arithmetic Mean Centre (AMC) . . . 98
5.2 Radial Density Profiles (RDP) and Half-Maximum Radius (HMR) . . . 99
5.2.1 King Profile . . . 101
5.2.2 Plummer Profile . . . 102
5.2.3 Elson-Fall-Freeman (EFF) profile . . . 102
5.3 WISE Far-Infrared Images . . . 105
5.4 Literature and Catalogue Search . . . 106

6 Gaia Proper Motion and Photogeometric Distance 109
6.1 Proper Motions . . . 111
6.2 Parallax Derived Distances . . . 112
6.3 Gaia Control Plots . . . 116
6.4 Example Plots - Unknown Candidates . . . 118
6.5 Gaia Cross-match Summary . . . 120

7 HDBSCAN Clustering Algorithm 121

7.1 Methodological Role and Avoidance of Circularity 122
7.2 Overview of the Algorithm 122
7.2.1 Transform the Data Space Using Density 123
7.2.2 Build the Minimum Spanning Tree 125
7.2.3 Build the Cluster Hierarchy 125
7.2.4 Condense the Cluster Tree 127
7.2.5 Extract the Clusters 128
7.3 Applying HDBSCAN to the Sample Candidates 130
7.3.1 HDBSCAN Example Output 133
7.3.2 HDBSCAN UGPS Cross-match 134
7.3.3 Comparison of HDBSCAN and Bonatto CMDs 141
7.4 HDBSCAN Summary 147

8 Summary, Conclusions and Further Work **149**
8.1 Summary 150
8.1.1 Sensitivity of the 'No Cluster' Fraction to the Applied Tests . . . 157
8.2 Conclusion 159
8.3 Further work 160

A Candidate Data **163**

Bibliography **166**

List of Tables

1.1 Cluster Taxonomy 28

4.1 Validation results for known clusters. 83

5.1 SIMBAD object classifications related to star formation activity used for candidate cross-matching 108

7.1 NGC-2401 - HDBSCAN-Bonatto Cross-match Statistics 134

7.2 NGC-2262 - HDBSCAN-Bonatto Cross-match Statistics 135

7.3 Candidate 15 - HDBSCAN-Bonatto Cross-match Statistics 136

8.1 Breakdown of Results by Class and Status 151

8.2 Unknown Candidates with WISE Distances 156

8.3 Summary of Successful Gaia and HDBSCAN Test Results 157

8.4 Number of Candidates Supported as Clusters by Each Test 158

A.1 Column Descriptions - *appendix-a_candidate_data.xlsx* 163

List of Figures

1.1 Classification of YSO Evolution 12

1.2 Near-Infrared Colour-Colour Diagram showing theoretical colour relationship for main-sequence and giant stars 16

1.3 Colour-Colour Diagrams for T-Tauri Stars Including Reddening Vector . . 17

1.4 NIR CCD of IC 348 highlighting infrared excess 18

1.5 Pre-main-sequence evolutionary tracks and isochrones between the birthline and the ZAMS 19

1.6 UKIDSS colour-magnitude diagrams for an uncrowded and a crowded field, showing the main-sequence and giant loci 20

1.7 NIR CMD of Spitzer-detected young stars in Orion demonstrating the presence of infrared excess 21

1.8 Theoretical zero-age main sequences for single and binary stars showing the offset caused by unresolved companions 22

1.9 The 8-parameter version of the Chabrier IMF showing the distribution of stellar masses 23

1.10 CMDs for six star clusters of varying ages over-plotted with PARSEC and MIST isochrones 25

1.11 Example J vs J-H and K vs J-K colour-magnitude diagrams for Teutsch 10 and Teutsch 25 26

2.1 The UKIRT atop Mauna Kea 32

2.2 Sky coverage of the UKIDSS GPS DR10 survey 34

2.3 WFCAM detector layout and field of view for the four IR arrays 37

2.4 Optical layout of the WFCAM instrument 37

2.5 WFCAM instrument installed at the UKIRT 38
2.6 The 1.3m 2MASS survey telescope at Mt. Hopkins, Arizona 39
2.7 Saturated bright star in a GPS field showing typical detector artefacts and diffraction spikes 42
2.8 SQL statements used to extract the cross-matched GPS and 2MASS data for magnitude comparison 44
2.9 UKIDSS vs 2MASS magnitude comparison for the outer Galaxy (l = 100-102°) showing the selection of magnitude cuts 46
2.10 UKIDSS vs 2MASS magnitude comparison for the inner Galaxy (l = 30-31°) showing the selection of magnitude cuts 47
2.11 The WISE flight system in survey configuration showing the cryostat and telescope components 48
2.12 WISE multi-band cluster candidate images at 10 and 30 arcmin fields showing environmental markers 50
2.13 WISE 1-degree RGB composite of the candidate field demonstrating large-scale environmental context 50
2.14 Gaia satellite subsystems and deployed sun shield 51
2.15 Geometric principles of Gaia parallax measurement showing the along-scan angle dependence 53
2.16 Illustration of the Gaia scanning law and the path of the viewing direction 54
2.17 Geometric intersection of Gaia scans enabling two-dimensional coordinate determination 54
2.18 Gaia proper motion and photogeometric distance plots for NGC-2360 highlighting member identification 56

3.1 K-band leavstack multiframe showing images from the four WFCAM detectors with stellar over density and nebulosity (candidate-3) 62
3.2 Distribution of the 403 cluster candidates (blue circles) across the UKIDSS GPS DR10 survey area (grey shaded) 63

3.3 3D decontamination CMD and CCD for Candidate 75 (NGC 7044) showing member selection 70
3.4 3D decontamination CMD and CCD for the emerging cluster Candidate 6 (NGC 1931) showing member selection and infrared excess 72

4.1 Effect of differential extinction on an embedded cluster (clus288) 80
4.2 Comparison of two distinct artefact failure modes (clus378 and clus170) . 82
4.3 Validation fits for the two controls with a blind fit 85
4.4 Validation fits for the remaining evolved control clusters 87
4.5 Validation fits for the remaining young control clusters 88
4.6 Grade A isochrone fit for an evolved discovery (clus307) 90
4.7 Grade A isochrone fit for a young discovery (clus321) 91
4.8 Grade B isochrone fit (clus59) 92
4.9 Grade C isochrone fit (clus235) 92

5.1 KDE density plot for Candidate 88 demonstrating robust central coordinate determination for an open cluster 98
5.2 Comparison of KDE and AMC centering methods for the complex star-forming region Candidate 380 (DR 21) 100
5.3 Multi-wavelength imaging of Candidate 380 (DR 21) explaining centering algorithm bias 101
5.4 Surface and radial density plots showing the output of the King, Plummer and EFF models 104
5.5 Comparison of UKIDSS GPS K-band and WISE multi-band imaging for Candidate 9 revealing an HII region 106

6.1 Gaia distance probability distribution functions (*PDFs*) for nearby and distant objects 114
6.2 Validation of Gaia proper motion and photogeometric distance measurements for the known open cluster NGC-6705 116

6.3 Gaia proper motion and photogeometric distance analysis for the open cluster NGC-2401 . . . 117
6.4 Gaia proper motion and photogeometric distance analysis for the new open cluster candidate 122 . . . 118
6.5 Gaia proper motion and photogeometric distance analysis for the emerging cluster candidate 211 . . . 119

7.1 Demonstration of the HDBSCAN algorithm on a sample dataset with overlapping clusters of varying densities . . . 123
7.2 Impact of the `min_samples` parameter on the sensitivity and smoothness of the recovered PDF . . . 124
7.3 Minimum spanning tree (*MST*) for the sample dataset illustrating density-based connections and mutual reachability distance . . . 126
7.4 Hierarchical dendrogram showing the initial single-linkage cluster tree derived from the minimum spanning tree . . . 127
7.5 Condensed cluster tree showing the influence of the `min_cluster_size` parameter on cluster stability . . . 128
7.6 Selection of stable clusters from the condensed tree hierarchy . . . 130
7.7 Final HDBSCAN cluster selection with membership probabilities and noise identification . . . 131
7.8 Example HDBSCAN output showing identified clusters in the field of the open cluster NGC-6846 . . . 133
7.9 Cross-match between HDBSCAN identified clusters and Bonatto members for the open cluster NGC-2401 . . . 134
7.10 Cross-match between HDBSCAN identified clusters and Bonatto members for the open cluster NGC-2262 . . . 135
7.11 Cross-match between HDBSCAN identified clusters and Bonatto members for Candidate 15 . . . 136
7.12 Candidate 29 - HDBSCAN-Bonatto Cross-match . . . 137
7.13 Candidate 86 - HDBSCAN-Bonatto Cross-match . . . 138

7.14 Candidate 105 - HDBSCAN-Bonatto Cross-match 139
7.15 Candidate 132 - HDBSCAN-Bonatto Cross-match 140
7.16 Candidate 30 - HDBSCAN-Bonatto Cross-match 141
7.17 Candidate 30 CMDs and CCDs - HDBSCAN and Bonatto 142
7.18 Candidate 5 - HDBSCAN-Bonatto Cross-match 143
7.19 Candidate 5 CMDs and CCDs - HDBSCAN and Bonatto 144
7.20 Candidate 108 - HDBSCAN-Bonatto Cross-match 145
7.21 Candidate 108 CMDs and CCDs - HDBSCAN and Bonatto 146

8.1 2x2 K-Band images of four previously unknown clusters with large WISE distances. 161

Chapter 1

Introduction

Star formation is one of the most fundamental processes in the universe and has been in action since the dawn of time. During their lifetime the resulting stars process hydrogen, helium and trace amounts of metals into all the heavier elements up to iron, with the elements beyond iron being created during the supernova death of the most massive stars. This enriched material becomes mixed with the vast quantities of hydrogen and other trace gases contained in the galactic disk and is recycled through generation after generation of stars. This continuing enrichment of the medium provides all the building blocks for the complex chemistry that eventually results in the evolution of life in the Universe as we see it today (Boesgaard and Steigman 1985 and references therein). Understanding the processes that lead to the formation of stars and their implications is of great interest and importance to us as we continue to learn about our place in the cosmos. Fortunately we are situated within a galaxy with much star formation activity, some of which are local enough to allow for detailed study.

Our Galaxy is a complex and dynamic environment. At the largest scale it consists of a relatively low density halo of older, metal-poor stars and small globular clusters with a radius of $\sim$ 20 kpc. These stars are kinematically distinct from the stars in the higher-density large-scale structures towards the plane of rotation. Within the halo, in the plane of rotation, is the Galactic disk containing $\sim$ 90% of the stellar population (Bland-Hawthorn and Gerhard 2016), it consists of both a thick element mainly composed of older, metal-poor stars that are intermediate between the disk and halo, and

a thin element that contains the structures we know as spiral arms, the molecular clouds of gas and dust, and the younger stellar population (Lin and Shu 1964). The disk has a radius of $\sim$ 15 kpc with a thick disk scale height of $\sim 1 - 1.5$kpc and a thin disk scale height of $\sim$ 300pc $-$ 1kpc. Finally, there is the central bulge of the Galaxy, an elliptical structure with a radius $\sim$ 1.5 kpc (Bland-Hawthorn and Gerhard 2016; Deason, Belokurov, and Sanders 2019). Star formation activity is primarily confined to the thin disk in the dense spiral arms, which contain the bulk of the molecular gas (Lin and Shu 1964 and subsequent publications), although a small portion ($\sim$ 10% of the total) is involved in star formation in a central molecular zone (CMZ) within the Galactic bulge (Immer et al. 2012). Approximately half of the volume of the Galaxy is filled with hot, ionised gas with a number density of approximately $n < 0.01 cm^{-3}$ and $T_K > 10^5 K$ while the other half contains some combination of warm ionised gas and warm neutral gas with number density of approximately $n \sim 0.1 - 1 cm^{-3}$ and $T_K >> 10^3 K$ (Cox 2005). The neutral gas partially segregates into warm and cold components, the cold having $n > 10 cm^{-3}$ and $T_K < 100 K$ (Kennicutt and Evans 2012).

Large-scale structures in the spiral arms are typically sheets and filaments of turbulent warm gas (Smith, Glover, and Klessen 2014; Smith, Glover, Klessen, and Fuller 2016; Cox 2005) and close to the mid-plane of the disk there are sub-sheets and filaments of turbulent cold dense material[1] which decrease in mean density with distance from the mid-plane and extend at least a kiloparsec above and below the plane. The remainder of the volume consists of lower density regions such as bubble interiors, cavities, and tunnels within the ISM formed by the intense winds from O and B class stars (HII regions) and supernovae with some, but not all, of these regions hot enough to be observable by their X-ray emission (Cox 2005; Kennicutt and Evans 2012; Blitz 1982; Blitz and Rosolowsky 2004). The presence of large regions of warm plasma within the Galaxy means that it is filled with strong but irregular magnetic fields which have an impact on the bulk processes and motion of material and is one of the causes of pressure waves and turbulence. The molecular clouds themselves are the coldest and densest

[1]The tendency for large clouds of self gravitating, turbulent, matter towards forming filaments is a universal phenomenon originally described by Lin and Shu (1964) and Springel (2005)

component of the ISM with n > 30cm^{-3} and $T_K \sim 10$K. It is the colder ($T_K < 10$) molecular clouds that are of interest for star formation, there are of the order of 2,000 GMCs formed by collisions between the many thousands of smaller clouds distributed within the spiral arms. The low temperature and relatively high density of these clouds makes them the only possible site for star formation activity.

Situated in the disk of our Galaxy we find that our view, in visible light at least, is limited when looking through the plane of the Galactic disk by the absorption and scattering effects of the giant molecular clouds (GMCs) of dust and gas. The dust comes from the many generations of stars that have been born, lived, died, and returned some of their heavy elements and energy to the Inter Stellar Medium (ISM) (Shu, Adams, and Lizano 1987). The advent of infrared astronomy has allowed us to see through this dust and gas and with high resolution instruments such as the UKIRT-WFCAM (United Kingdom InfraRed Telescope, Wide-Field CAMera) and VISTA (Visible and Infrared Survey Telescope for Astronomy), and the new generations of space based observatories such as WISE (Wide-field Infrared Survey Explorer), Spitzer SIRTF (Space InfraRed Telescope Facility) and the JWST (James Webb Space Telescope), we are now able to observe the formation process throughout most of its stages in some detail.

It is thought that the vast majority of stars form in clusters following the hierarchical gravitational collapse of GMCs predominantly in the Galactic plane (Lada and Lada 2003). Observing these young clusters in situ can reveal much about the process of cluster evolution, such as the detailed form of the initial mass function (IMF) where questions still remain. For instance, Lyo et al. (2006) found that the young η Chamaeleontis cluster shows an apparent lack of low mass stars and brown dwarfs. This despite a photometry completeness limit that was able to reveal low-mass members down to the brown dwarf and planet boundary of $\approx 13 M_{Jupiter}$.

This appears to contradict the established models and observations of the IMF and observing cluster structure across a large sample, in a variety of environments, can give a better statistical knowledge of the initial conditions of cluster formation, and give insight into mysteries of this kind (Allison, Goodwin, Parker, de Grijs, et al. 2009; Allison,

Goodwin, Parker, Zwart, et al. 2010). This is especially true for unusual clusters such as young, massive clusters (YMC) which are rare, but very important in Galactic terms because they host the massive stars that supply most of the UV flux that dominates stellar feedback into the ISM. In addition to this, we are interested in cluster location within the molecular filaments and their associated GMCs as this, in combination with tracers of imminent or recent star formation such as ultra-compact HII regions, dense clumps, methanol masers, etc., is an interesting probe of the propagation of star formation in these environments (Kennicutt and Evans 2012; Dirienzo et al. 2012; Peretto and Fuller 2009). For this reason there have been many surveys looking to identify young infrared (IR) star clusters in our Galaxy using data primarily from the WISE All-WISE (Camargo, Bica, and Bonatto 2015, Majaess 2012) , Spitzer GLIMPSE (Mercer and Clemens 2005), 2MASS All-Sky (Dutra et al. 2003; Bica et al. 2003; Froebrich, Scholz, and Raftery 2007; Froebrich and Schmeja 2010), UKIDSS GPS (Solin, Ukkonen, and Haikala 2012) and the Vista Variables in the Via Lactea (Borissova, Bonatto, et al. 2011; Borissova, Ivanov, et al. 2018) projects.

From these surveys a large number of new clusters have been identified but each of these projects had its limitations of sensitivity, resolution and coverage meaning that they are limited to specific depths in specific regions of the Galaxy. Camargo, Bica, and Bonatto (2015) used WISE data to identify 437 embedded clusters and groups in the $145° < l < 290°$ $-25° < b < 20°$ region. Dutra et al. (2003) identified 179 new IR clusters but only looked in the $230° < l < 350°$ region using the 2MASS survey which is limited to K-band sources of approximately magnitude $K = 14.3$ (Skrutskie et al. 2006 - see Chapter 2). While much work has been done to understand the detail of this process, and to survey the sites of star formation visible to us, there are still many questions to be fully answered and more data to be added to the catalogues of knowledge as well as new or improved techniques for analysing the data, with some of these discussed in more detail in section 1.3.

1.1 Gravitational Instability and Mechanisms of Cloud Collapse

1.1.1 Isothermal Collapse

The exteriors of the molecular clouds themselves are visible to us via the relatively strong 2.6 mm line of $^{12}C^{16}O$, although the clouds are optically thick to this wavelength. Viewing the internal structure requires the use of a less common isotope such as $^{13}C^{16}O$ which has a much lower column density, although this results in a much lower strength line. Observations using the radial velocity profile of the $^{13}C^{16}O$ line have proved valuable in making clear the far from uniform distribution of mass and kinetic energy within the clouds (Solomon and Rivolo 1989; Stahler and Palla 2008). The turbulence resulting from the assimilation of smaller clumps, large scale gravitational action, pressure from surrounding warmer material, shock fronts, star formation activity and radiation pressure all having an effect on the internal kinematic motion (e.g. Lada and Lada 2003; Kennicutt and Evans 2012) with the distribution of clump velocities seen to be randomly distributed around the mean velocity of the GMC itself. This supports the notion that the GMCs are built up from dense cores, in fact we find that in all complexes the mass distribution of the cores can be described using a combination of power laws. Olmi et al. (2018)) found that the clump mass function (CMF), in its most basic form, follows the simple power-law below:

$$\frac{\delta N}{\delta M} \propto M^{\alpha} \tag{1.1.1}$$

where $\frac{\delta N}{\delta M}$ is the number of clumps per unit mass interval, M is the clump mass and α is the power-law index, typically in the range $1.5 - 2.5$. Typically though, the CMF exhibits a break at a characteristic mass M_c, giving:

$$\frac{\delta N}{\delta M} \propto \begin{cases} M^{-\alpha_1}, & M < M_c \\ M^{-\alpha_2}, & M \geq M_c \end{cases} \tag{1.1.2}$$

where $\alpha_1 \sim 1.5$ and $\alpha_2 \sim 2.3$, which is similar in fact to the value derived by Salpeter for the slope of the IMF (see Salpeter 1955 and section 1.2). Reid and Wilson (2006) for

instance, observed that submillimeter clump CMFs in high-mass star-forming regions fit well to broken power laws, with steeper high-mass slopes, however, Smith, Clark, and Bonnell (2008) found that for low-mass clumps, especially in quiescent or early-stage clouds, the CMF may follow a log-normal distribution and this is often associated with turbulent fragmentation:

$$\frac{\delta N}{\delta \log M} \propto \exp\left(-\frac{(\log M - \log M_0)^2}{2\sigma^2}\right) \tag{1.1.3}$$

The same processes that result in the GMCs clumpy, filamentary structure also work on smaller scales creating mass segregation along the filaments themselves, resulting in a string of dense cores (McKee and Ostriker 2007; André et al. 2010; Men'shchikov et al. 2010). As these cores evolve over time internal turbulence and pressure from the surrounding material can increase the local density, causing the core to become gravitationally unstable should it exceed the local Jeans / Bonnor-Ebert mass.

It is interesting to note that the form of the Core Mass Function (CMF) across a filamentary region with a number of cores in the process of gravitational collapse appears to be very similar to that of the Stellar Initial Mass Function (André et al. 2010). It is likely that the distribution of core masses directly influences the distribution of stellar / cluster masses and this is an active area of research (Könyves et al. 2015; Pezzuto et al. 2021).

For large, unbounded clouds where the dominant forces are internal the process of gravitational instability and collapse is best described by the *Jeans Mass*, M_J, named after James Jeans who originally derived it (see Jeans 1902). For a dense core or a compact volume within a star-forming region which is pressure-confined by its surrounding material, it is more appropriate to use the *Bonnor-Ebert Mass*, M_BE as it explicitly includes the influence of the surrounding cloud (Bonnor 1956, Ebert 1955). The two approaches are essentially the same in that they both describe how an isothermal system behaves in terms of the competing forces of gravity and internal pressure, and both provide a way to determine the threshold for gravitational collapse. The choice of method depends on the specifics of the case under consideration, with Jeans Mass being suited more to large, essentially unbounded, clouds and Bonnor-Ebert being suited more

to filaments and cores.

As an example we can consider the case of a large, unbounded, optically thin, isothermal cloud in which case the gravitational stability in any given region can be described by the local Jeans Mass, M_J, thus:

$$M_J = \frac{9}{4} \times \left(\frac{1}{2\pi n}\right)^{\frac{1}{2}} \times \frac{1}{m^2} \times \left(\frac{kT}{G}\right)^{\frac{3}{2}} \tag{1.1.4}$$

where n is the particle number density, m is the mass of the average particle and T is the gas temperature.

If the mass in the any region exceeds this value the cloud becomes unstable to its own gravity and begins to collapse. The size of the stable region in which gravitational attraction is balanced by internal pressure (assuming a quiescent state) is defined by the *Jeans Length*, L_J, where:

$$L_J = \sqrt{\frac{\pi c_g^2}{G\rho}} \tag{1.1.5}$$

and c_g is the speed of sound in the gas, which is dependent on temperature and molecular weight of the gas molecules, G is the gravitational constant and ρ is the density of the gas.

During condensation of the region its angular momentum is conserved causing an increase in turbulence that can lead to fragmentation. These smaller regions can then sometimes exceed the local Jeans Mass and begin to collapse under their own self gravity (McKee and Ostriker 2007; Lada and Lada 2003; André et al. 2010). This process is iterative and leads to the hierarchical fragmentation of large clouds down to a lower limit of approximately $10M_{Jupiter}$ (see Silk, Silk, and J. 1977; Smith, Glover, and Klessen 2014; Smith, Glover, Klessen, and Fuller 2016; Adams and Myers 2001). We will consider the formation of stellar groups, or clusters, in the next section but first it is useful to review the process of individual star formation and the observational characteristics thereof.

Taking now the case of a higher density region bounded by a less dense region which is exerting a confining pressure, we would use the Bonnor-Ebert equation to define the

local threshold for gravitational instability thus:

$$M_{BE} \simeq 1.18 \frac{c_s^4}{\sqrt{G^3 P_{ext}}} \tag{1.1.6}$$

where G is the gravitational constant, P_{ext} is the external pressure confining the gas sphere and c_s is the isothermal sound speed and characterizes the thermal motions in the gas, defined by:

$$c_s = \sqrt{\frac{k_B T}{\mu m_H}} \tag{1.1.7}$$

where k_B is the Boltzmann constant, T is the temperature, μ is the mean molecular weight, and m_H is the mass of a hydrogen atom.

Once a dense core has passed the threshold for gravitational instability it begins its collapse as an isothermal gas in free-fall. The free-fall time, t_{ff}, gives the time it would take a pressure-less sphere of gas to collapse under its own gravity and is useful for describing this initial stage of collapse:

$$t_{ff} = \sqrt{\frac{3\pi}{32 G \rho}} \tag{1.1.8}$$

where G is the gravitational constant and ρ is the density of the gas. Typical number densities for a core about to form a star are in the range of $n \approx 10^5$ cm^{-3} and using a mean molecular weight of 2.3 (which includes both the hydrogen and trace molecules) we can see that t_{ff} is approximately 10^5 years (for an idealised system of average mass). The conversion of gravitational potential energy into thermal motion results in the collisional excitation of the molecules, which then relax and radiate the energy as molecular-line photons. The gas at this stage is optically thin so this process is very efficient allowing a rapid phase of collapse with a relatively stable temperature, typically between 10 and 30 K (see Elia et al. 2017). In the early stages it is a combination of trace molecules such as carbon monoxide (CO) or formaldehyde (H_2CO) and continuum emission from the dust itself that provide much of the radiative cooling as the lowest molecular transitions of hydrogen molecules occur at temperatures of the order of 100 K (Levy and Lunine 1993; Evans 1999; Stahler and Palla 2008). The cooling efficiency at this stage is still far higher

than the energy generated by the increase in kinetic energy, such that temperatures do not rise significantly during this phase, and this remains true until the gas becomes opaque and has exhausted the dissociation and ionisation energies of all the molecules.

1.1.2 Formation of the First Hydrostatic Core

When the central region's density reaches approximately 10^{13} g cm^{-3} the gas becomes optically thick, radiative cooling becomes much less efficient and it becomes adiabatic, at which point thermal pressure starts to build up. As the internal thermal pressure increases it eventually becomes high enough that it balances the gravitational infall and halts the collapse. At this point the first hydrostatic core is formed, a temporary, quasi-stable object with a typical radius of about 5 AU (7.48×10^8 km) and an initial mass around 5 $M_{Jupiter}$ masses (9.5×10^{27} kg) see Bate (2011), Levy and Lunine (1993), and Stahler and Palla (2008) et al.

Observationally these objects are hard to detect due to their short lifetime (of the order of 1000 years), low surface temperature (of the order of 100 K), and correspondingly low luminosity in the far-infrared/sub-millimetre wavelengths (Young et al. 2018; Bate 2011). In fact it is doubtful that any have yet been observed, however, the cooler spectral energy distribution of these objects should be useful in distinguishing them from other young stellar objects.

As the core continues to accrete mass the temperature rises. When it reaches approximately 2000 K, molecular hydrogen (H_2) begins to dissociate - a process that absorbs energy and reduces the gas pressure, leading to a second phase of gravitationally driven collapse.

1.1.3 Second Collapse, Protostar Formation and Evolution

The onset of (H_2) dissociation triggers a rapid second collapse forming the first protostar, a Class 0 Young Stellar Object (YSO - see Fig. 1.1). The Class 0 YSO is the first observationally detectable stage of protostar evolution, at this stage the object is still embedded within a dense envelope of gas and dust and is still accreting material at a high

rate from the surrounding cloud, whilst simultaneously driving large, highly collimated bipolar outflows. This is the stage during which the protostar will gain the majority of its final mass. See Andre, Ward-Thompson, and Barsony (1993) and Myers et al. (1998).

It is important to keep in mind the influence and contribution of the clouds angular momentum and magnetic fields to its evolution. The conservation of angular momentum causes a rotational spin-up during collapse which can in fact slow or even prevent gravitational collapse by preventing the radial accretion of material. The angular momentum needs to be removed or redistributed into the cloud to allow collapse to proceed and this is where the magnetic fields within the cloud become important.

Magnetic fields are integral during the collapse of molecular cloud cores. They facilitate the transfer of angular momentum away from the infalling material through a process known as magnetic braking. This mechanism allows the material to lose angular momentum efficiently, promoting accretion onto the protostar and aiding in the formation of a circumstellar disk. However, overly efficient magnetic braking can suppress disk formation by removing too much angular momentum (known as the "magnetic braking catastrophe"). See Mellon and Li (2007), Ray and Ferreira (2021) and Joos, Hennebelle, and Ciardi (2012) for instance.

The protostar is now undergoing Kelvin-Helmholtz contraction, so although the energy released by further compression of the central region increases the temperature, resisting further contraction, the outer layers are still able to radiate energy efficiently and allowing continued collapse. This process slows the rate of contraction and outputs significant radiation to the enveloping dust and gas, which is absorbed and re-radiated in the far-infrared and submillimeter bands.

This emission gives the Class 0 protostar its characteristic cold black-body spectral energy distribution (SED) with no emission below 10 μm, in fact the emission is entirely that of warmed dust despite the temperature of the young, shrouded, protostar. See the first row in Fig. 1.1 for a typical SED and physical properties. The Class 0 phase is relatively short-lived, lasting approximately $2 - 6 \times 10^4$ years. Towards the end of this phase the protostar eventually disperses the surrounding envelope to reveal a more

evolved Class I protostar.

The Class I protostar is still enveloped by substantial amounts of gas and dust in rotating circumstellar envelope. At this stage the angular momentum has spun-up the protostar and the infalling material is forming an accretion disk while a combination of the angular momentum and magnetic fields drive the formation of bipolar outflows and collimated jets via a Blandford & Payne-type magneto-centrifugal mechanism (Blandford and Payne 1982).

Angular momentum is also being redistributed outward via Alfvén waves. These waves are magnetohydrodynamic waves propagating along magnetic field lines generated by turbulent motions and interactions within the accretion disk and the protostar's magnetosphere. The waves transfer angular momentum from the inner regions of the disk outward and the reduction in momentum in the inner disk facilitates accretion onto the protostar (Jatenco-Pereira 2013).

At this stage the envelope mass is comparable to or exceeds the mass of the central protostar. Despite the fact that they have cleared a good deal of the enshrouding material the protostar still exhibits high accretion rates, leading to significant luminosity. Accretion shocks on the protostellar surface release energy, heating the surrounding material and contributing to the observed SED which peaks in the infrared region, reflecting the thermal emission from the warm dust in the envelope and disk. The presence of a significant infrared excess distinguishes them from more evolved stars (see Fig. 1.1). This strong infrared emission often includes features corresponding to silicate dust (Andre, Ward-Thompson, and Barsony 1993; Myers et al. 1998).

Many Class I protostars exhibit bipolar outflows and jets which are observable through emission lines such as Hα and forbidden lines such as [O I] and [S II]. These features indicate active mass loss and are one of the main ways the protostar interacts with the surrounding medium, although the mass loss rate is only about 10% of the accretion rate through all of the YSO phases. The balance between accretion and mass loss during this phase impacts the final mass of the star and the evolution of the surrounding disk, potentially leading to planet formation. Variations in luminosity at this

stage are still common and result from changes in accretion rates and ongoing structural alterations in the circumstellar material. The Class I phase of evolution lasts for between $\approx 10^5 \sim 10^6$ years; see e.g. Evans et al. (2009) and Dunham et al. (2014).

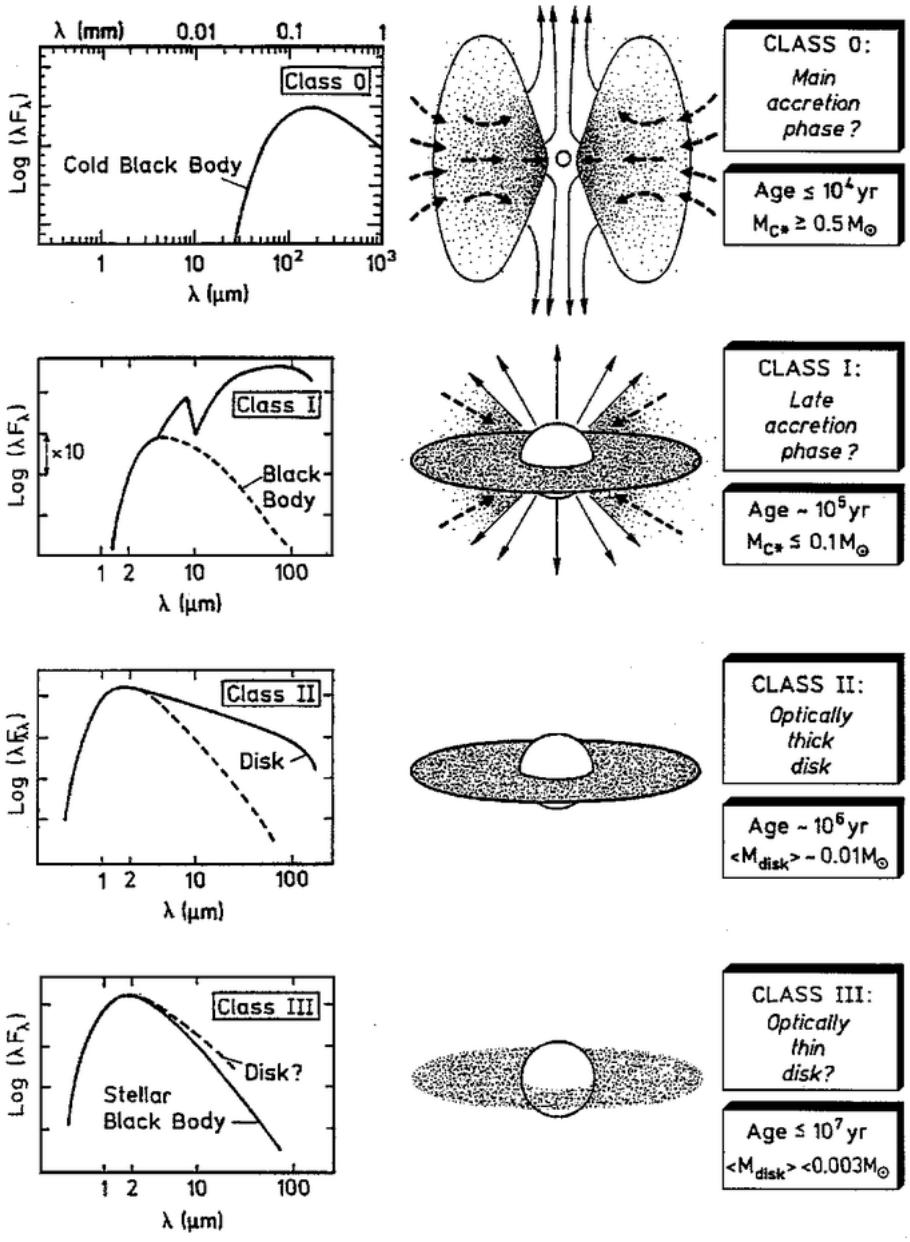


Figure 1.1 Diagram showing the different phases (classes) of protostellar (YSO) evolution. The classification scheme was developed over many years by Lada, Wilking, André, et al. (Lada and Wilking 1984; Lada 1987; André et al. 2010) and is based on the characteristic spectral energy distribution of the objects. As the YSO evolves and its overall temperature increases its emission spectrum shifts to shorter wavelengths. Image credit: Palla (1996).

1.1.4 Young Stellar Object Evolution onto the Main Sequence

The gradual dissipation of the protostellar envelope and the emergence of a protoplanetary disk is what drives the transition of the class I protostar to a class II YSO. Class I protostars as described above are still deeply embedded within dense envelopes of gas and dust, which obscure the central protostar. As evolution proceeds, these envelopes dissipate as a result of processes such as accretion onto the protostar, outflows, and radiative heating, which over time lead to increased visibility of the central object. With the clearing of the envelope, the circumstellar material flattens into a rotating disk structure. This protoplanetary disk becomes the dominant feature in Class II YSOs and is the site where planet formation may occur.

Observationally this is characterised by a shift in the objects SED from the significant infrared excess due to the thermal emission from the envelope of the Class I protostar to an SED that peaks at shorter wavelengths, reflecting the reduced contribution from circumstellar material and the increasing contribution of the stellar photosphere (see fig. 1.1). The overall infrared emission decreases and this marks the transition from Class I to Class II. This change is indicative of the reduced optical depth and the emergence of the star and disk system. The Class II YSO can still exhibit bipolar outflows and jets, however, these features tend to weaken as the system transitions to Class III, with the corresponding decline in accretion activity.

The Class II YSO stage corresponds to classical T-Tauri stars (CTTS) for low-mass objects with $M < 2M_{\odot}$ and Herbig Ae/Be stars for intermediate-mass objects with $2 \leq M_{\odot} \leq 10$. They still have a substantial circumstellar disk composed of gas and dust which is optically thick and plays a crucial role in ongoing accretion processes and potential planet formation. The SEDs exhibit infrared excess emission indicative of thermal radiation from the circumstellar disk and it is this excess that distinguishes them from main-sequence stars, whose SEDs peak at shorter wavelengths. Accretion from the disk onto the central star continues during the Class II phase, albeit at reduced rates compared to earlier stages. This accretion manifests as variability in luminosity and emission lines in the star's spectrum and also results in movement around the infrared

colour-magnitude or colour-colour diagrams.

In NIR colour-colour diagrams (CCDs) main-sequence stars and giants typically occupy distinct regions (see Fig. 1.2). CTTSs, however, exhibit excess emission in the infrared due to the thermal radiation from their circumstellar disks. This excess causes them to deviate from the intrinsic stellar loci, forming a distinct sequence known as the *CTTS Locus*. This locus extends from the main stellar locus toward redder H-K colours, reflecting the contribution of disk emission. See Figure 1.3, Greene and Meyer (1995) and Meyer, Calvet, and Hillenbrand (1997).

Class II objects can also still exhibit bipolar outflows and jets. The YSO typically stays in this phase for between ~ 1–5×10^6 yr. See e.g. Haisch, Lada, and Lada (2001) and Fedele et al. (2009). The transition from the Class II to the Class III YSO happens when the circumstellar disk dissipates and the star progresses toward the main sequence. This transition is observed in both low-mass stars ($M < 2M_\odot$), such as Classical T-Tauri Stars (CTTS), and intermediate-mass stars $2 \leq M_\odot \leq 10$, like Herbig Ae/Be stars. The transition is the result of the gradual clearing of the circumstellar disk through accretion onto the star, planet formation, or dispersal by stellar winds and this results in a reduction in the infrared excess as the disk dissipates and the accretion rate declines (see the reduction in the infrared emission above $\sim 10, \mu$m in the bottom row of fig. 1.1). This also leads to weak or absent emission lines, such as those seen in Weak-line T-Tauri Stars (WTTS). The Class III phase typically lasts between ~ 5–10×10^6 yr, although high mass stars with ($M > 8M_\odot$) evolve much faster, often reaching the main sequence in less than 1×10^6 yr (Lada 1987; Andre, Ward-Thompson, and Barsony 1993; Myers et al. 1998; Bate 2011 etc).

For low and intermediate mass YSOs the transition from Class III YSO to main-sequence star marks the final stage of pre-main-sequence evolution. During this phase, the star's internal structure, energy generation and observable properties undergo significant changes leading to stable hydrogen fusion in the core. This then marks the start of the main sequence life of the star. The energy source changes from Kelvin-Helmholtz contraction to hydrogen fusion (from either the PP chain for stars with

$M < 2M_{\odot}$ or CNO cycle for stars with $M > 2M_{\odot}$) which stabilises the contraction. The infrared excess and emission lines disappear as the disk is dissipated and accretion stops.

The evolutionary pathways of YSOs of various mass are shown in the Hertzsprung-Russell (HR) diagram in Figure 1.5. As can be seen in the figure, the higher mass YSOs ($M > 6M_{\odot}$) evolve almost straight onto the main sequence and maintain an almost constant luminosity, whereas the intermediate mass YSOs ($2 \leq M_{\odot} \leq 6$) start on the Henyay track (of horizontal movement to higher T_{eff}) and move towards the main sequence. The lower mass YSOs ($M < 2M_{\odot}$) initially follow the vertically descending Hayashi tracks of constant T_{eff} before turning to follow the Henyay track onto the main sequence, although those below about $M < 0.5M_{\odot}$ never make it to the Henyay track and descend straight down the Hayashi track to the main sequence (see Lada and Lada 2003; Myers et al. 1998; Stahler and Palla 2008).

As described above, a key observational characteristic of protostars and YSOs is the large degree of variability in their infrared luminosity and reddening during their evolution. Observationally this manifests as a wide spread of colour excess values in both the colour-colour and colour-mag diagrams and occurs on timescales as diverse as days to thousands of years (Myers et al. 1998; Lada 1987). This variability is a result of changes in the rate of mass accretion, clumpy in- or out-flows, disk warping and precession as well as intrinsic structural events during evolution of the protostar itself that affect its luminosity. Figure 1.2 shows the theoretical relative colours of main-sequence stars and giants, while Figure 1.4 shows the actual distribution for a small, local cluster. The majority of the stars in this example lie along the main sequence tracks as expected, but a number show a degree of reddening that corresponds to a displacement along the reddening vector of $A_V \lesssim 5$ mag. In addition to this however, a number of stars exhibit even greater displacement along the reddening vector whilst some are below the dashed line, showing an infrared-excess typically associated with young, embedded protostars or YSOs whose emission has been shifted to longer wavelengths by the enveloping dust and gas. Figure 1.6 shows examples of high and low extinction fields from the UKIDSS survey demonstrating the significant reddening effect

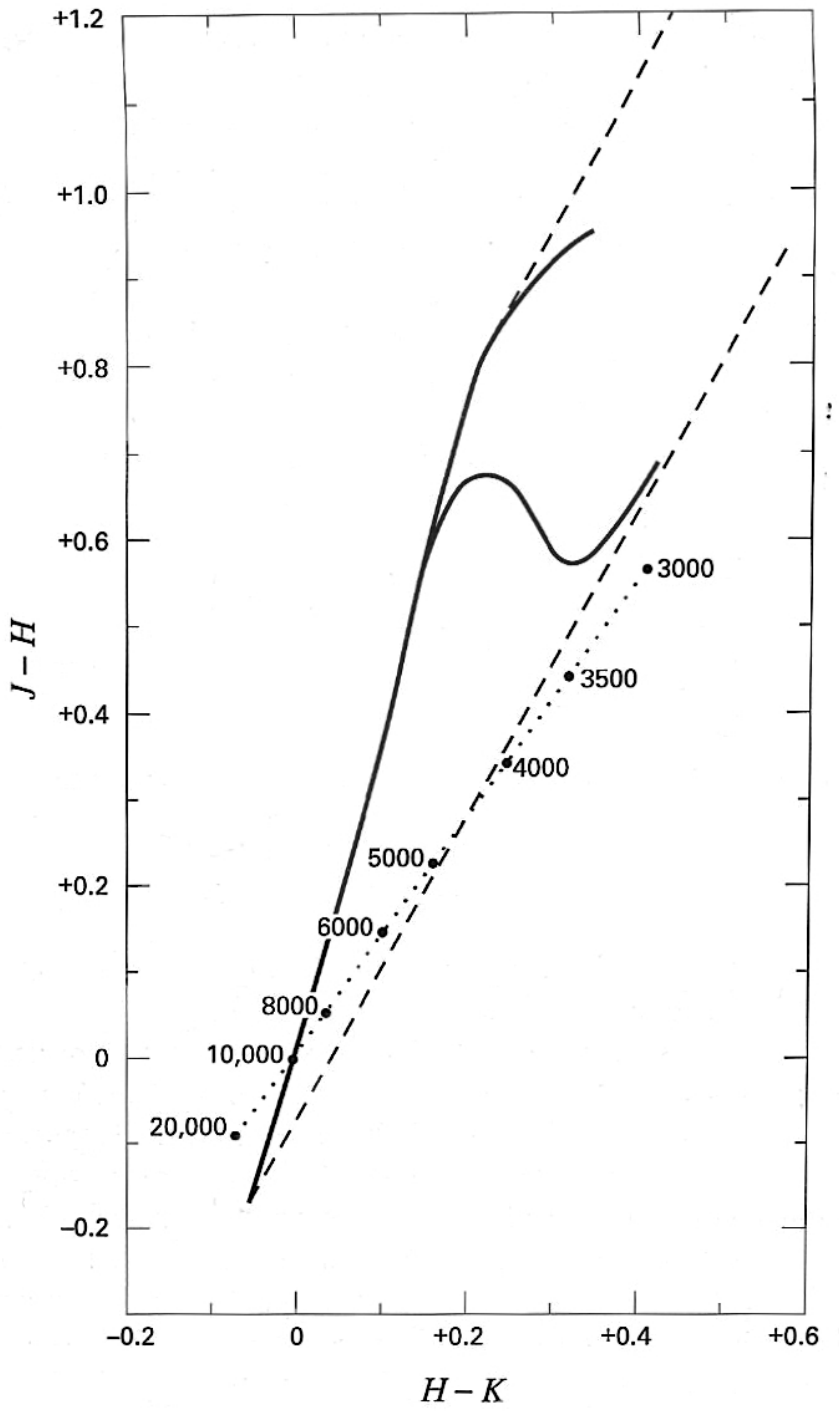
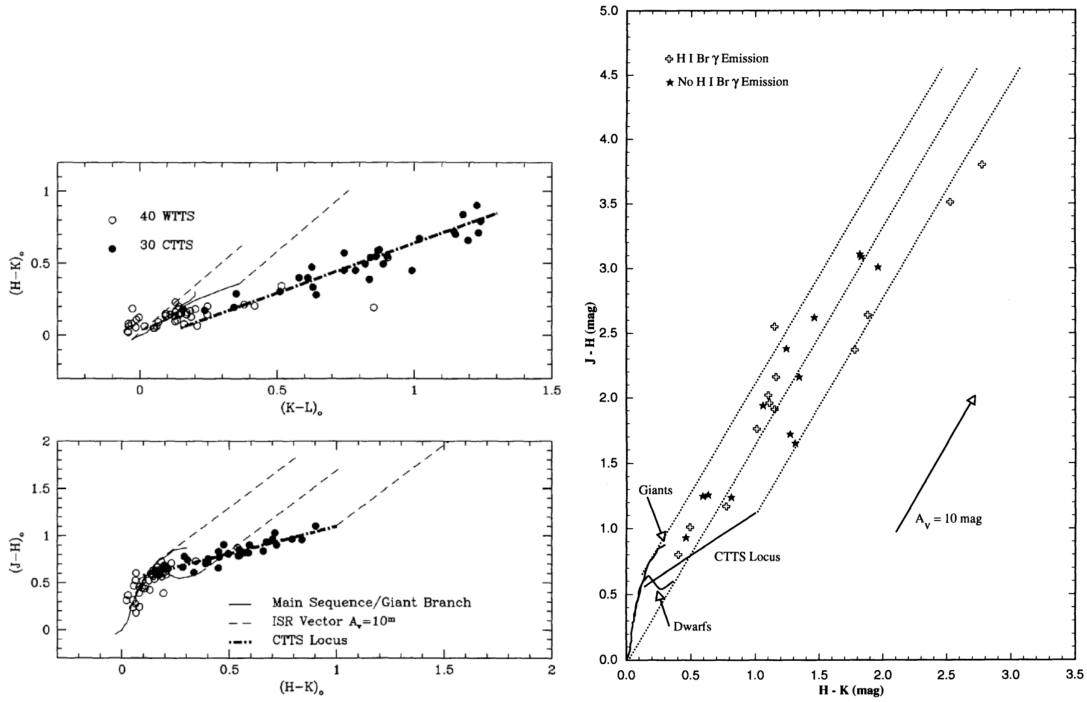


Figure 1.2 Near-infrared colour-colour diagram. The $J - H$ and $H - K$ colour indices are displayed as the vertical and horizontal axis respectively. The solid curve shows the relation between these for main-sequence stars (lower branch) and giants (upper branch). The dotted curve shows the colours of blackbody spectra at the indicated temperatures. Straight dashed lines indicate the relative colour changes due to interstellar reddening. Image and caption credit: Stahler and Palla (2008).



(a) Image Credit: Meyer, Calvet, and Hillenbrand (1997) **(b)** Image Credit: Greene and Meyer (1995)

Figure 1.3 Panel *a* above shows de-reddened colour-colour diagrams for weak-line T-Tauri stars (WTTS), top, and classical T-Tauri stars (CTTS), bottom, from Taurus-Auriga showing clear loci for both populations. The magnitude and slope of the interstellar reddening vector was taken from Cohen and Kuhi (1979) for the CIT photometric system. The intrinsic colours of giant and dwarf stars were taken from Bessell et al. (1988) and transformed into the CIT system. Panel *b* shows a sample of stars from ρ Ophiuchi (L1688) with the theoretical dwarf and giant branches and the CTTS locus marked. The reddening vector of $A_V = 10$ Mag has been added. Sources within or to the left of the reddened dwarf and giant loci have small near-IR excess whilst those to the right have large excesses, well above those one would expect from a stellar photosphere.

of dust in some fields (Lawrence and Warren 2007). This is also described in Figure 1.3, Greene and Meyer (1995) and Meyer, Calvet, and Hillenbrand (1997).

Whilst the excessive reddening and infrared colour-excess of YSOs is helpful for identifying young, embedded clusters it can be ncomplicated by the distribution of dense gas and dust in the vicinity of a cluster or star-forming region (SFR). Clumpy, dense gas in these situations can exhibit significant variability in optical depth, even at near-infrared wavelengths, which causes differential extinction and reddening across the cluster area and along the line of sight. This differential extinction leads to colour discrepancies among stars that are intrinsically similar, complicating the analysis of stellar populations

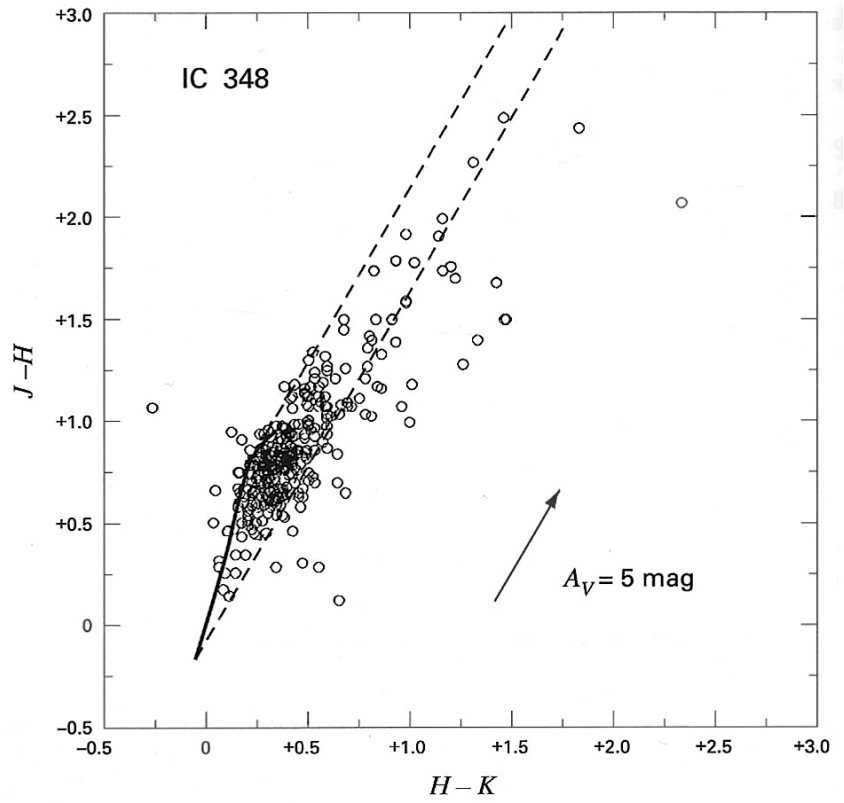


Figure 1.4 Near-infrared colour-colour diagram for IC 348, a compact cluster located in the Perseus Molecular Cloud, at a distance of 320 pc. This system contains numerous optically visible stars but many more are still embedded. Of the 342 sources plotted, comparison with adjacent fields indicates that about 60% should be cluster members. The majority of the stars in the diagram have near-infrared colours that are close to, but slightly above, the main-sequence curve, with a displacement along the reddening vector that corresponds to $A_V \lesssim 5$ mag. A smaller number of sources have significantly larger A_V, while some 20% lie well outside the dashed lines altogether. These latter stars, which are redder in $H - K$ than would be expected given their $J - H$ colours, are said to exhibit an *infrared excess*. Image and caption credit: Stahler and Palla 2008.

by distorting the shape of the infrared colour magnitude diagram or colour-colour diagram and produces a spread toward greater values of $J - H$ or $H - K$ (Arellano Ferro et al. 2023; Platais et al. 2012). Figure 1.7 is a CMD showing young stars in the Chandra Orion Ultra-deep Project (COUP) field. This field is relatively close and located off the Galactic plane toward the outer galaxy meaning it has minimal line-of-sight extinction and shows the differential reddening effect of the cluster's dust and gas cloud, as well as the infrared excess of the young stars.

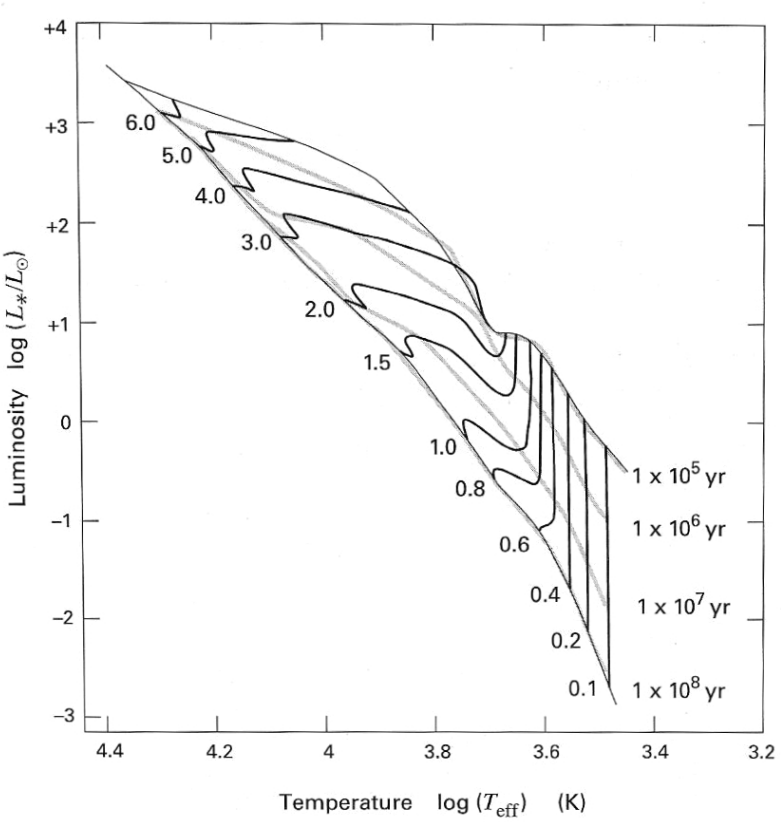


Figure 1.5 Pre-main-sequence evolutionary tracks. Each track is labelled by the stellar mass in units of $M_\odot$. The grey curves are isochrones, labelled in *yr*. The $t = 0$ isochrone coincides with the birth-line, the lighter solid curve at the top. Note that the $t = 1 \times 10^8$ yr isochrone nearly matches the ZAMS, the lighter solid curve at the bottom. Image credit: Stahler and Palla (2008).

Another common issue arises from the fact that a significant proportion of young stars, especially where $M > M_\odot$, are born in binary systems. In this case we have stars that appear to be brighter than they should be given their spectral class, although this depends on the mass ratio, separation and evolutionary stage. In the case of spectroscopic binaries the apparent excess luminosity means that these stars form a 'binary-sequence' in HR or NIR CMDs, appearing above the theoretical main-sequence as shown in Figure 1.8 (Hurley and Tout 1998; Straižys and Lazauskaitė 2009).

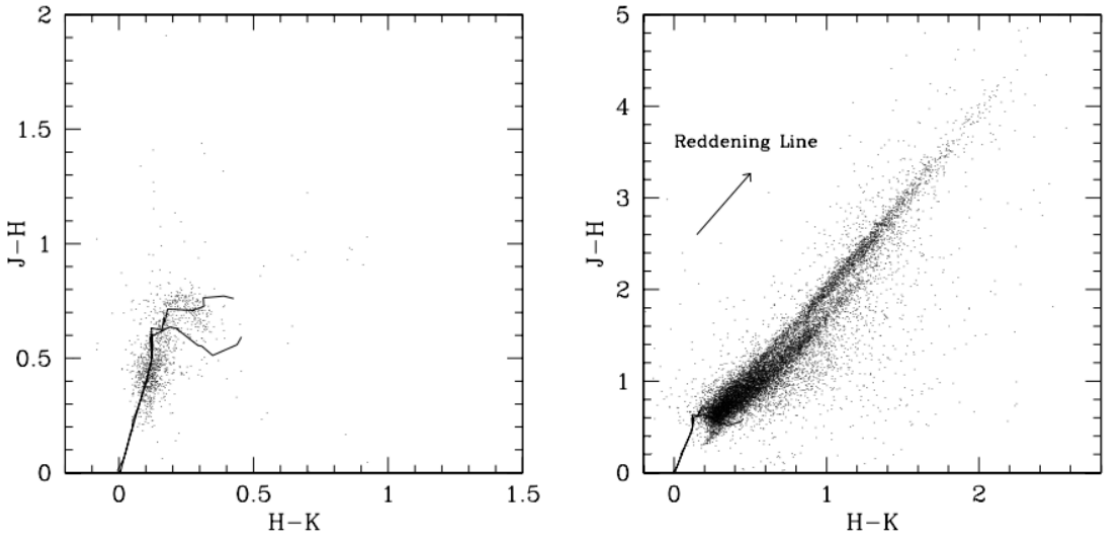


Figure 1.6 Left hand panel - an uncrowded UKIDSS field with low extinction at l = 171.2, b = 4.8 showing 1236 sources. Right hand panel - a crowded UKIDSS field with high extinction at l = 15.0, b = 0.0, showing 12954 sources. In each panel the lower curves shows the locus of the main sequence, and the upper curve shows the locus of luminosity class III giants. Even in the crowded field the photometric precision is good enough to show the 0.25 mag vertical separation between the reddened giant sequence and the reddened main sequence stars. An approximate reddening vector is plotted in the right hand panel. Some curvature of the reddening sequences is seen in the data, which is attributed to the change in the effective bandpasses at high extinction. Image and caption credit: Lawrence and Warren (2007).

1.2 Cluster Formation

As previously described the vast majority of stars are born in clusters, or the sparser associations, that form during the hierarchical fragmentation and gravitational collapse of Galactic GMCs. Stars do not form in a purely random manner; within a group or cluster there are characteristic features to the process that apply to all so-far observed instances of star formation within the Galaxy. Principally, this is the tendency of collapsing clouds to produce stars with a characteristic distribution of masses, known as the initial mass function (IMF). This mathematical form of the IMF was first proposed by Salpeter (1955), as:

$$\Phi \log m = \frac{\partial N}{\partial \log m} \propto m^{-\Gamma} \tag{1.2.9}$$

This introduced the IMF as a power law function describing the number of stars

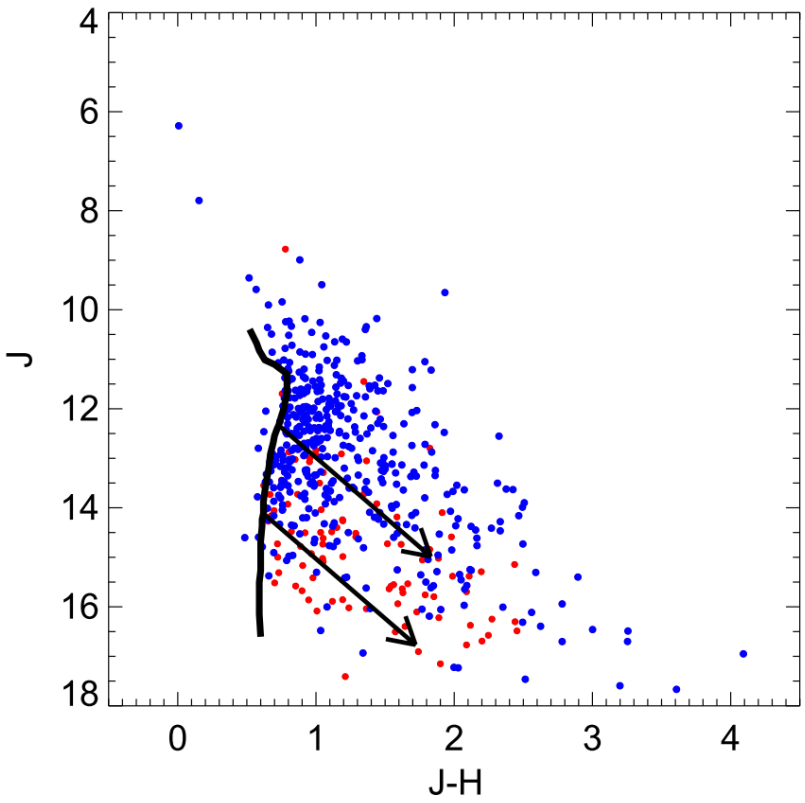


Figure 1.7 NIR CMD showing young stars detected by Spitzer Space Telescope (red points) in the Chandra Orion Ultradeep Project (COUP) field (blue points). The black vertical curve is the 1 Myr isochrone from Baraffe et al. 1998 and the two black arrows are extinction vectors extending to $1A_k$ for 1 Myr stars with masses of $0.25M_\odot$ and $0.08M_\odot$. Although this diagram is incomplete for very low mass stellar and sub-stellar members $< 0.25M_\odot$ it clearly shows how the infrared excess of these young stars presents in the CMD. Image credit Megeath et al. (2016).

formed for any given mass range. This formulation has a single logarithmic slope of $\Gamma = 1.35$ and this formulation is typically known as the Salpeter IMF. It became clear that the IMF was more complex than a single power law, and work done by Kroupa, Tout, and Gilmore (1993) showed that the IMF was better described by a combination of several power laws, each describing a particular mass range. Typically this relates to fewer high-mass stars and many more low-mass stars than would be predicted by the Salpeter IMF. These segmented power law IMFs are generally referred to as Kroupa IMFs. Chabrier (2003) reviewed many observational constraints and formulated an eight parameter version of the IMF using the central limit theorem. This is shown below in Figure 1.9.

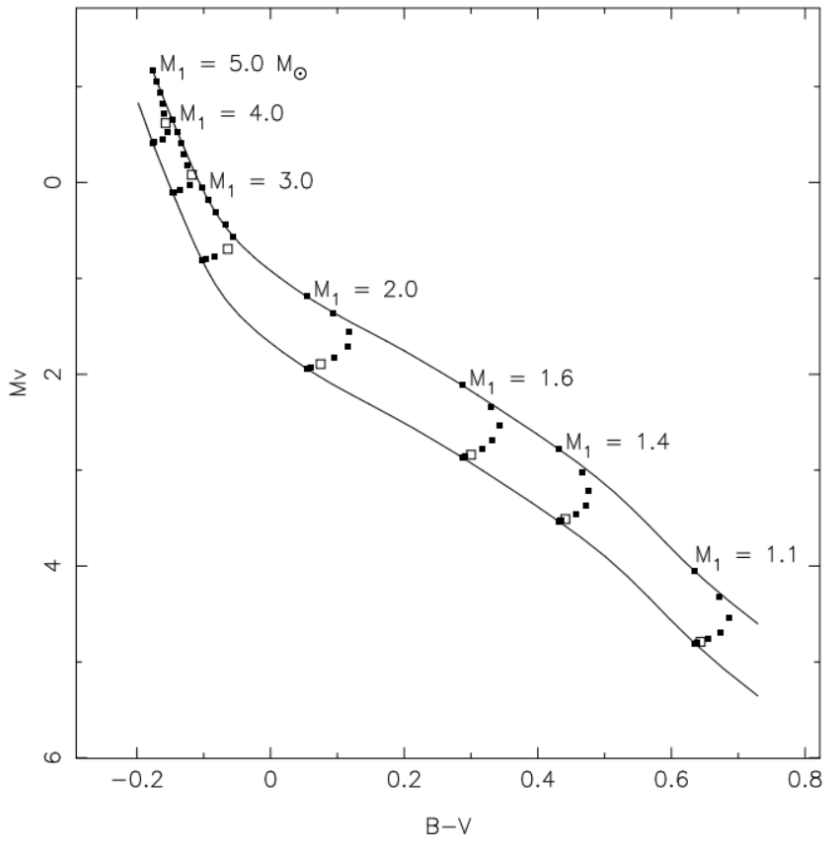


Figure 1.8 Theoretical zero-age single star and equal-mass binary main sequences covering a range of M_V . Also plotted for a range of primary masses, M_1, are binary points with $M_2 = qM_1$ where q ranges from $1.0 \rightarrow 0.0$ in increments of 0.1. The point at $q = 0.5$ is an open square. Image and caption credit: Hurley and Tout (1998).

This is an ongoing area of research, with the number of stars forming in these regions and the distribution of their masses being an area of great interest. At this time the IMF is not known to depend on the mass of the initial cloud, apart from the obvious upper limit on possible stellar masses placed by the total cloud mass and formation efficiency. It is possible however that the mass of the parent cloud has an influence on the mass distribution of the stars formed but this has not been observed yet. This is discussed in Liptai et al. (2017), Peretto and Fuller (2009), Offner et al. (2013), and Bastian, Covey, and Meyer (2010)) and references therein. In general a large proportion of work on the IMF has only measured it down to 1 $M_\odot$. There are also outstanding questions with regard to the propagation of star formation across GMCs (Bastian, Covey, and Meyer 2010)), tracers of imminent star formation such as dense clumps and water or methanol

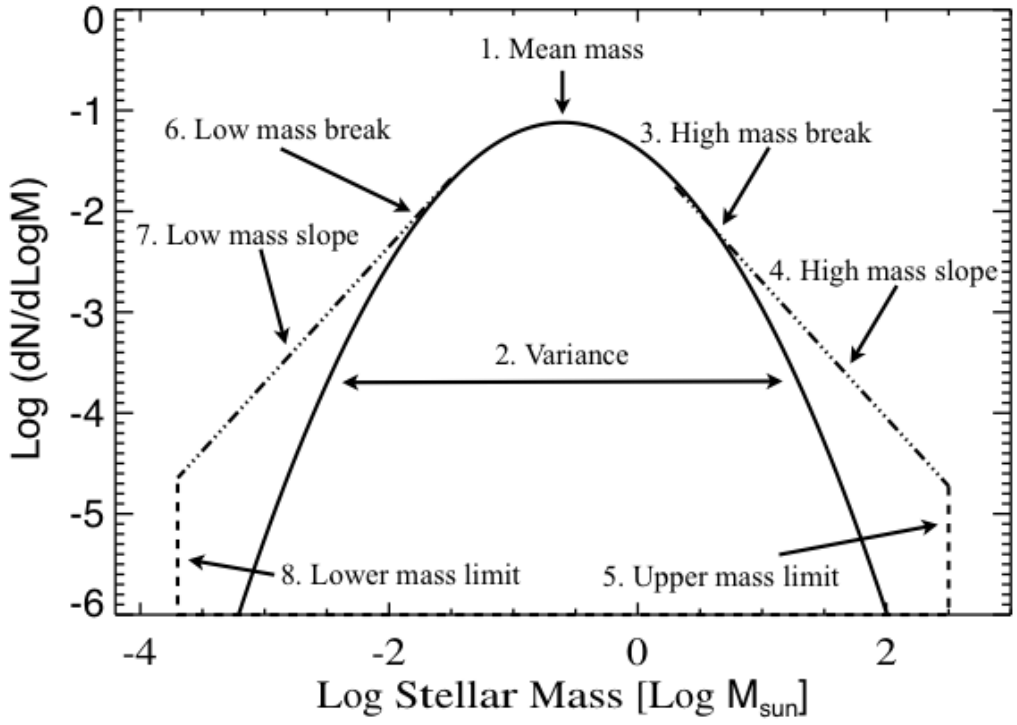


Figure 1.9 The 8 parameter version of the Chabrier IMF. Diagram taken from Bastian, Covey, and Meyer 2010.

masers (Peretto and Fuller 2010; Peretto, Fuller, et al. 2013; Caswell and Breen 2010). The greater sensitivity and resolution of the UKIDSS GPS survey (see Chapter 2) offers the possibility of extending IMF measurements down below 1 $M_{\odot}$ to test these hypotheses.

How star formation within a particular cluster proceeds is governed by a number of factors, such as the initial mass, density, clumpiness and turbulence of the parent cloud and these will influence the survival time and evolutionary path the cluster follows. Cluster isochrones are theoretical curves plotted on the HRD or CMD representing stars of varying masses but identical ages within a star cluster. These curves allow us to determine the age and evolutionary stage of a cluster by fitting the observed stellar data to the isochrones (Bressan et al. 2012; Choi et al. 2016). Figure 1.10 below shows examples of infrared isochrones from two of the leading theoretical models, PARSEC (Nguyen et al. 2022) and MIST (Dotter 2016; Choi et al. 2016). It can be seen that with increasing age the main-sequence turn-off moves from the upper left-hand corner of the plot

downwards, and to the right. This is because the more massive, more luminous, stars that inhabit the upper left of the plot are the first to turn off the main sequence due to their rapid evolution. The rate of change of the position of the turn off slows dramatically as we look at longer timescales due to the much longer lifetimes of lower mass stars. Figure 1.11 shows examples for two well known clusters.

Equation 1.2.10 gives the relationship between luminosity, mass and main-sequence lifetime of a star, in this case normalised to solar parameters. Because L_* increases very rapidly with M_*, t_{ms} falls correspondingly fast. Our Sun is approximately 4.6×10^9 years old which is about half way through its t_{ms} of about 1×10^{10} years whilst an O-type star of $40M_\odot$ has a t_{ms} of only around 4.3×10^6 years (Stahler and Palla 2008).

$$\begin{aligned} t_{ms} &\approx 5 \times 10^{-4} \frac{M_* c^2}{L_*} \\ &= 1 \times 10^{10} \frac{(M_*/M_\odot)}{(L_*/L_\odot)} \text{yr} \end{aligned} \tag{1.2.10}$$

The evolution of the cluster begins with an embedded phase in the same way as individual stars as described above. This phase typically lasts ≤ 1 Myr with the cluster beginning to form within the dense molecular cloud. Emission shows significant infrared excess and the cluster is dominated by Class 0 and Class I protostars and young T-Tauri stars with most of the stars yet to reach the main-sequence (Lada and Lada 2003; Megeath et al. 2016).

Between about 1–10 Myr of age the young stars begin core Hydrogen burning and the cluster enters the pre-main-sequence phase. The stars are contracting toward the zero age main-sequence (ZAMS) following the Hayashi and Henyay tracks (see fig 1.5. The massive O and B type stars evolve faster onto the main-sequence while the lower mass PMS stars form a broad sequence above the ZAMS, and a distinct main-sequence turn-on point can be seen in the distribution. Those high mass O and B type stars produce high levels of ultraviolet radiation and have strong outflows which is fed back into the cloud dissociating H_2 molecules and producing HII regions, thus beginning the process of photoevaporation, a process that will first halt star formation and eventually remove the cloud completely, leaving an open cluster (Faustini et al. 2009; McKee and Ostriker

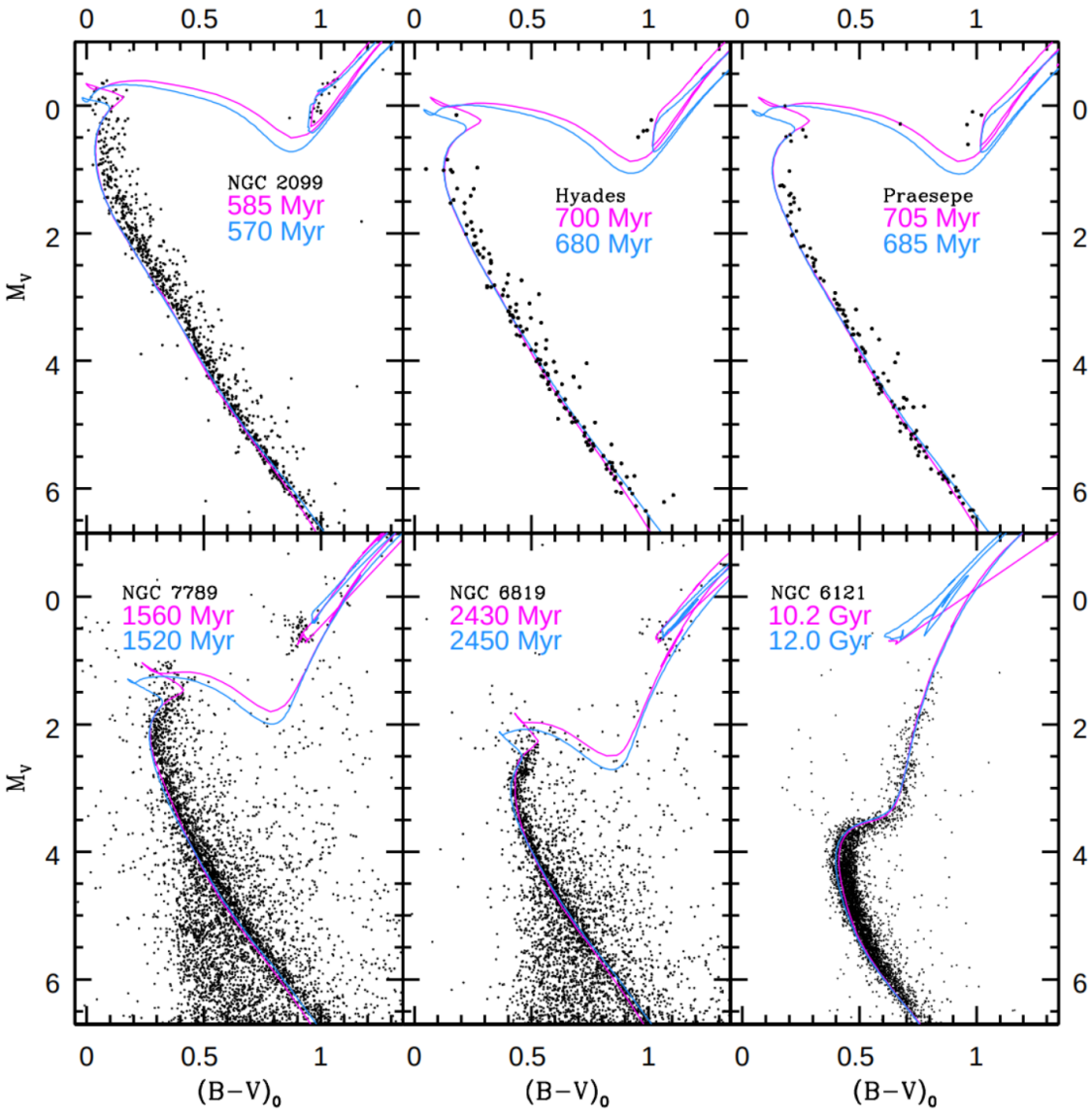


Figure 1.10 Examples of colour magnitude analysis from Cummings et al. (2018). The panels show CMDs for six star clusters of varying age over-plotted with isochrones from two leading theoretical models. PARSEC isochrone ages are shown in magenta, and the non-rotating MIST isochrone ages are shown in blue.

2007; Lada and Lada 2003). The majority of the binding mass of the cluster ($\sim$ 60%) resides in the gas from which it is forming (Moeckel and Bate 2010). Once the gas has evaporated the gravitational binding energy of the cluster is significantly reduced, in some cases by enough to allow the cluster to quickly dissolve, although the more massive, or denser clusters remain gravitationally bound, but less strongly.

The clusters are still dynamic in nature and as the stars interact over time will begin

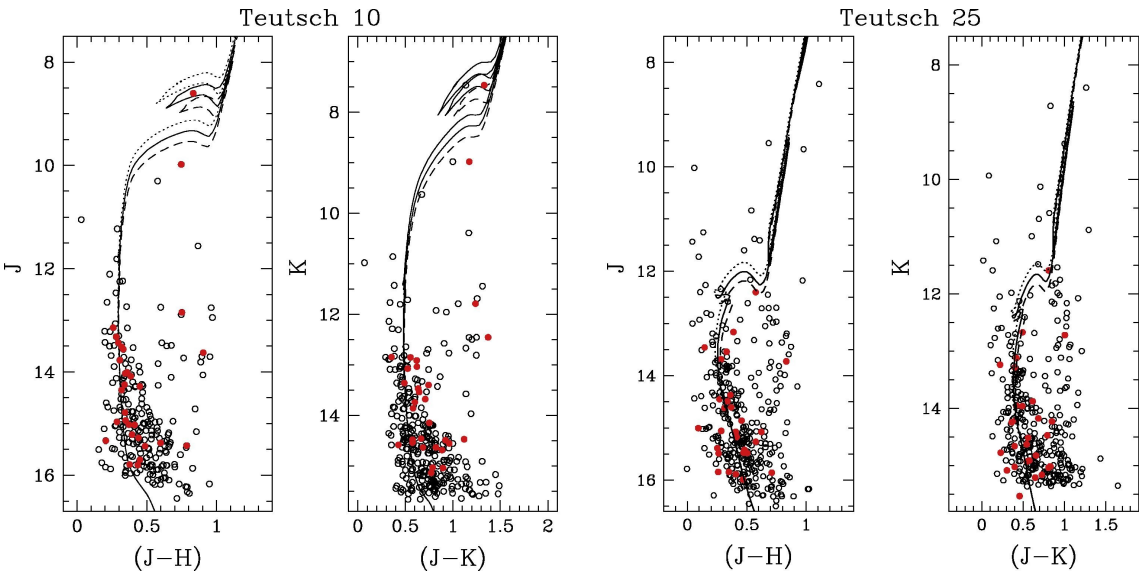


Figure 1.11 Examples of J vs J-H and K vs J-K colour magnitude diagrams for two clusters, Teutsch 10 and Teutsch 25. These panels show the fitting the Solar metallicity isochrones of $(log(age) = 7.80, 7.85, 7.90)$ for the cluster Teutsch 10 and $(log(age) = 8.90, 8.95, 9.00)$ for the cluster Teutsch 25. These isochrones are taken from Bressan et al. (2012) and corrected for reddening. In this case, most likely cluster members are represented by red dots.

to energetically relax as the kinetic energy of the cluster is redistributed from the more massive stars to the less massive. This process of equalisation across stars of different masses causes a mass-segregation, with higher-mass stars imparting momentum to lower mass stars resulting in the gradual movement of the higher mass body towards the centre of the cluster (Allison, Goodwin, Parker, de Grijs, et al. 2009). Over time the removal of the bulk gas mass and resulting reduction in binding energy, coupled with various gravitational interactions, work to gradually remove stars from the cluster with perhaps only 10% remaining after a period of 10^8 years. This dynamical evolution continues with the smaller, open clusters ($\sim 10^4 M_\odot$) surviving for approximately 100 Myr and massive open clusters ($\sim 10^5 M_\odot$) about 1×10^9 yrs (Gieles 2011; Portegies Zwart, McMillan, and Gieles 2010; Stahler and Palla 2008).

1.2.1 Cluster Taxonomy

It is useful at this point to define the different terms used to describe groups of associated stellar objects and give some typical properties. There are several opinions as to how to

define the variously sized groupings of stars, one definition of a cluster is given by Lada and Lada 2003 as group of physically related stars that survive longer than 10^8 yr as a bound system. This requires at least 35 stars and a stellar mass density greater than 1 $M_\odot$ pc^{-3}. Adams and Myers 2001 indicate three distinct groupings in which stars form: isolated singles and binaries, stellar groups (10 to 100 members), and clusters (> 100 stars).

These three modes represent different kinds of stellar systems that exhibit very different kinds of dynamical behaviour and evolution. Cluster environments subject their stellar and proto-stellar constituents to many disruptive and constructive influences as a result of the higher density of objects, whereas smaller groups have relatively little effect on the star formation process. It is also clear that dissolution times for groups are far shorter than those of true clusters which are more strongly gravitationally bound (Adams and Myers 2001).

Adams and Myers also provide a very clear, quantitative distinction between groups and clusters. Adams and Myers classification is well defined using physical properties rather than an arbitrary number, so this will be adopted for the purposes of discussion in this work. However, while this works well for a general classification when looking at the younger objects that are visible in the near infrared it is also useful to further differentiate.

Camargo, Bica, and Bonatto (2015) used a scheme of five classifications (repeated in the list below) which closely match the variety of clusters found in the observations used for this project. Therefore I will adopt a system based on these definitions with some modifications to suit my data set. Those objects established in the literature (i.e. with more than 10 references) and that have $\sim$ 20 or more decontaminated members are labelled as clusters, those with $\leq$ 10 references and typically less than $\sim$ 20 members are cluster candidates. The threshold of 10 references is used as a guiding principle for classification rather than a strict quantitative rule. I have also added a 'possible' label for objects that have the strong visual look of a cluster but lack the detections and references to be included in this work. These objects will need to wait for a higher resolution, deeper survey before they can be investigated. In addition to these there is the EGr class

for regions clearly showing star formation activity but without obvious signs of a cluster. Throughout this work the term 'clusters' will be used whenever referring to the sample as a whole.

Table 1.1 Cluster Taxonomy

OC	Open Cluster	Mainly dust-free objects, minimal JHK-band nebulosity, identified in > 10 papers, $n \geq 20$.
OCC	Open Cluster Candidate	Mainly dust-free objects, minimal JHK-band nebulosity, identified in ≤ 10 papers, $n \geq 20$.
POC	Potential Open Cluster	Poorly resolved objects that require deeper photometry or higher resolution, unconvincing CCD & CMDs, mainly dust-free, minimal nebulosity, ~ 0 papers, $n \leq 20$.
EC	Emerging Cluster	Still enshrouded in their molecular cloud, JHK-band nebulosity, identified in > 10 papers, $n \geq 20$.
ECC	Emerging Cluster Candidate	Less populous but still enshrouded in their molecular cloud, JHK-band nebulosity, identified in ≤ 10 papers, $n \leq 20$.
PEC	Potential Emerging Cluster	Poorly resolved objects that require deeper photometry or higher resolution, unconvincing CCD & CMDs, still enshrouded in their molecular cloud, ~ 0 papers, $n \leq 20$.
EGr	Embedded Group	Star-forming low-density groups, still enshrouded in their molecular cloud.
NC	No Cluster	Despite the initial visual suspicion of a cluster, no supporting data is found or the data is bad.

1.3 Outstanding Questions

As outlined in the preceding sections the star formation process including the initial conditions, evolution and dissolution within the Galactic plane is of great interest as it is

one of the key mechanisms that drives galaxy evolution and is responsible for fundamental processes such as the synthesis of the heavy elements and the formation of planets. This is a complex process that we are only now beginning to be able to investigate in detail thanks to modern instruments. However, despite the relatively large number of surveys in recent years there are still many unanswered questions around the process of cluster formation. This is an ongoing area of research, with the number of stars forming in these regions and the distribution of their masses an area of great interest.

At this time the IMF is not known to depend on the mass of the initial cloud, apart from the obvious upper limit on possible stellar masses placed by the total cloud mass and formation efficiency. It is possible however that the mass of the parent cloud has an influence on the mass distribution of the stars formed but this has not been observed yet. There is also interest in how the IMF varies across different environments, what are the key factors that influence it. These questions are discussed in Liptai et al. (2017), Peretto and Fuller (2009), and Offner et al. (2013) and Bastian, Covey, and Meyer (2010)) and references therein. In general a large proportion of work on the IMF has only measured it down to 1 $M_{\odot}$ so additional data from sites of lower mass star formation, and a wider variety of initial conditions, would be valuable.

There are also outstanding questions with regard to the propagation of star formation across GMCs (Bastian, Covey, and Meyer 2010)), tracers of imminent star formation such as dense clumps and water or methanol masers (Peretto and Fuller 2010; Peretto, Fuller, et al. 2013 and Caswell and Breen 2010). In addition to the factors that can initiate star formation what actually determines the star formation rate (SFR) and star formation efficiency within a galaxy? Understanding the factors that regulate the rate at which gas in a galaxy converts into stars is crucial. This includes the roles of galactic dynamics, feedback mechanisms, and environmental conditions (Krumholz 2014; Liptai et al. 2017; Offner et al. 2013)

Elucidating how star clusters form, evolve, and either disperse or remain bound over time is also vital. This encompasses the study of young massive clusters (YMCs) and their connection to globular clusters, and their role in investigating the wider role of

clusters in the context of galaxy formation and evolution, including their impact on the interstellar medium and feedback processes, see Cantat-Gaudin, Anders, et al. (2020), Longmore et al. (2014), and Lada and Lada (2003) among others.

The greater sensitivity and resolution of the UKIDSS GPS survey (see Chapter 2) offers the possibility to investigate these questions to greater depth. From extending IMF measurements down below 1 $M_{\odot}$, to resolving small (or distant) young or more evolved open clusters, and providing the initial data needed to develop these hypotheses further.

1.4 Project Aims

The primary goal of this work is to utilise the high spatial resolution and sensitivity of the UKIRT-WFCAM to extend the sample of known young, compact, Galactic star clusters in a parameter space where previous surveys such as 2MASS and Spitzer have been unable to provide data due to their limited resolution or depth. This is even true for modern surveys such as Gaia, where WFCAM actually has better sensitivity for young, embedded clusters or distant clusters with high extinction. This work should be able to identify more visually compact clusters, either at an early stage of evolution or more distant than before, as well as being able to identify compact regions of star formation within dark clouds and filaments, which are typically difficult to resolve. A lack of small, compact objects would also be an interesting result however as, for instance, it could indicate that small clusters have an unusually high rate of 'infant mortality' as described by Lada and Lada (2003), where small young clusters are rapidly disrupted by dynamic interaction and the rapid gas dispersal by feedback from the young stars.

To summarise the aims of the project:

1. Identify and provide basic characterisation for new star clusters and cluster candidates found during the quality control checking of the UKIDSS GPS DR10 data release.

2. Determine whether the UKIDSS GPS can be used to extend the sample of young, compact clusters into new parameter space with its greater resolution and depth.

3. Use data from ESAs Gaia DR3 release to investigate the proper motion and distances of both the existing and new / candidate clusters identified in 1 above.

4. Investigate the usefulness of a published, unsupervised machine learning technique to provide additional support to the cluster hypothesis for the new / candidate clusters where possible.

5. Provide confirmation of previously identified candidate clusters where possible.

Chapter 2

Data Sources and Observations



Figure 2.1 The UKIRT atop Mauna Kea. Image Credit: Jason Chu

The primary data set that I used for this investigation is the United Kingdom Infrared Deep Sky Survey (UKIDSS, see Lawrence and Warren 2007) Galactic Plane Survey (GPS, see Lucas et al. 2008) which consists of both an image archive and curated database. These are described in more detail in the following section. This core data set has then been combined with other survey results, where appropriate, in order to enhance the luminosity and wavelength range of the data, gain a better understanding of the environmental context, and to add valuable proper motion and parallax information. All

of this adds context to the observations and aids the confirmation and basic characterisation of the newly discovered clusters.

Near Infrared (NIR) observations (i.e. those between 0.8 and 2.5μm) are particularly suited to the identification of young star clusters and star forming regions (SFRs) because, at the longer end of these wavelengths (>≈1.5μm) the dust that typically obscures these regions starts to become transparent, and we are able to glimpse the young, relatively cool, objects forming within.

2.1 The United Kingdom Infrared Deep Sky Survey's Galactic Plane Survey (UKIDSS GPS)

The GPS was one of five public surveys undertaken by the UKIDSS consortium beginning in 2005 and using the Wide-Field Camera (WFCAM, see Casali, Adamson, et al. 2007 and Hewett et al. 2006) on the United Kingdom Infrared Telescope (UKIRT), which is located atop Mauna Kea on the Big Island of Hawaii. The UKIRT began operations in 1979 and has a 3.8m diameter primary mirror making it one of the largest infrared survey instruments in the world. The addition of the Wide-field Imager WFCAM in 2004 facilitated the UKIDSS programme of surveys and is described more fully below in section 2.1.1.

The GPS is designed to cover the entire northern and equatorial Galactic plane accessible to the United Kingdom Infrared Telescope (UKIRT) in a 10 degree wide band across the sky covering approximately 1800 degree^2 of the northern Galactic plane in the J, H, and K pass-bands (Lucas et al. 2008). In addition to this the GPS includes a narrower southern extension towards the Galactic Centre and a 200 degree^2 survey of the Taurus-Auriga-Perseus (TAP) molecular cloud complex, following the contours of the ^{13}CO gas mapped by Ungerechts and Thaddeus 1987. The southern extension is narrower because of its poorer accessibility from Mauna Kea. In Galactic coordinates, the in-plane regions are:

- $141° < l < 230°,\ b = \pm 5°$

- $15° < l < 107°,\ b = \pm 5°$
- $-2° < l < 15°,\ b = \pm 2°$

The section of the plane between $107° < l < 141°$ is not covered because it lies north of 60° declination, a region of the sky that UKIRT cannot reach due to a fundamental limitation of the telescope design. The data I used for this research comes from the latest (and last) GPS data release, DR10. GPS DR10 contains just under 1 billion sources and achieved depths of J= 19.8, H= 19.0, K= 18.1 (see Lawrence and Warren 2007 and Lucas et al. 2008). Fig. 2.2 shows the sky coverage for the GPS DR10 data release as a blue shaded area.

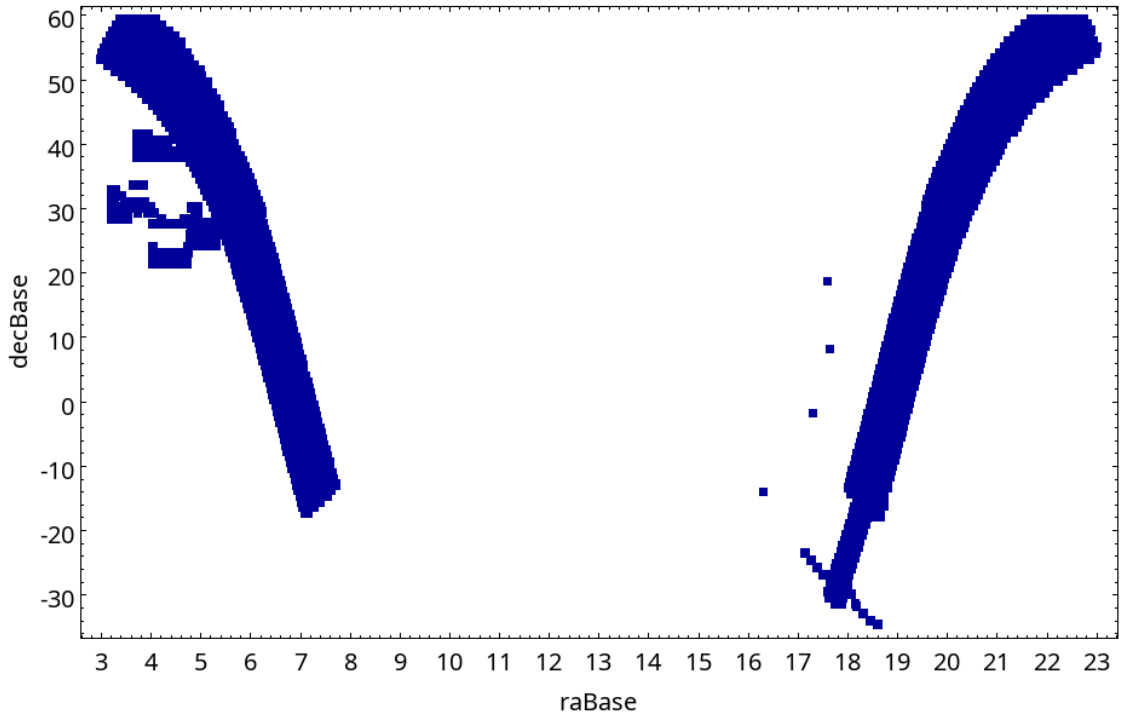


Figure 2.2 RA/DEC plot showing sky coverage area for the UKIDSS GPS survey DR10 in blue

2.1.1 The Wide Field Camera (WFCAM)

As mentioned previously, observations were made using the WFCAM instrument and this is what gives the GPS its impressive depth. For the deep UKIDSS surveys, such as the UDS, it was able to achieve depths of J > 22 and K > 21 (Dye et al. 2006). The

median 5σ depths achieved by WFCAM for the GPS are also given in Dye et al. 2006, as well as Lucas et al. 2008; Lawrence and Warren 2007, as J = 19.77, H = 19.00, K = 18.05 which is a significant improvement over its predecessor, the 2-Micron All Sky Survey (2MASS), which achieved 10σ depths of J = 15.8, H = 15.1, and K = 14.3 in the point source catalogue (see Skrutskie et al. 2006). Ultimately the survey depth is limited by source confusion with seeing of about 0.6 arcsec (Full Width Half Maximum - FWHM) at best.

WFACM was built by the UK Astronomical Technology Centre (UK ATC) in Edinburgh as a common user instrument for UKIRT and the UKIDSS programme owes its feasibility to the large field of view of the camera. The WFCAM field of view is ((0.9°) and this coupled with its four 'large-format' detectors gives it significant surveying performance, in fact it is able to survey at > 100 times the rate of 2MASS (Casali, Adamson, et al. 2007).

For the GPS, WFCAM uses the J (1.17-1.33μm), H (1.46-1.635μm) and K (2.03-2.37μm) band-pass filters from the Mauna Kea photometric system (Tokunaga, Simons, and Vacca 2002) with total on-source integration times of 80s, 80s and 40s respectively. These integrations are made up of shorter exposures, employing a sub-pixel dithering strategy to fully sample the image profile, as described in Casali, Adamson, et al. 2007.

The data captured by WFCAM passes through an onsite pipeline to make the initial quality checks and ensure the telescope is operating correctly before being shipped to the Cambridge Astronomical Survey Unit (CASU) for further reduction and collation into the archived science products available via the WFCAM Science Archive (WSA). The CASU is part of Cambridge University Institute of Astronomy in Madingley and provides data pipeline processing and archiving of data products for several survey project instruments, including WFCAM. The data pipeline is based on the VISTA Data Flow System (VDFS, see Irwin et al. 2004) developed at CASU by Mike Irwin and others for WFCAM and other data.

Each WFCAM exposure produces four FITS files (Flexible Image Transport

System), one for each detector, and each night of observation creates about 200GB of data which is transferred to the CASU where they are combined into a Multi-Extension FITS files (MEF). The data consists of three types of file: science frames (including the calibration), flat frames and dark frames. Dark frames are used to ascertain and remove thermal and other detector noise while flat frames, taken with a uniform light source, are used to normalise the per-pixel sensitivity of the detectors. The MEF files are then further processed into the various data products before being stored in, and accessed via, the WSA. See Hambly and Collins 2008, Irwin et al. 2004 and Dye et al. 2006 for full description of the WSA and the CASU data pipeline.

WFCAM is a wide field imaging camera, primarily designed to carry out large scale semi-automatic surveys of the sky. It operates in the near IR between $1 - 2.5\mu$m and uses four 2048x2048 pixel focal plane detector arrays. These are arranged in a 2x2 pattern, spaced by approximately 90% of a detector spacing. Four separate stepped exposures are taken and combined to produce a single image frame, or tile, covering approximately (0.9°) x (0.9°).

Figure 2.3 shows the layout of the focal plane, together with the angular field coverage of the optical system. Figure 2.4 shows the optical layout of the WFCAM and figure 2.5 shows the WFCAM installed on the UKIRT.

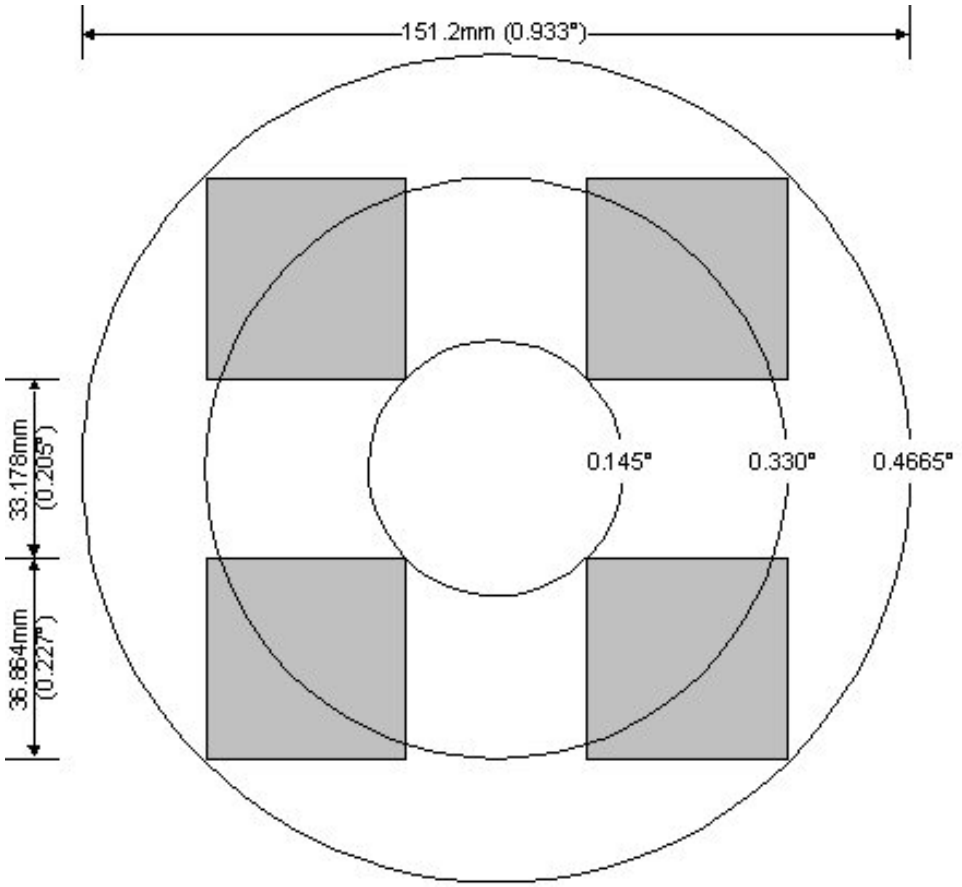


Figure 2.3 WFACM Detector Layout and Field of View. Image Credit: Royal Observatory Edinburgh, Astronomy Technology Centre

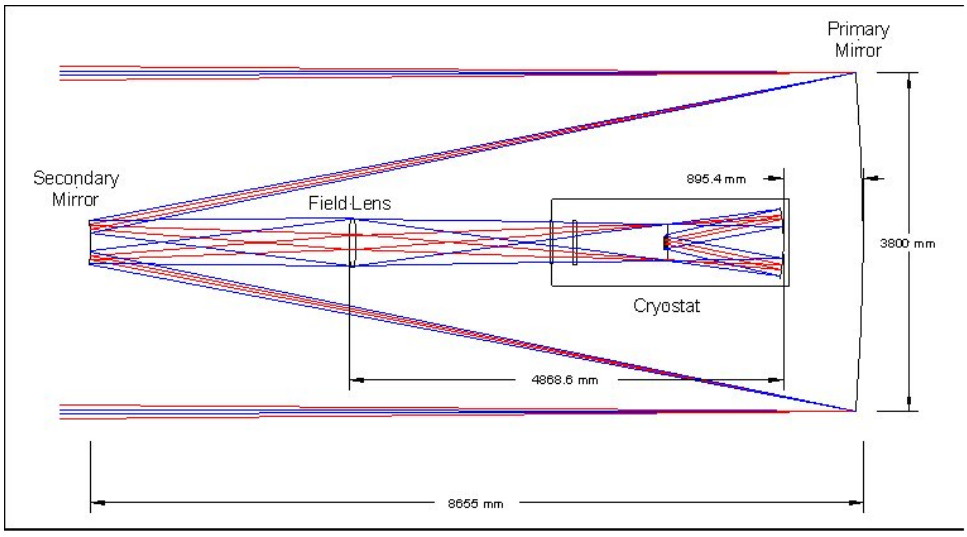


Figure 2.4 WFCAM Optical Layout. Image Credit: Royal Observatory Edinburgh, Astronomy Technology Centre



Figure 2.5 WFCAM Installed on the UKIRT. Image Credit: Laurie Chu and Jessica Stasik

2.2 The Two-Micron All-Sky Survey (2MASS)

While the GPS provides excellent depth and resolution it is limited at brighter magnitudes by its saturation limit of approximately magnitude 13. This ultimately limits our ability to identify potential cluster members using just GPS data alone. 2MASS was a precursor survey using almost identical band-pass filters but with lower resolution and sensitivity, but a much brighter saturation limit of approximately magnitude 4. Thus by combining the GPS and 2MASS data sets I was able to extend the dynamic range of my dataset significantly. This is described fully in section 2.2.1.

The 2MASS project ran from 1997 to 2001 and scanned the entire sky in three near-infrared bands (see Skrutskie et al. 2006). As well as the science enabled by its coverage, improved sensitivity and resolution it was intended to be a feeder project for larger ground and space-based telescopes such as Spitzer Space Telescope, Hubble Space Telescope, Keck and Gemini. To get whole sky coverage it used two highly-automated 1.3m telescopes, one at Mt. Hopkins, AZ, and one at CTIO, Chile. Each telescope was equipped with a three-channel camera, each channel consisting of a 256×256 array of

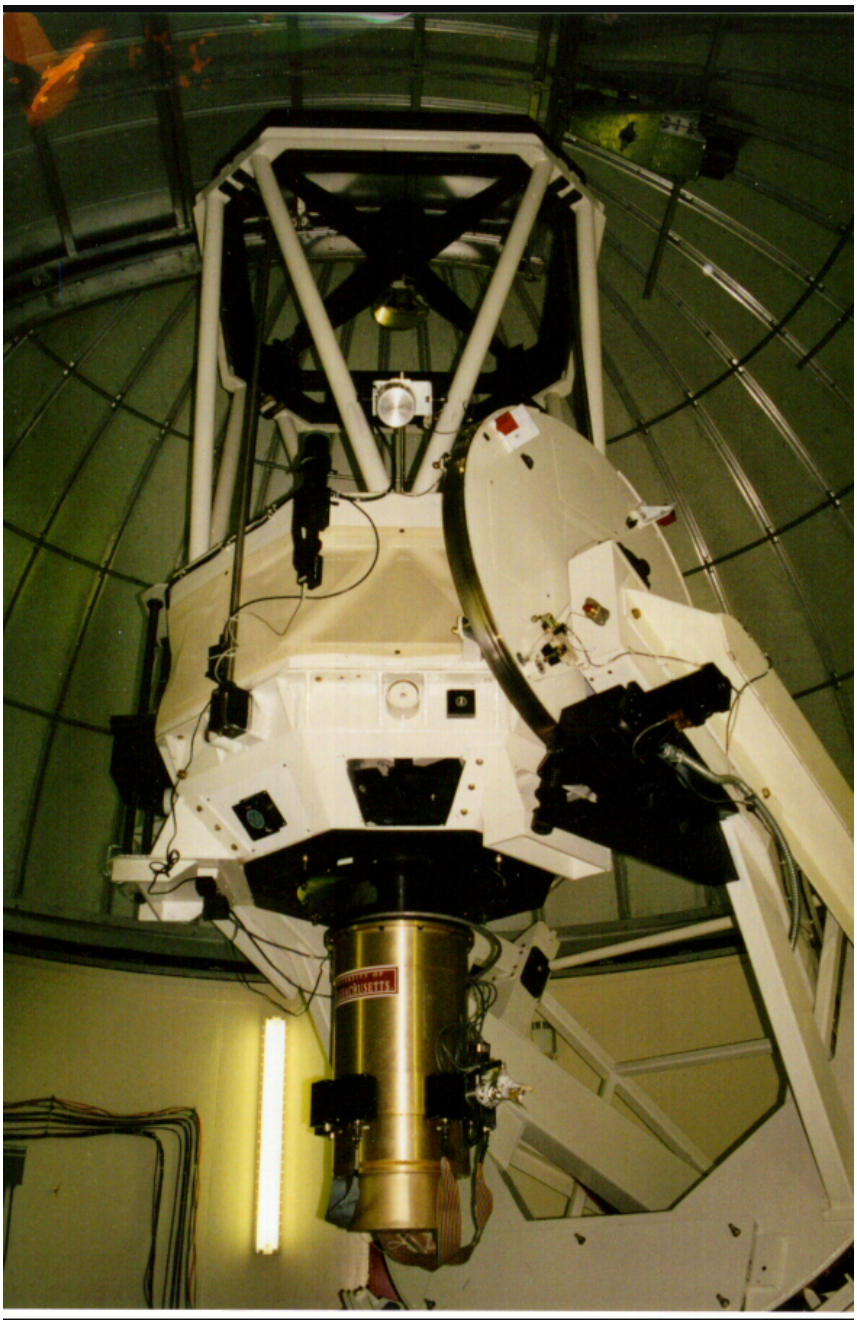


Figure 2.6 The 1.3m telescope at Mt. Hopkins, AZ, one of the two, 2MASS instruments. Image Credit: M3 Engineering and Technology Corp.

HgCdTe detectors and observed simultaneously at J (1.25μm), H (1.65μm), and K_s (2.17μm).

The 2MASS arrays imaged the sky while the telescopes scan smoothly in declination at a rate of ~1´ per second. The 2MASS data "tiles" are 6° long in the declination direction and one camera frame (8.5´) wide. While the entire telescope scanned in the

declination direction, the telescope's secondary mirror tilted opposite the scan direction to momentarily freeze the focal plane image. At the end of a 1.3s exposure the secondary flew back to its start position and froze a new frame slightly displaced from the previous one. The dead time between frames used for secondary fly-back and array reset was less than 0.1s. The camera field-of-view shifted by ~1/6 of a frame in declination from frame-to-frame. The camera images each consisted of six pointings on the sky for a total integration time of 7.8s, with sub-pixel dithering, which improves the ultimate spatial resolution of the final Atlas Images. When accounting for dead time and the time to point the telescope and initiate a scan, the 2MASS observing system integrated on the sky approximately 84% of each night.

The University of Massachusetts was responsible for the overall management of the project and for developing the infrared cameras and on-site computing systems at both facilities. The California Institute of Technology Infrared Processing and Analysis Center (IPAC) is responsible for all data processing through the production pipeline, and construction and distribution of the data products. The 2MASS project involves the participation of members of the science team from several different institutions. See Kleinmann et al. 1994 and Skrutskie et al. 2006. 2MASS data products are available via the IPAC InfraRed Science Archive (IRSA) Gator service in a similar manner to the WSA.

2.2.1 Managing saturation limits of the GPS with 2MASS

Although the GPS is able to achieve a good depth it does suffer from saturation of the brighter sources and this can create a number of issues ranging from spurious sources to source confusion and loss of data. In order to increase the optical range of observations available for me to use I needed to identify the upper magnitude limit of reliability for the data, both in the direction of the galactic centre, where high source confusion leads to premature saturation, and towards the outer galaxy, where the source density is lower and I also needed to cover the full range of GPS latitudes to ensure we are accounting for the increased reddening towards the plane of the galaxy. While 2MASS has a shallower

depth it has a much brighter saturation limit and thus can provide photometry for the brighter sources, as described below, and this offers an opportunity to use a larger sample size improving the completeness of the study. Firstly, it is important to make a comparison between 2MASS and GPS to determine the point at which GPS mags start to diverge from 2MASS (i.e. they become less reliable), and therefore set a more realistic limit for the lower magnitude.

2MASS was used to provide the primary photometric calibration for WFCAM (see Hodgkin et al. 2009 and refs therein) and has a whole-sky calibration accuracy of $\leq 2\%$ (Nikolaev et al. 2000), so is a perfect companion to the GPS needing little to no correction. Both 2MASS and WFCAM use the Vega (Johnson) magnitude system which mean that the J, H & K pass bands overlap strongly there are only subtle differences in mag. The fact that 2MASS uses the K_s band vs the WFCAM K band results in a difference of only around 0.01 mag. Hodgkin et al. 2009 gives details in §4 of the photometric transformation for DR2 onwards as:

$$J_w = J_2 - 0.065(J_2 - H_2)$$

$$H_w = H_2 + 0.070(J_2 - H_2) - 0.030$$

$$K_w = K_2 + 0.010(J_2 - K_2)$$

Hodgkin et al. also looked at the effect of Galactic extinction in §4.3, where at lower latitudes it becomes significant, moving intrinsically blue stars to the red and reddening the giants further. While high levels of extinction can significantly affect the relationship between different photometric systems in the visible spectrum, the effect on the J, H, and K NIR bands used here is negligible, and we can therefore assume the 2MASS / GPS relationship remains robust.

One important point mentioned above was that the saturation limits in the GPS vary depending on the sky background and to help users assess and work around saturation the WSA provides the `ppErrBits` field. `ppErrBits` contains data on the reliability of the source and whether it is saturated, and a full description of the values can be found in (Lucas et al. 2008 and Hambly and Collins 2008). Figure 2.7 shows an example of a field containing a saturated star with diffraction spikes and the toroidal shape as described in

Lucas et al. 2008. This data is provided for each of the band-passes (`jppErrBits`, and so on) and will have a value < 65536 for unsaturated sources. It should be noted that heavily over-saturated stars can cause multiple spurious detections in each filter, and the centre often has a lower count level than the sky background leading to the 'hollow' image profile. This is however a general feature of data taken with near infrared arrays in the correlated double sampling mode (CDS) and since the saturated region is typically larger than the size of the stellar image profile used by the pipeline source detection algorithm, they are often erroneously listed as resolved sources with `mergedclass=+1` (i.e. galaxy) rather than stars. Conservatively, it is safest to make cuts at $J < 13.25$, $H < 12.75$ and $K < 12.0$ in order to avoid issues with unreliable photometry (Lucas et al. 2008).

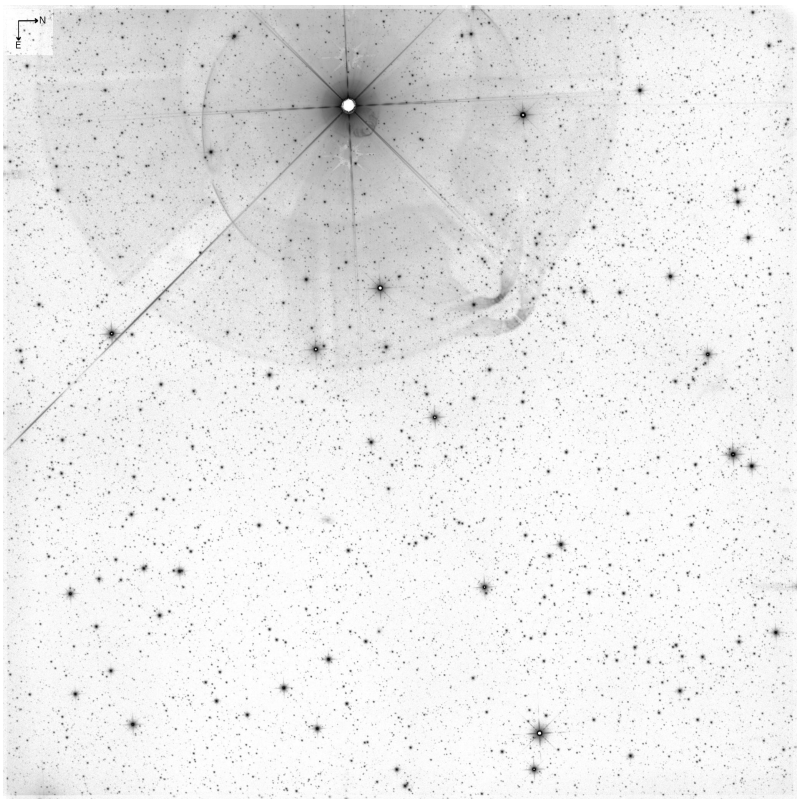


Figure 2.7 Example of a saturated bright star in a GPS field. Note the bright white centre caused by saturation of the detector. There are typically large diffraction spikes and other spurious artefacts visible such as the rings and ghosting shown in this image. Saturation effects are discussed in §A2 of Lucas et al. 2008 and are now always confined to the detector of the saturated source. Most GPS images will have saturated sources but most will not affect the quality of data, unlike this example

As previously described, 2MASS has a strong similarity to the GPS in that the

wavelength bands are almost the same and dynamic ranges overlap to a certain extent, with 2MASS extending to the brighter magnitudes and GPS the fainter. By making a cross-match between the two it is possible to look at how the reported magnitudes of objects vary between the surveys and by plotting these against each other as a function of magnitude we are able to see the point at which the two begin to diverge. While both surveys report the same values of magnitude for the same objects the plot will be a straight line, the point at which one domain begins to deviate from the centre line indicates the point at which the magnitudes become unreliable and hence provide the limiting magnitude. The only difference between the UKIDSS and 2MASS bands is in the K-band, with 2MASS using a K_s (or K-short) filter which has a band-pass of 2 - 2.31μm (Skrutskie et al. 2006) vs the UKIDSS Mauna Kea Consortium specification filter band-pass of 2 - 2.37μm

As mentioned, the dynamic range of the GPS is dependent on the sky background level and this will vary as a function of galactic longitude because of variability in source density. Ideally, we would make measurements for each small increment of latitude and calculate the saturation limit dynamically. This is time consuming and in most cases unnecessary because of the overlap in depth between the two surveys. What would be ideal is a cut-off which would satisfy the saturation limit in the direction of highest source density, i.e. towards the Galactic centre, and still be high enough not to exclude too many sources, and thus affect completeness, at the less dense longitudes towards the outer Galaxy. To make this comparison I selected one longitude band towards the inner galaxy, 30 – 31°, and one towards the outer galaxy, 100 – 102°. Rather than cross match data across all magnitudes for these bands which would make the plots harder to read, I opted to use a limited range of mag 10-13 as this is the range in which we should see the divergence.

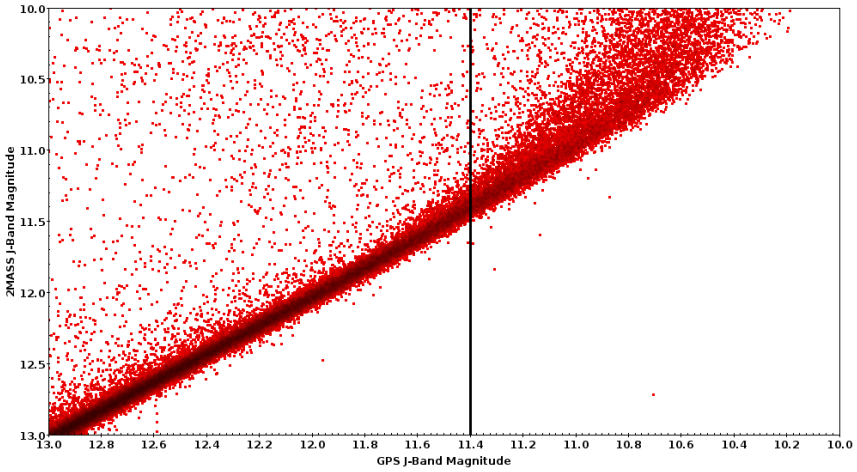
The WSA conveniently provides cross-matched data for a number of complementary surveys and this is available via 'neighbour' tables using the free-form SQL interface. Figure 2.8 lists the SQL queries used to extract the GPS-2MASS cross-matched source data.

```
SELECT k_1apermag3,k_2apermag3,k_m
FROM gpssource AS gps, TWOMASS..twomass_psc AS psc, gpsSourceXtwomass_psc AS x
WHERE x.masterObjID=gps.sourceID AND x.slaveObjID=psc.pts_key AND
      x.distanceMins<0.016 AND k_m BETWEEN 10 AND 13
      AND gps.l BETWEEN 100 AND 102

SELECT japermag3,j_m
FROM gpssource AS gps, TWOMASS..twomass_psc AS psc, gpsSourceXtwomass_psc AS x
WHERE x.masterObjID=gps.sourceID AND x.slaveObjID=psc.pts_key AND
      x.distanceMins<0.016 AND j_m BETWEEN 10 AND 13
      AND gps.l BETWEEN 100 AND 102

SELECT hapermag3,h_m
FROM gpssource AS gps, TWOMASS..twomass_psc AS psc, gpsSourceXtwomass_psc AS x
WHERE x.masterObjID=gps.sourceID AND x.slaveObjID=psc.pts_key AND
      x.distanceMins<0.016 AND h_m BETWEEN 10 AND 13
      AND gps.l BETWEEN 100 AND 102

SELECT k_1apermag3,k_2apermag3,k_m
FROM gpssource AS gps, TWOMASS..twomass_psc AS psc, gpsSourceXtwomass_psc AS x
WHERE x.masterObjID=gps.sourceID AND x.slaveObjID=psc.pts_key AND
      x.distanceMins<0.016 AND k_m BETWEEN 10 AND 13
      AND gps.l BETWEEN 30 AND 31

SELECT japermag3,j_m
FROM gpssource AS gps, TWOMASS..twomass_psc AS psc, gpsSourceXtwomass_psc AS x
WHERE x.masterObjID=gps.sourceID AND x.slaveObjID=psc.pts_key AND
      x.distanceMins<0.016 AND j_m BETWEEN 10 AND 13
      AND gps.l BETWEEN 30 AND 31

SELECT hapermag3,h_m
FROM gpssource AS gps, TWOMASS..twomass_psc AS psc, gpsSourceXtwomass_psc AS x
WHERE x.masterObjID=gps.sourceID AND x.slaveObjID=psc.pts_key AND
      x.distanceMins<0.016 AND h_m BETWEEN 10 AND 13
      AND gps.l BETWEEN 30 AND 31
```

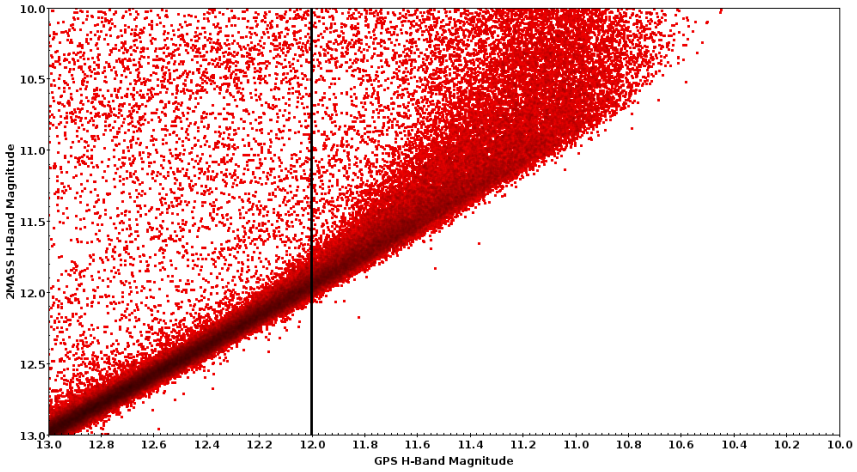
Figure 2.8 SQL statements used to extract the cross-matched GPS and 2MASS data for magnitude comparison. Statements cover JHK bands both toward and away from the galactic centre.

The results are more clearly shown in figure 2.9 and figure 2.10. In these plots we are looking to see the point at which the magnitudes reported by the GPS become divergent from those reported by 2MASS as we move toward the brighter end. This indicates the upper reliable level for GPS magnitudes, and therefore the point at which we should make a cut. This is preferable to relying on the conservatively set `ppErrBits` flag as it gives the largest GPS sample possible and we can then merge the 2MASS data to give a single dataset with a higher dynamic range.

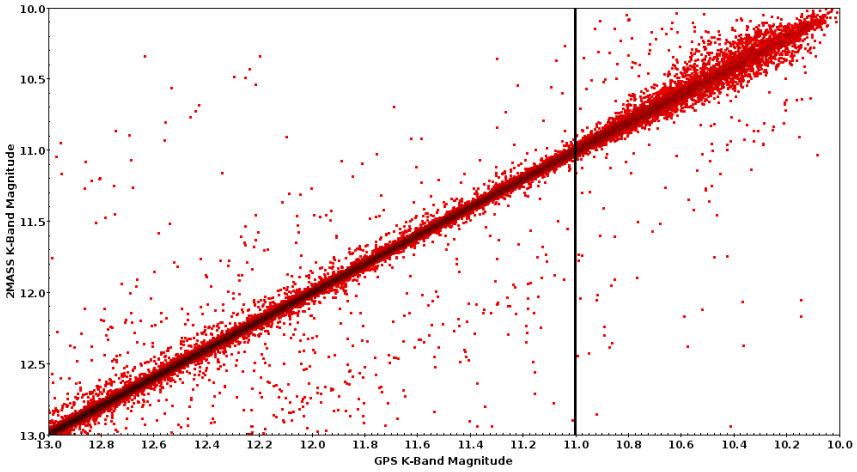
In figure 2.9 we are looking at sources in the galactic longitudes $100 - 102°$ towards the outer galaxy. Plot *a* shows us the J-band, plot *b* the H-band and plot *c* the K-band. We can see in plot *a* that the GPS data begins to diverge at around J= 11.4, plot *b* around H= 12 and plot *b* around K= 11. These values also hold within a reasonable limit for figure 2.9 where we are looking at longitudes $30 - 31°$ towards the galactic centre, although the much higher source density makes this much harder to determine. It was however desirable to find a limit that would work in both regions to make automated data processing realistic.



(a) UKIDSS GPS J-Band Magnitudes vs 2MASS J-Band Magnitudes

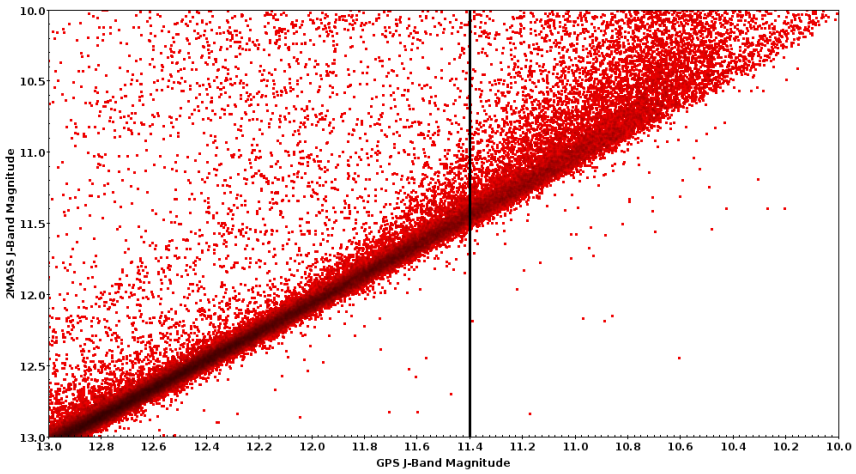


(b) UKIDSS GPS H-Band Magnitudes vs 2MASS H-Band Magnitudes

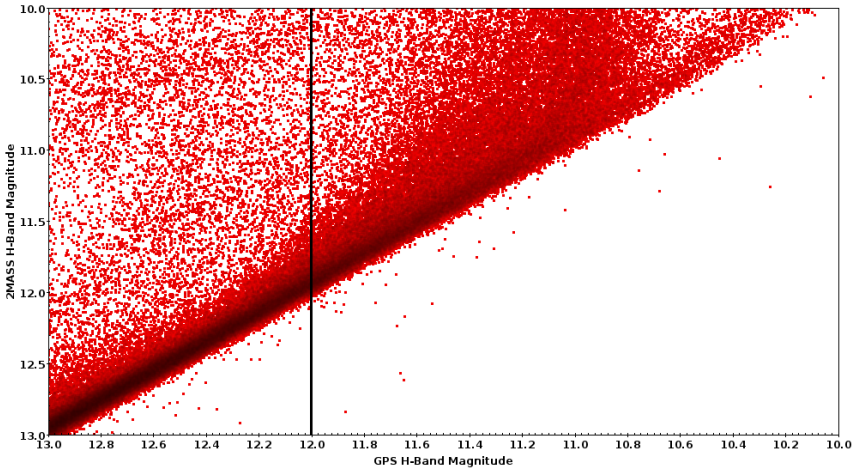


(c) UKIDSS GPS K-Band Magnitudes vs 2MASS K-Band Magnitudes

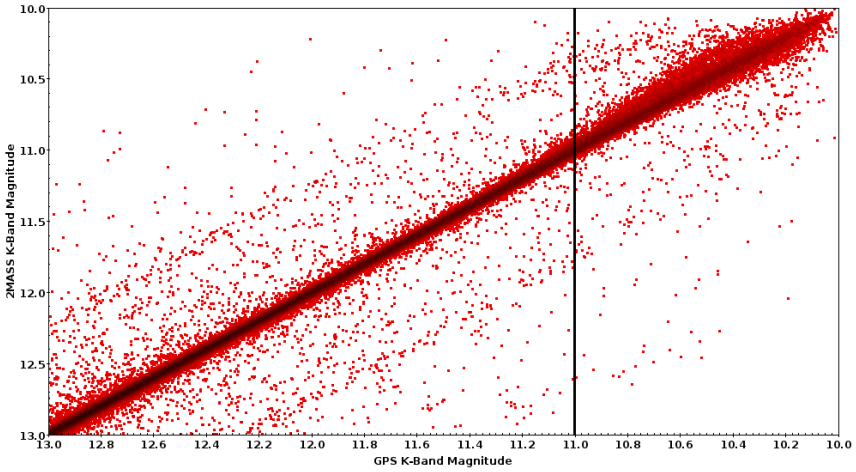
Figure 2.9 Comparison between UKIDSS `apermag3` JHK magnitudes plotted against 2MASS JHK_s magnitudes for cross-matched sources between Galactic longitude 100 – 102°, towards the outer galaxy. The vertical line represents the selected magnitude cut based on the plot.



(a) UKIDSS GPS J-Band Magnitudes vs 2MASS J-Band Magnitudes



(b) UKIDSS GPS H-Band Magnitudes vs 2MASS H-Band Magnitudes



(c) UKIDSS GPS K-Band Magnitudes vs 2MASS K-Band Magnitudes

Figure 2.10 Comparison between UKIDSS `apermag3` JHK magnitudes plotted against 2MASS JHK_s magnitudes for cross-matched sources between Galactic longitude 30 – 31°, towards the inner galaxy. The vertical line represents the selected magnitude cut based on the plot.

2.3 The Wide-field Infrared Survey Explorer (WISE)

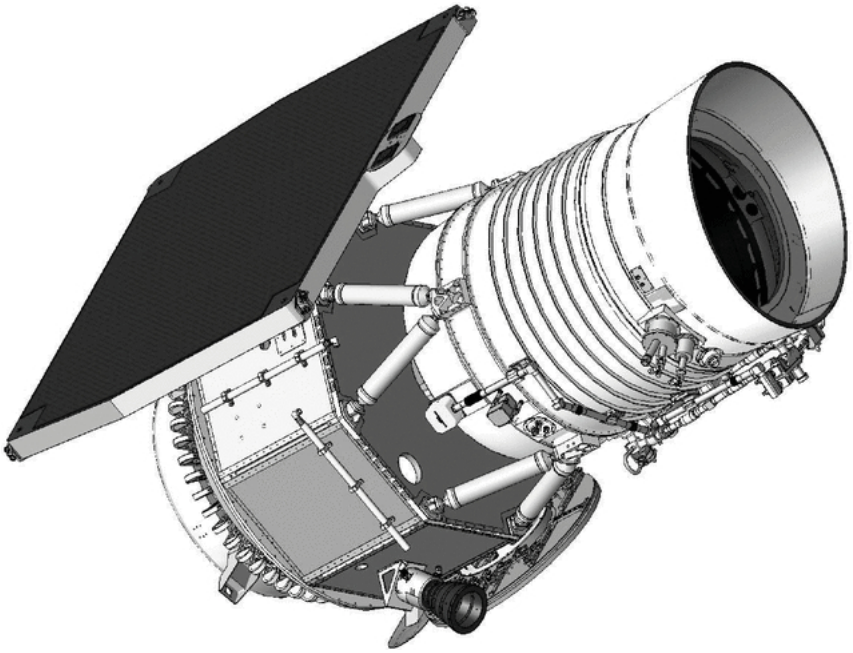


Figure 2.11 The WISE flight system in survey configuration with cover off. Image Credit: NASA / UCB

The Wide-field Infrared Survey Explorer (WISE) is a space-based observatory launched in late 2009 to survey the full sky in four IR wavelength bands (3.4, 4.6, 12 and 22μm). This it did until the frozen hydrogen cooling the telescope was depleted in September 2010 when it was put into hibernation before being repurposed in 2013 as NEOWISE to study Near Earth Objects.

The WISE instrument itself consists of a 40cm telescope and imaging optics. It has a 0.8° field of view and uses HgCdTe and SiAs 10242 detector arrays with a plate scale of 2.75 arcsec pixel^{-1} giving it a resolution of 6 arcsec (12 arcsec at 22μm). WISE has adopted the 2MASS strategy of moving the telescope at a constant inertial rate and using a scan mirror to freeze the sky on the arrays for the integration time. It initially used a two-stage solid-hydrogen cryostat with an expected 10-month lifetime to cool focal planes and optics to ~7°K. The WISE mission is described more fully in Wright, Eisenhardt, et al. 2010.

WISE provides an all-sky survey which is up to 500 times more sensitive than the original IRAS survey. WISE's goals were to find the most luminous galaxies in the

Universe, the closest stars to the Sun, detect most main-belt asteroids larger than 3km, and extend the 2MASS survey into the thermal infrared. WISE enables a wide variety of studies ranging from the evolution of protoplanetary debris discs to the history of star formation in normal galaxies. The WISE data products are managed by the IPAC IRSA in a similar way to the 2MASS data as described above and made available via a web portal including an image viewer (see 2.12).

Given its specific set of science goals the WISE source extraction pipeline was not optimised for the Galactic plane and this ultimately limits its use in this project, however it does provide the ability to use the IRSA Viewer to provide a wide field view of a cluster candidates environment at longer wavelengths which, for younger clusters particularly, can provide supporting evidence of star formation activity such as HII regions and clumps of relatively cool material, both at the coordinates of interest and in the local environment. The location of the WISE pass-bands in the mid-infrared matches the wavelength range where the excess emission from cooler circumstellar disk or envelope material in young stellar objects (YSOs) begins to become significant in relation to the stellar photosphere.

YSOs are typically found in regions that are bright in the mid-infrared from thermal dust emission or polycyclic aromatic hydrocarbon (PAH) dust, or regions that are dark as in the case of infrared dark clouds (Peretto and Fuller 2009) or in dense clustered regions. This fact means that WISE can be in fact be useful as a tool to find young stellar objects (YSOs) and other star formation tracers such as HII regions (Anderson et al. 2013) in a similar way to the work carried out with the Spitzer Space Telescope (SST) by Allen et al. 2004, Gutermuth, Myers, et al. 2008 and others.

Figure 2.12 shows an example of the IRSA Viewer showing a cluster candidate environment with a 10 arcmin (top row) and 30 arcmin (bottom row) field. The long wavelength nebulosity is a strong indicator of star formation activity and thus can be used to support the other lines of evidence such as CMDs and proper motion to identify this as a new cluster.

Figure 2.13 is an RGB render using W2 band (4.6μm) as red, W3 band (12μm) as

green and W4 band (22μm) as blue. This composite allows us to assess the environment and we can clearly see the green nebulous dust surrounding the cooler, denser blue core tightly formed around a cluster of stars.

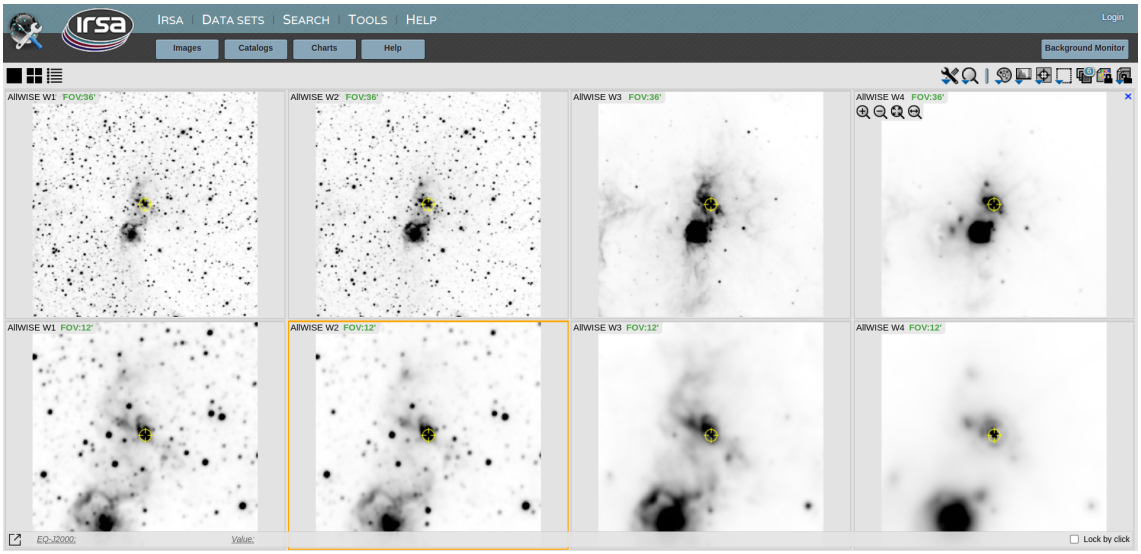


Figure 2.12 IRSA Viewer showing wise images for a cluster candidate in four wavebands at both 10 and 30 arcmin fields.

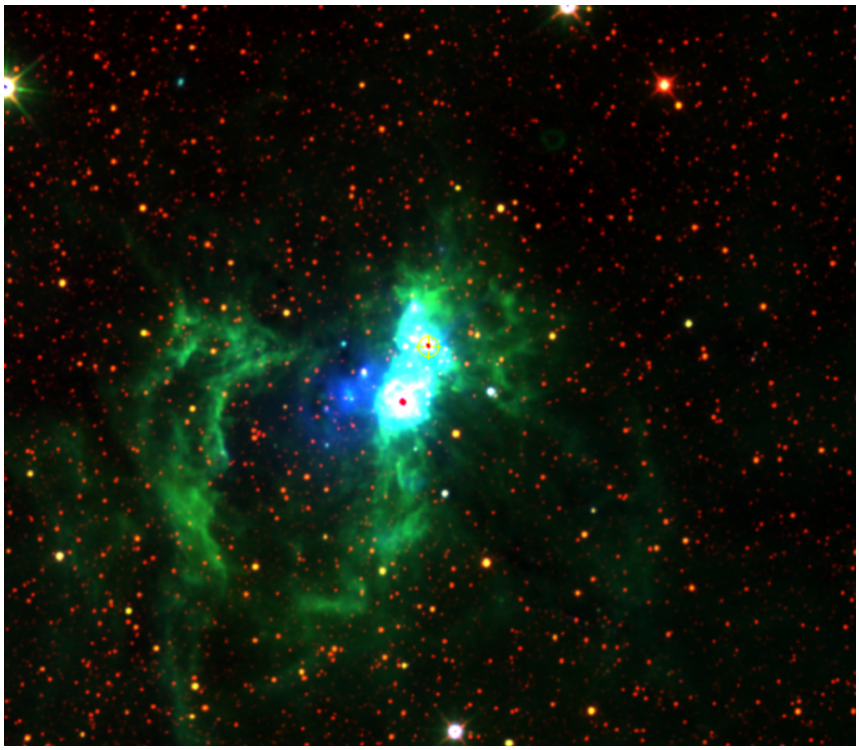


Figure 2.13 IRSA Viewer showing a 1 degree RGB composite of the same field.

2.4 The Gaia mission

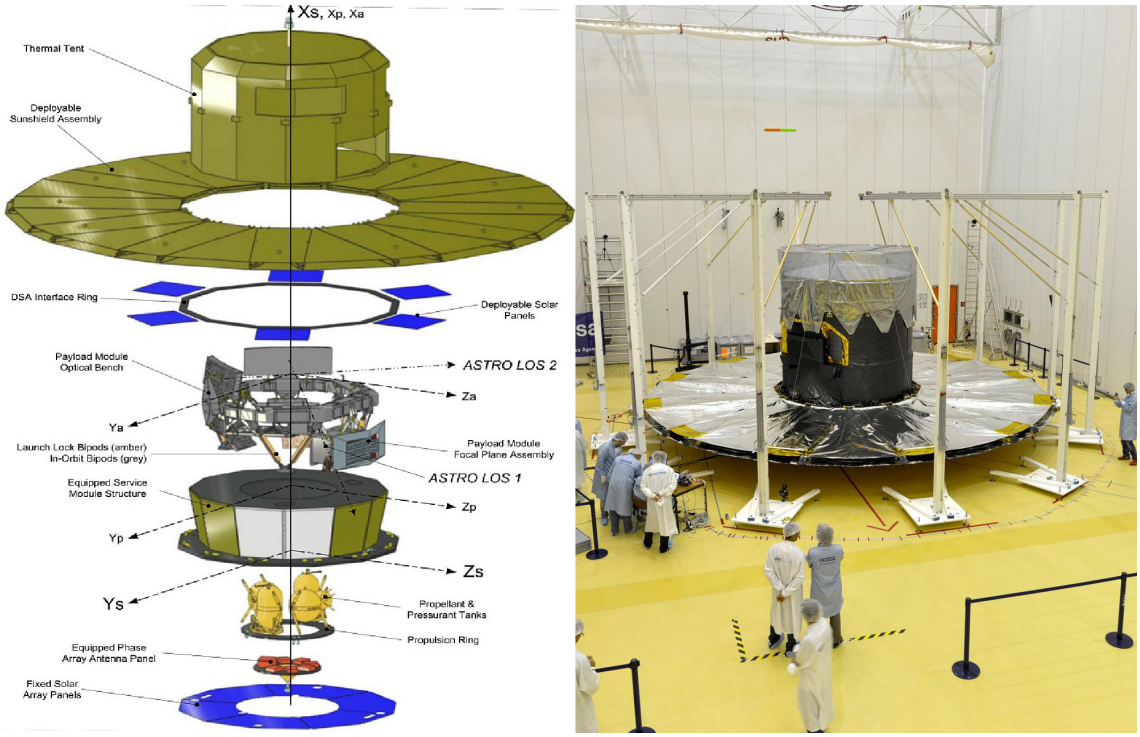


Figure 2.14 Left - Exploded diagram of Gaia subsystems. Right - Gaia with deployed sun shield during ground testing. Image Credits: Rudolph et al. (2014)

Gaia is a cornerstone mission in the science programme of the European Space Agency (ESA). Launched in December 2013, it builds on the success of the Hipparcos mission and was tasked with measuring the three-dimensional spatial, and the three-dimensional velocity distribution of approximately 1.8 billion stars. In addition to this it was to determine their astrophysical properties, such as surface gravity and effective temperature and from this allow us to map and understand the formation, structure, and evolution of our Galaxy down to at least a visual magnitude of 20 in the Gaia G-Band, as described in Prusti et al. 2016 and Rudolph et al. 2014.

The Gaia satellite (fig.2.14) was built under an ESA contract by Airbus Defence and Space (D&S) and was a truly European project. The payload module was built by Airbus D&S in Toulouse, the mechanical service module by Airbus D&S in Friedrichshafen, and an electrical service module by Airbus D&S in Stevenage.

Like the Herschel and Planck missions that came before it, Gaia flies on an Lissajous

orbit around the second Lagrange point (L2) in the Earth-Sun-Moon system which provides the stable thermal environment that is required to maintain high levels of accuracy, and minimises obstructions of the sky by the Earth and Moon. The Lissajous orbit is named after the French physicist Jules Antoine Lissajous, and is an orbit around a Lagrangian point in a three-body system that requires minimal propulsion input to maintain, thus making it ideal for spacecraft on long term station.

To achieve its astrometry goals as described in Rudolph et al. 2014 and Prusti et al. 2016, Gaia is equipped with two telescopes with astrometric viewing directions separated by a large angle, called the basic angle, of ~106.5°. The two telescopes share the same focal plane, in which the CCD detectors are located. With this arrangement the position of stars can be measured and compared between the two lines of sight which are arranged in a plane perpendicular to the spacecraft spin axis. The spacecraft performs a continuous scanning motion around the spin axis, which allows the fields of view to sweep across the sky with a speed of ~1° min^{-1}.

The measurement principle of Gaia is derived from the 'global-astrometry' concept successfully demonstrated by Hipparcos (Perryman 1989) and uses the principle of scanning space astrometry (SSA) Lindegren and Bastian 2010). SSA relies on a slowly spinning spacecraft that accurately measures the crossing times of targets transiting the focal plane by synchronising the optical measurements with an onboard atomic clock. These observation times represent the one-dimensional, along-scan (AL) stellar positions relative to the instrument axes (see fig. 2.15). The astrometric catalogue is built up from a large number of such observation times, by an astrometric global iterative solution (AGIS) process (e.g. Lindegren, Lammers, Hobbs, et al. 2012; Lindegren, Lammers, Bastian, et al. 2016), which also involves a simultaneous reconstruction of the instrument attitude as a function of time, and of the optical mapping of the focal plane detector elements (pixels) through the telescope(s) onto the celestial sphere (geometric calibration, see 2.16). The spin axis of the spacecraft rotates around the Sun direction with a period of ~63 days and this slow precession allows for full coverage of the sky (see 2.17). The speed of the precession is as small as possible to limit the across-scan

smearing of images when they transit the focal plane yet it is just large enough to ensure that subsequent loops overlap meaning that there are at least six distinct epochs of observations per year for any object in the sky.

In addition to the astrometric instruments Gaia also incorporates a radial velocity spectrometer and photometric instruments. The radial velocity spectrometer provides the third component of the star's velocity and the photometric instrument measures the spectra of each star several times allowing the determination of the physical and chemical characteristics of the observed stars.

This presents an opportunity to take new cluster candidates and, using a cross-match of sky coordinates, obtain proper motion and parallax information that can be used to help identify or confirm cluster members and in some cases provide a distance estimate ahead of more detailed follow-up study.

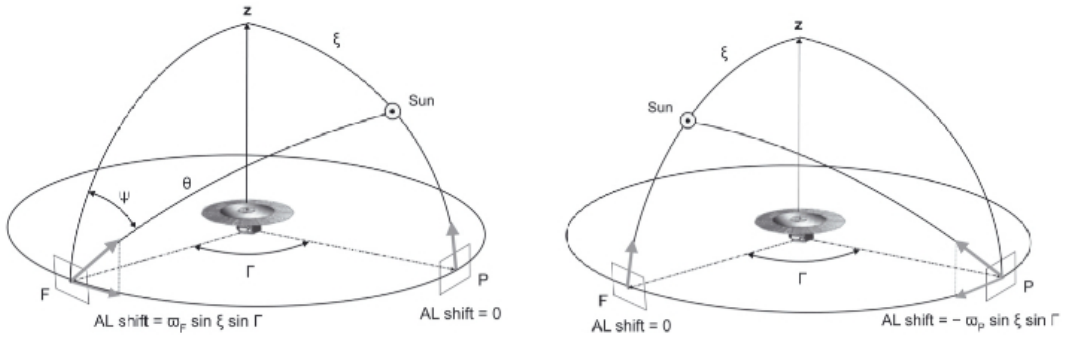


Figure 2.15 Measurable, along-scan (AL) angle between the stars at P and F depends on their parallaxes ω_P and ω_F in different ways depending on the position of the Sun. This allows us to determine their absolute parallaxes, rather than just the relative parallax $\omega_P - \omega_F$. Wide-angle measurements also guarantee a distortion-free and rigid system of coordinates and proper motions over the whole sky. See Lindegren and Bastian 2010.

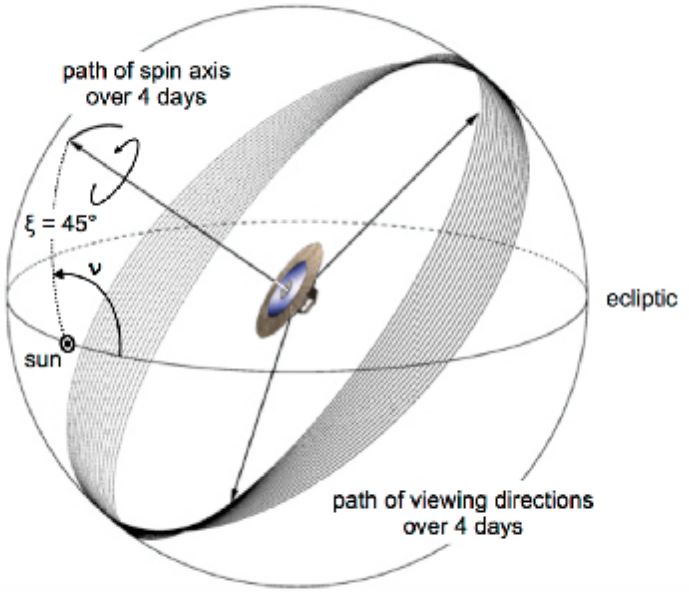


Figure 2.16 Illustration of the scanning law of Gaia, showing the path of the spin axis (z), and the corresponding path of the preceding viewing direction, during four days. For clarity, the path of the following viewing direction is not shown. See Prusti et al. 2016.

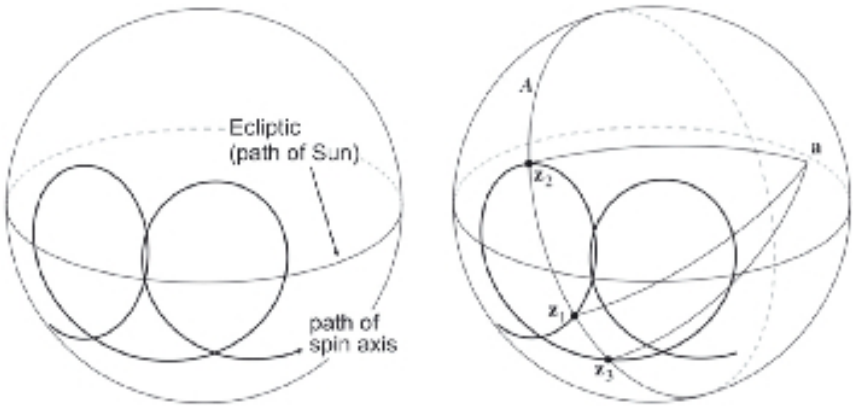
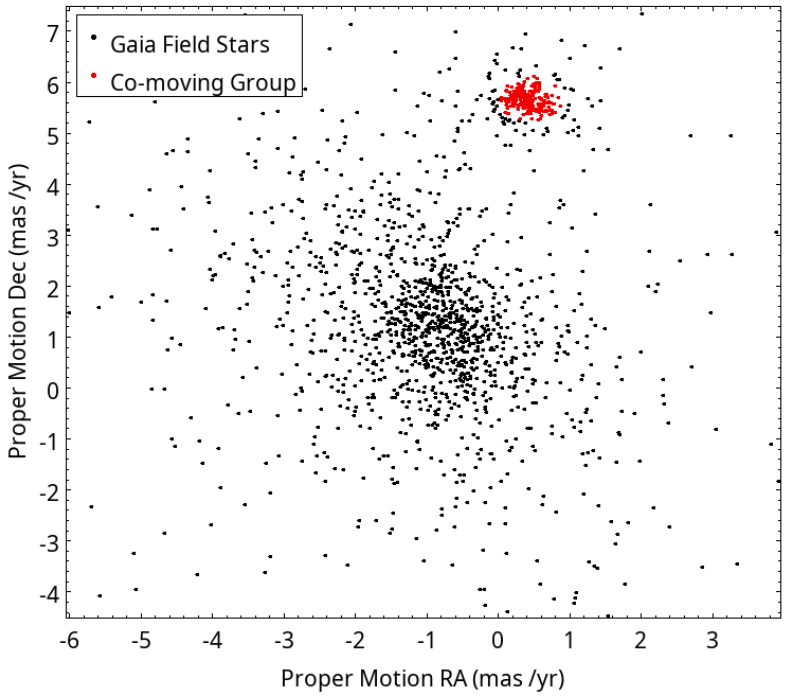


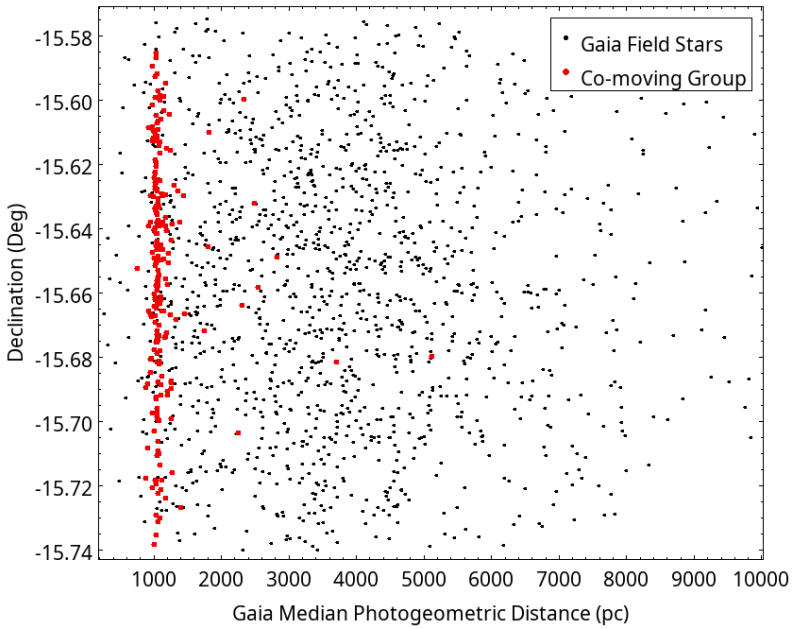
Figure 2.17 Spin axis z makes loops around the Sun, which must overlap. A star at point a may be scanned whenever z is 90° from a, i.e. on the great circle A at z1, z2, z3, etc. The scans intersect at a large angle, which allows the determination of two-dimensional object coordinates from one-dimensional measurements. See Prusti et al. 2016.

2.4.1 Using Gaia data in the confirmation and basic characterisation of cluster candidates.

Figure 2.18 provides an example of how Gaia proper motion and parallax data can be used to provide additional insight into the cluster candidates, both in terms of confirmation and characterisation. Figure 2.18 shows proper motion and photogeometric distance plots (see Bailer-Jones, Rybizki, Fouesneau, Demleitner, et al. 2021) for NGC-2360, an open cluster in the constellation Canis Major. Fig 2.18 subplot (*a*) shows the proper motions of all Gaia stars in a 4 arcmin radius field of the cluster centre. There can be seen a distinct sub-group of co-moving stars clumped in pm space around pmra of 0.45 mas yr^{-1}, pmdec of 5.6 mas yr^{-1}, toward the top right of the plot. This sub-group is moving independently of the main field of stars and this would strongly suggest they are related. The red highlighted stars lie within a radius of 0.5 mas yr^{-1} of (0.45,5.6) mas yr^{-1}, this figure was chosen as it captured the core of the clump. Fig 2.18 subplot (*b*) shows a declination vs photogeometric distance plot of the same 4 arcmin field with the red subgroup. We can see a clear vertical line of red points from the subgroup at approximately 1 kpc which is indeed the distance reported in the literature for NGC-2360 and clearly demonstrates how we can use Gaia proper motion and photogeometric distance data as part of the confirmation process for new cluster candidates. In many cases it may be impossible to make use of this technique as not all candidates will have members with Gaia data available, especially any distant or dense clusters with low visible luminosities. As a point of interest, it is worth noting than until the Gaia data was available the distance estimates for this cluster ranged from approximately 3.5 kpc to 6.2 kpc. See chapter 6 for a full discussion of the Gaia cross-match, plotting and distance estimates, including reliability.



(a) Proper motion plot of Gaia stars in 4 arcmin radius of cluster centre. Black points are Gaia field stars, note clump centred on pmra of 0.45 mas yr^{-1}, pmdec of 5.6 mas yr^{-1}, this is a sub-group with different bulk motion to the rest of the field. Red points are radius of 0.5 mas yr^{-1} centred on (0.45,5.6) mas yr^{-1}.



(b) Declination vs Photogeometric Distance plot of the same Gaia field. Red points are the sub-group from plot *a* above and clearly show almost all lie at the same distance of approximately 1 kpc.

Figure 2.18 Proper motion and photogeometric distance plots for NGC-2360. Full discussion is in §2.4.1.

2.4.2 Strasbourg astronomical Data Centre (CDS) and Online Astronomical Databases

The final sources of data used in this project are the Strasbourg astronomical Data Centre (Centre de Données astronomiques de Strasbourg or CDS) and the on-line Unified Cluster Catalogue (UCC). The CDS is dedicated to the collection and worldwide distribution of astronomical data and related information and hosts the SIMBAD astronomical database, the reference database for the identification of astronomical objects; VizieR, the catalogue service for the CDS reference collection of astronomical catalogues and tables published in academic journals; and the Aladin interactive software sky atlas for access, visualization and analysis of astronomical images, surveys, catalogues, databases and related data. The CDS mission is to collect useful information concerning astronomical objects, upgrade these data by critical evaluations and comparisons, distribute the results to the astronomical community, and conduct research using these data. The UCC was recommended by Ernst Paunzen (one of the WEBDA maintainers) as an alternative whilst WEBDA is offline. The UCC is the largest catalogue of stellar clusters currently listing nearly 14,000 objects (Perren et al. 2023). The UCC lists any object that was catalogued as an open cluster in the literature. If the objects have been classified differently in other articles (e.g. moving group or association) it will remain listed in the UCC because at least one article at some point indicated that it was an open cluster. The UCC is regularly updated to include new research.

SIMBAD (an acronym for Set of Identifications, Measurements and Bibliography for Astronomical Data), which is a database providing information on astronomical objects of interest which have been studied in scientific articles (Wenger et al. 2000). The related VizieR catalogue tool (Ochsenbein, Bauer, and Marcout 2000) provides the most complete library of published astronomical catalogues, tables and associated data (with verified and enriched data) accessible via multiple interfaces. Query tools allow the user to select relevant data tables and to extract and format records matching given criteria. Currently, 25839 catalogues are available.

Originating in 1972, SIMBAD was created by merging the Catalog of Stellar

Identifications (CSI) and the Bibliographic Star Index as they existed at the Meudon Computer Centre until 1979. The resulting data base was then expanded by the addition of source data from the many catalogues connected to the CSI and by new literature references. The database was extended to galaxies and other non-stellar objects in 1983. It has seen several iterations over the years until reaching its current form in 2006 as SIMBAD 4, developed in JAVA language, and using the open source PostgreSQL DBMS. SIMBAD was developed by Marc Wenger (retired), Anaïs Oberto, Grégory Mantelet (CDS, Strasbourg) along with help from their students.

The database is actively maintained by the research community and the CDS team and provides the bibliography, as well as available basic information such as the nature of the object, its coordinates, magnitudes, proper motions and parallax, velocity/redshift, size, spectral or morphological type, and the multitude of names (identifiers) given in the literature. The CDS team also performs cross-identifications based on the compatibility of several parameters, in the limit of a reasonably good astrometry.

The SIMBAD interface allows users to search for references based on a number of input parameters such as co-ordinates, known identifiers, object type, or combination thereof. It also provides a TAP (Table Access Protocol) service to allow ADQL (Astronomical Data Query Language) queries as well as remote queries via Virtual Observatory connection from data tools such as TOPCAT (Taylor et al. 2005).

In addition to the UCC I also used the DAML02 catalogue (Dias et al. 2002) as this has a significant cross over with SIMBAD/Visier but are valuable sources of data on open clusters. Although SIMBAD contains some of the Gaia proper motion data, DAML02 is also useful as it includes additional data such as proper motions from other sources which are essential for helping to separate co-moving groups and clusters from the field. The DAML02 catalogue, formally titled "New Catalogue of Optically Visible Open Clusters and Candidates," represented a significant advancement in the compilation of open cluster data, particularly by integrating kinematic information that was largely absent in earlier compilations.

The primary objective behind the compilation of the DAML02 catalogue was to

update and expand upon existing open cluster catalogues, notably WEBDA, Mermilliod (1995). A critical innovation of DAML02 was the inclusion of new data on kinematics, specifically proper motions, which had not been systematically present in previous compilations. Kinematic information is fundamental for distinguishing true cluster members from foreground and background field stars and by incorporating this, DAML02 significantly improved the reliability of cluster membership determination and, consequently, the accuracy of derived cluster parameters such as distance, age, and metallicity.

I have also made use of available catalogue sources to gather physical properties of known clusters and candidates in my sample. Whilst there is inevitable cross-over between the catalogues, cross matching them gives a more complete set of parameters. Catalogue sources included the FSR catalogue (Froebrich, Scholz, and Raftery 2007), the Hunt & Reffert 2023 catalogue of Gaia DR3 open clusters (Hunt and Reffert 2023) and the Milky Way Star Cluster Catalogue (Kharchenko et al. 2013).

Chapter 3

Initial Search, Classification and Field Star Decontamination

The list of cluster candidates identified as part of this work emerged from the visual quality control checking of the UKIDSS GPS Data-Release 10 (DR-10) multiframe images, prior to their release to the WSA. As described in chapter 2, each of the leavstack files contains four images, one from each of the four WFCAM detectors at the observed wavelength. Quality checking these images involves visually looking for aberrations, missing data and other quality related issues.

During these visual checks I noticed a number of potential young clusters, primarily in the K-band multiframes, as small over-densities with some associated K-band nebulosity (see Fig. 3.1 for example). Cross-checking this against astronomical databases such as SIMBAD, Visier and WEBDA revealed that these were previously unknown objects, and suggesting that the DR-10 data held important new discoveries. From this I embarked on a systematic review of all the images to collate a catalogue of candidates for follow-up.

The initial selection criteria were fairly simple, I was looking for stellar over-densities within the J, H or K band images, with or without associated K-band nebulosity or dark IR clouds (these show up as regions of very low stellar density, particularly in K-Band images). If an over-density appeared to cross the detector

boundary I would look for an image cut-out from the WSA using coordinates that were as close to a visual centre as possible (the reported central coordinates were determined later using the methods stated in section 5.1). There are however limitations to the size and density of clusters that can be determined by eye. The source images are typically 10 x 10 arcmin or smaller and this imposes an upper limit to the angular size, although clusters with a larger angular diameter can be identified if they have a dense central concentration, and indeed this was the case for a number of NGC clusters with angular diameters > 10 arcmin identified as part of the search. There will however be cases where less populous open clusters are not identified because their size is on a scale larger than the typical image size, or the stellar radial density lacks a strong central concentration.

It should therefore be noted that previous attempts have been made to search UKIDSS data for clusters using automated techniques, such as Lucas P. W. et al. (2013) and Solin, Ukkonen, and Haikala (2012) using a Bayesian algorithm or Gaussian Mixture Model (GMM) respectively. While these attempts have had some success they are ultimately limited by the algorithms inability to select candidates where there may be significant K-band nebulosity, extended sources or obscuration by associated dust. It is therefore likely that many of the more open, less obscured clusters will have been discovered already by these techniques and whilst I have recorded all the candidates visible in the images, really the focus of this work is on identifying those candidates and clusters that are apparent, currently, only to the human eye. This is important in order to extend the completeness of my sample of young, emerging, Galactic clusters and feed into further work on Galactic star formation.

In order to identify sub-clusters that might be related to, or part of, the same structure any of the candidates within 2 arcmin of each other were treated as related and merged in the list, with the central coordinates being updated to reflect this (see more detailed description in section 5.1). Star forming regions which extend over many arcmins' can of course contain multiple clusters, and these are treated as separate entities in this work, provided the determined coordinates are > 2 arcmin distant and there is no obvious visual overlap of members between them.

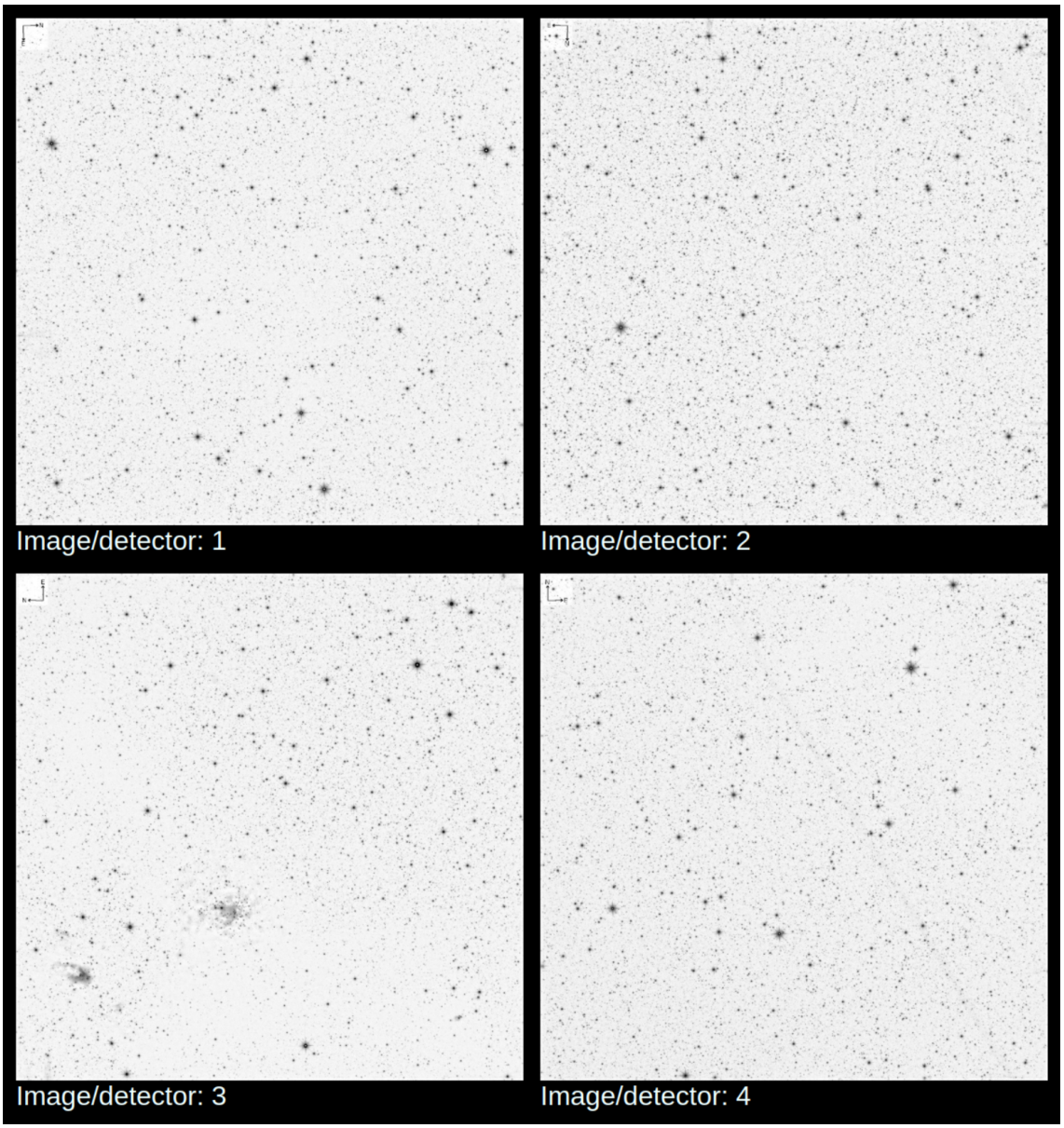


Figure 3.1 A K-band leavstack multiframe showing images from the four WFCAM detectors (candidate-3). Two distinct over-densities of stars with associated nebulosity can clearly be seen towards the lower left of detector 3. This would be marked for follow up as described in the text.

Figure 3.2 shows the basic distribution of the candidates across the galactic plane. As can be seen there is a good distribution across the galactic plane within the survey boundary. The GPS is defined over $|b| < 5°$, but because it was observed as discrete WFCAM tiles the realised footprint extends marginally beyond this nominal limit, reaching $|b| \approx 6°$ across most of the surveyed longitude range, with more substantial extensions towards the Galactic centre ($l < 20°$) and across $140° < l < 200°$. The

distribution has three notable areas of lower density, one at approximately 0 to 25 degrees around the galactic centre, where the fields become extremely crowded and visually identifying over-densities becomes much harder, the other two are between 160 and 170 degrees towards the galactic anti-centre and an under-density in the mid-plane from $50 < l < 107$. Both of these were areas covered by a previous search of the DR4 data release that I completed in conjunction with my supervisor, Prof. Lucas.

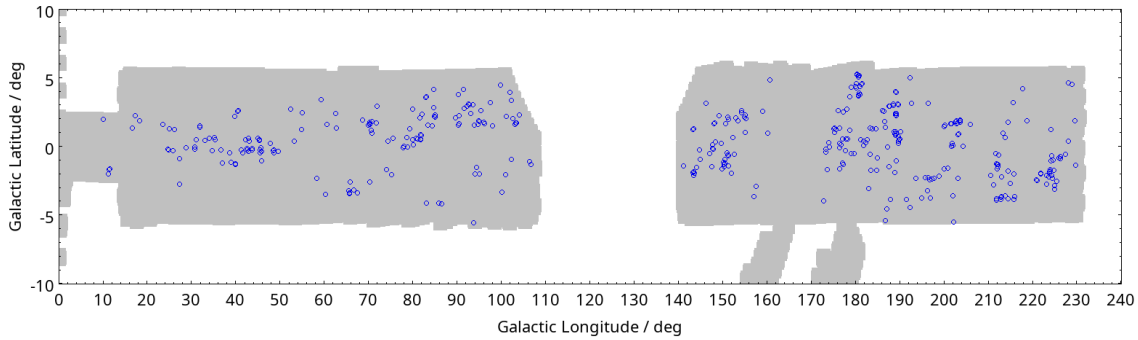


Figure 3.2 Galactic Lat/Long plot showing GPS DR10 coverage as the grey area and the distribution of the 403 cluster candidates overlaid as blue circles. The shaded footprint is drawn from the individual WFCAM tile boundaries.

3.1 Initial Coordinates and Apparent Radius

In order to allow further investigation of the candidates I needed to establish a more robust initial set of coordinates for each. The eye can give a fairly good approximation based on visual density, but this is subject to individual bias. An algorithmic approach removes human bias but can find young, partially obscured, clusters problematic because of gaps in the data. This then presents us with a dilemma: in order to make a proper calculation of the cluster centre we ideally want to remove the field stars to ensure the algorithm is not biased by stars unrelated to the cluster, but, in order to give it the decontaminated cluster stars we initially have to use an estimated centre.

The compromise here was to take data in a 5 arcmin radius of the visual centre and run this through an algorithm to give a repeatably determined initial centre for the decontamination algorithm to use. This was then recalculated after decontamination and

the decontamination re-run if the difference between the two was greater than 25% of the apparent (decontaminated) radius, and this process is described in more detail in §3.2.4. Several different algorithms were tested to determine which was best able to manage the variety of morphologies in the sample and this process is described in detail in §5.1.

In addition to determining central coordinates for the decontamination it is necessary to determine an initial apparent radius to each candidate. This is used by the algorithm to set the target area vs the comparison field annulus. As with the central coordinates I initially assigned an approximate apparent radius based on visual assessment of the stellar distribution around the visual centre. I initially tried a fixed radius across all candidates but this tended to give poor quality results with spurious member selections. Spurious member selection is a constant problem when using a colour-based algorithm and is caused by a number of factors which are discussed further in §3.2.5.

In practice getting both of these parameters right required some test iterations to become comfortable that the initial coordinate selection and target radius were reliable and did not affect the outcome of the decontamination process. Once the decontamination was complete I was then able to algorithmically determine a more reliable centre, and this process is described fully in §5.1.

3.2 Field-star decontamination

Field-star decontamination is the principal problem when trying to confirm and characterise new cluster candidates and this is particularly true in crowded fields towards the Galactic plane or centre and for those candidates with a large fraction of faint stars or obscuring dust and gas (see e.g. Bonatto and Bica 2007, Froebrich and Schmeja 2010, Borissova, Ivanov, et al. 2018). For open clusters it is also easy to confuse a potential cluster with a natural field over-density and vice versa, especially for more evolved clusters in the process of dissolution. The end goal of the decontamination is not to produce a complete list of cluster members as this would require detailed spectroscopic study of each star in the target area. Instead what we want to achieve is a group of statistically likely member stars that can reveal the shape of the intrinsic CMD and CCD

of the cluster. These can then be used in conjunction with other techniques, such as isochrone fitting, to obtain an approximate age and mass distribution, among other things, and provide a starting point for more detailed analysis.

The list of cluster candidates produced by the initial search contains a wide range of morphologies with potential open clusters of varying spatial density, as well as young, rich, clusters with high and variable extinction. There are many approaches to field-star decontamination (just referred to as *decontamination* from here on), ranging from the old visual light eyeball selection used by Bergh and Hagen (1975) through to complex statistical methods such as Bonatto and Bica (2007). For this project I looked at two loosely related statistical methods, one relatively straight forward and one fairly complex. Later I also looked at another, modern, machine learning technique using HDBSCAN to see if this could offer any improvement to the detection quality, and this is described in Chapter 7. After all, it is not expected that any method will select all true cluster members. The aim is simply to provide a sample that approximates the real CMD/CCD to enable further analysis, and I am using a combination of techniques to try to build a convincing case for any new clusters.

3.2.1 Colour-Colour-Magnitude Method (CCM)

This decontamination process is a Colour-Colour-Magnitude (CCM) statistical technique developed by Froebrich and Schmeja (2010) and is based on the procedure developed by Bonatto and Bica (2007) and described below. This approach uses the J, H and K-band magnitudes for all stars in the initial target region (the radius from the central coordinates believed to contain the cluster) and compares these to the magnitudes in a control region, separate from the target region. The comparison is made via a vectorisation of the root of the sum of the differences between the colours in the regions.

I ran a series of tests across a sample from all of the candidate morphologies and while it is easy to work with and works well in the case of open clusters and open cluster candidates I found that the results for the emerging clusters had fewer members identified, more spurious inclusions and less convincing CMDs.

3.2.2 Bonatto Method

The decontamination algorithm I decided to use for the bulk analysis is also a statistical decontamination technique is one developed by Charles Bonatto and Eduardo Bica (see Bonatto and Bica 2007 and references therein for details). This method has a well established history and has provided good results across a number of different environments, see Bonatto and Bica 2010 and references therein. I made some minor changes to the code (provided by C. Bonatto, private comm.) to ensure it worked correctly with the UKIDSS band passes and, as with the CCM code, I produced a second, modified version that uses just the H and K-band photometry for those candidates where there is a large amount of extinction and no J-band detections.

The Bonatto decontamination method is also a statistical algorithm which uses the relative number densities of field and cluster stars based on colour comparison between a field region and a target region approximately matching the clusters apparent radius. The field in this case is defined as an annulus of 1 arcmin radius around the target region which is local enough that the extinction should remain fairly constant, although with deeply embedded young clusters the extinction varies over very small apparent distances. These cases have to be handled with care as decontamination using an automated process may be impossible. Cases where this is suspected have been highlighted in the list in appendix A.

As previously mentioned the cluster candidates in this sample have a wide variety of apparent radii and morphologies so making a robust determination of the central coordinates and radius can be challenging. As already outlined I used an iterative process starting with an empirical estimate of the centre and radius for the decontamination which was then iterated to achieve the best signal to noise ratio in the resultant CMDs and CCD. I investigated several different quantitive methods to determine reliable central coordinates and radius and this process is described fully in sections §5.1 and §5.2.

At a high level the Bonatto algorithm works in three stages. First it takes the full range of CMD colours and magnitude and produces a grid with axes of J, J-H and J-K. It then looks at the field stars in the control region (annulus in this case) and computes an

expected density of field stars for each comparable cell in CCM space. Finally, it randomly subtracts the number of field stars it would expect to find in a given cell. All magnitudes are UGPS `apermag3` magnitudes.

To illustrate this in more detail consider a cell with dimensions:

$$\Delta J = 0.5, \Delta(J-H) = \Delta(J-K) = 0.2 \tag{3.2.1}$$

and whose sides have coordinates in the $(J, (J-H), (J-K))$ space of:

$$\left(J_c \pm \Delta\frac{J}{2}, (J-H)_c \pm \Delta\frac{(J-H)}{2}, (J-K)_c \pm \Delta\frac{(J-K)}{2} \right) \tag{3.2.2}$$

where $(J_c, (J-H)_c, (J-K)_c$ are the cell's central coordinates. It then uses a Gaussian distribution of magnitude and colours:

$$P\left(J, \bar{J}\right) = \frac{1}{\sqrt{2\pi}\sigma_J} e^{-\frac{1}{2}\left(\frac{J-\bar{J}}{\sigma_J}\right)^2} \tag{3.2.3}$$

and so on for the other two dimensions, $P\left(J-H, \bar{J-H}\right)$ and $P\left(J-K, \bar{J-K}\right)$, to compute the probability of a star with CMD coordinates of

$$\left(\bar{J} \pm \sigma_J, \left(\bar{J}-\bar{H}\right) \pm \sigma_{(J-H)}, \left(\bar{J}-\bar{K}\right) \pm \sigma_{(J-K)}\right) \tag{3.2.4}$$

to be found within that cell, where σ is the `apermag3err` uncertainty in the `apermag3` magnitude. We assume that probable cluster stars are located within a radial distance of the cluster centre so that the radial distance from the central spatial coordinates of the cluster is given by $R_{cl*} < R_{lim}$ with area of A_{cl}, where R_{lim} is the radius of the target area discussed above. Field stars are taken from the annulus (or offset region) where $R_{lim} < R_{fs} < R_{max}$ with area of A_{fs}, where R_{max} is the outer radius of the field annulus.

The expected number density of field stars $\left(\rho_{fs}^{cell}\right)$ in a given cell is calculated by summing the probability of all field stars to be found in a cell, divided by the annulus field area

$$\rho_{fs}^{cell} = \frac{P_{fs}^{cell}}{A_{fs}} \tag{3.2.5}$$

where

$$P_{fs}^{cell} = \sum_{i=1}^{N_{fs}} \iiint P_i\left(J, \bar{J}_i; (J-H), \left(\bar{J}-\bar{H}\right)_i; (J-K), \left(\bar{J}-\bar{K}\right)_i\right) \, dJ \, d(J-H) \, d(J-K) \tag{3.2.6}$$

$P_i\left(J, \bar{J}_i; (J-H), \left(\bar{J}-\bar{H}\right)_i; (J-K), \left(\bar{J}-\bar{K}\right)_i\right)$ is the probability of the ith offset field star, with CMD coordinates of $\left(J, \bar{J}_i; (J-H), \left(\bar{J}-\bar{H}\right)_i; (J-K), \left(\bar{J}-\bar{K}\right)_i\right)$ with uncertainties $\left(\sigma_{J_i}, \sigma_{(J-H)_i}, \sigma_{(J-K)_i}\right)$ to have the magnitudes and colours $(J, (J-H), (J-K))$. The integrals are calculated across the cell's extension in each dimension, i.e.

$$J_c - \frac{\Delta J}{2} \leq J \leq J_c + \frac{\Delta J}{2} \tag{3.2.7}$$

$$(J-H)_c - \frac{\Delta(J-H)}{2} \leq (J-H) \leq (J-H)_c + \frac{\Delta(J-H)}{2} \tag{3.2.8}$$

$$(J-K)_c - \frac{\Delta(J-K)}{2} \leq (J-K) \leq (J-K)_c + \frac{\Delta(J-K)}{2} \tag{3.2.9}$$

The process is then repeated in the cluster region, $R < R_{lim}$ where the number density of cluster and field stars is calculated for each cell

$$\rho_{obs}^{cell} = \frac{P_{obs}^{cell}}{A_{cl}} \tag{3.2.10}$$

and the expected number of field stars in the cell is given by

$$n_{fs}^{cell} = \frac{\rho_{fs}^{cell}}{\rho_{obs}^{cell}} \times n_{obs}^{cell} \tag{3.2.11}$$

where n_{obs}^{cell} is the number of observed stars (field and cluster) in the cell and the number of probable cluster member stars in the cell is

$$n_{cl}^{cell} = n_{obs}^{cell} - n_{fs}^{cell} \tag{3.2.12}$$

We can then calculate the total number of probable cluster members by summing n_{cl}^{cell} over all CMD cells

$$N^{cl} = \sum_{cell} n_{cl}^{cell} \tag{3.2.13}$$

This whole process is then repeated for three different grid sizes in each dimension in order to minimise any selection effect from the input values. this means that for instance a CMD grid starting at J_o with cell width of ΔJ there would be additional runs for $J_o \pm \frac{1}{3}\Delta J$ and this applies to all three dimensions. The combination of 3 grid sizes and 3 starting positions in 3 dimensions results in $3^6 = 729$ different configurations and from these runs it then calculates the average number of probable cluster stars $\langle N_c l \rangle$.

This standard method is referred to as the 3-dimensional (3D) method as it uses the J, J-H and J-K values for the decontamination. As mentioned above I also produced a second version of the code that uses only H and H-K values which is referred to as the 2D method. This appears to offer some additional advantage when selecting probable members for young, emerging clusters with significant extinction. The typically high extinction in the J-band of emerging clusters means that detected sources strong in the J band are disadvantaged when the code is making the comparison between sample and field areas. This deficiency adversely affects the results and can produce poor member selection of highly reddened objects, not to mention many of which could have no J-band detection at all.

3.2.3 Decontamination Examples

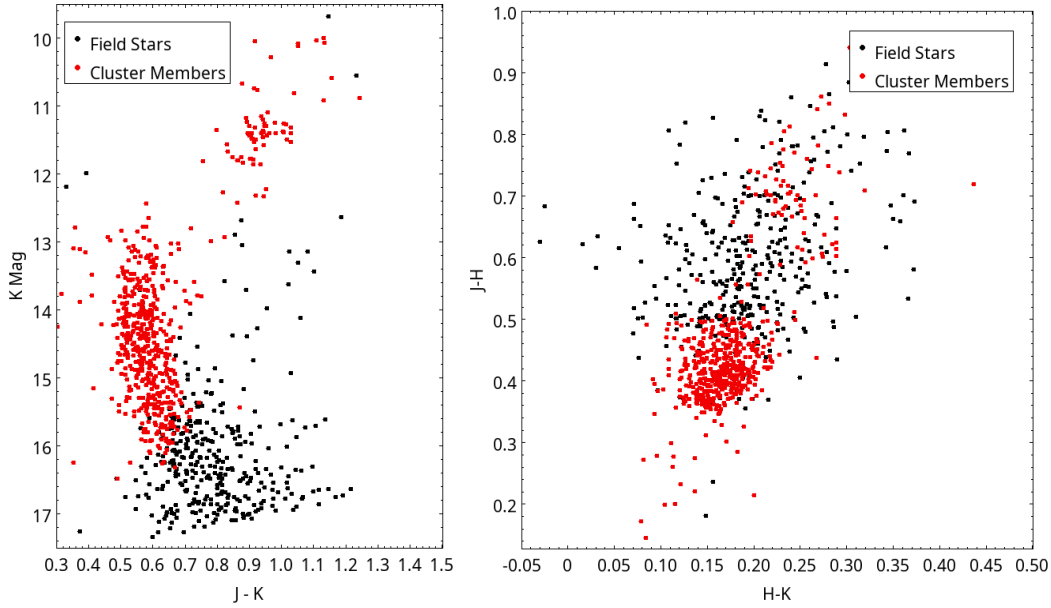
The range of candidate morphologies is great but can be broken down into the five categories described in §1.2.1. Below I have included and example of a known Open Cluster and a known emerging cluster and the decontamination output, along with a description of what we would expect to see for clusters of the given age and an analysis of the decontamination output.

Open Cluster (OC) - Candidate 75

Candidate 75 is known open cluster NGC 7044 in Cygnus. It has an apparent diameter of 6 arcmin and an estimated 742 visible light members (Casali, Magrini, et al. 2020, Aparicio et al. 1993 and Tarricq et al. 2021). The cluster has an age of approximately 1.7×10^9 years (Tarricq et al. 2021) which would mean that it is well evolved and we should see a main-sequence turn-off, a Red Giant Branch (RGB) and a Red Giant Clump (RGC).

The GPS data extraction and pre-cleaning results in 1774 stars with good JHK photometry in the 3 arcmin radius of the estimated central coordinates. The number of cluster members is expected to be in the range of 740, or approximately 60% of the GPS stars (making allowance for poor photometry). The high-probability members (P_{high}) are

shown as red points in the plots, with the field stars as black points.



(a) K vs J-K CMD, Bonatto Algorithm **(b)** J-H vs H-K, CCD, Bonatto Algorithm

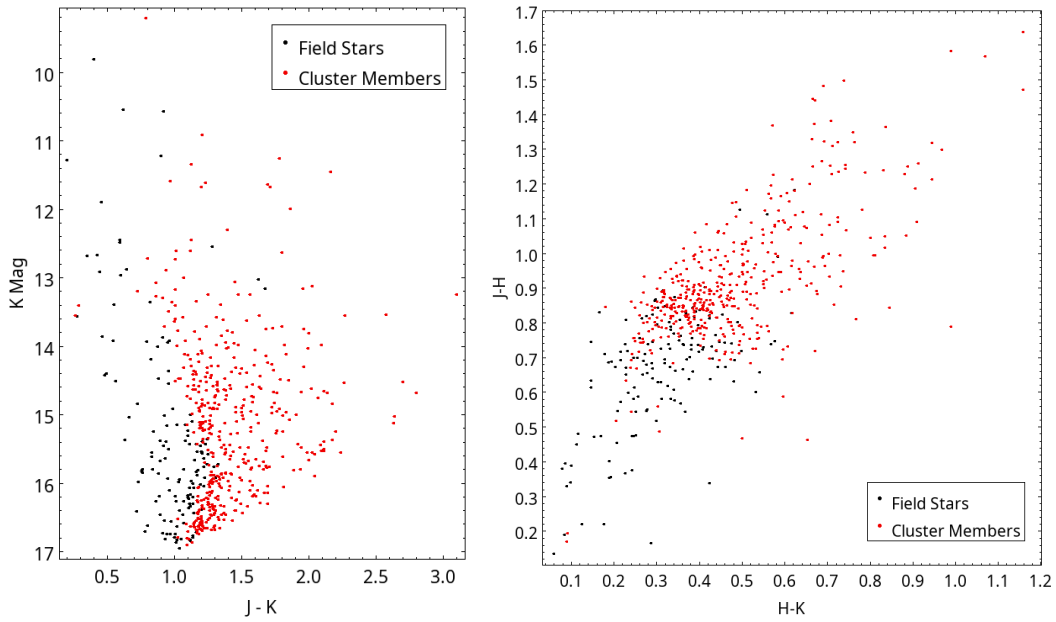
Figure 3.3 Candidate 75 (NGC 7044) - 3D decontamination plots from the Bonatto code showing the JK CMD and the CCD. The CMD shows field stars from the 3 arcmin radius target area (black) plotted against the high probability cluster members (red). The high-probability members lie primarily in the main sequence as expected for an open cluster, however in this instance the code has selected almost all the stars in the RGB and RGC. This is certainly in line with the expectation for a cluster of this age and stage of evolution.

The CMD in Figure 3.3 shows the outputs of the Bonatto algorithm as a K vs J-K plot in sub-plot *a* and we can see that the selected member points lie primarily in the main sequence as expected for an evolved cluster, but the selection of almost all of the brighter members in RGB and RGC is pleasing as this is exactly what would be expected of a cluster of this age and stage in its evolution. The CCD in sub-plot *b* shows a tight grouping of the (P_{high}) stars in colour space with minimal reddening, as would be expected for an open cluster with uniform extinction and also shows the redder population of evolved stars, likely from the RGB and RGC, further to the top right of the diagram. The algorithm selected 1106 members of the 1774 in the target region and this feels about right given the 742 members identified in visible light using data which is of lower resolution and depth.

Embedded Cluster (EC) - Candidate 6

Candidate 6 is known emerging cluster NGC 1931 in Auriga. It is both a reflection and emission nebula around a young star cluster and contains a small group of hot young stars. The exact number of members is not yet clear as it is an active star forming region, Pandey et al. (2013) found 22 probable visual members versus Bonatto and Bica (2009) who found ≈ 230 members for instance, which highlights the additional value high resolution infrared can bring to the investigation of this type of young cluster (although in this case the data came from 2MASS so we would expect an increase in detections with the higher resolution and depth of the GPS data). It has an apparent diameter of 4 arcmin and is approximately 10×10^6 years old (Bonatto and Bica 2009). For a cluster at this early stage of evolution we would expect to see a relatively small number of massive, luminous stars just starting to reach the main-sequence turn-off along with many more low mass stars still in the YSO phase, just making their way towards the main-sequence. There would likely be high and variable extinction across the cluster which could cause problems for colour-based decontamination. The GPS has approx 1102 stars with good photometry in the 2 arcmin radius of the estimated central coordinates.

The CMD in Figure 3.4 shows the outputs of the Bonatto algorithm as a K vs J-K plot in sub-plot *a* and we can see that there are a handful of brighter members with relatively low reddening which would be the higher mass stars beginning to evolve off the main sequence. We would expect them to have cleared their local region of the majority of the progenitor gas and dust, meaning that the colours are far less likely to result from extinction, and this is supported by the fact that they are the member stars with low colour excess in the CCD of sub-plot *b* (although this is not differentiated in these plots). The fainter members all show significant reddening in the CMD and significant colour-excess in the CCD indicating they are heavily extincted and this is what we would expect for an emerging cluster still partially shrouded in gas and dust. The decontamination selected 453 likely members which is a significant improvement on the 230 members identified by Bonatto and Bica (2009) and is likely the result of the better angular resolution and depth of the GPS data. The 22 visible light members identified by



(a) K vs J-K CMD, Bonatto Algorithm

(b) J-H vs H-K, CCD, Bonatto Algorithm

Figure 3.4 Candidate 6 (NGC 1931) - 3D decontamination plots from the Bonatto code showing the JK CMD and the CCD. The CMD shows field stars from the 2 arcmin radius target area (black) plotted against the high probability cluster members (red). The majority of the high-probability members are significantly reddened with a few brighter members slightly reddened and this is what we would expect from a cluster of this age. The CCD shows a similar picture with selected members showing significant reddening with some members below the reddening corridor showing significant colour excess associated with very young stars. This is certainly in line with the expectation for a cluster of this age and stage of evolution.

Pandey et al. (2013) also aligns with the approximately 15 brighter sources in the CMD.

3.2.4 Bulk Decontamination Process

Both the tests using known clusters of different ages were successful and demonstrate the ability of the algorithm to determine a reasonable set of high probability cluster members. Manually applying this process to the 403 candidates in my sample would be tedious and error prone so I developed an automated process that takes source data from GPS and 2MASS, merges it and removes any bad data (as described previously in §2.2.1) then processes each candidate using the seed coordinates and apparent diameter as described in §3.1.

Once the initial decontamination is run the algorithm described in §5.1 is used to

recalculate the centre and the decontamination re-run if the difference between the two was greater than 25% of the apparent (decontaminated) radius. After any iterations of coordinates the decontamination automation outputs a series of data files and CMD, CCD plots, and these are available in the accompanying file archives.

3.2.5 Summary of the Decontamination Process

Initially I looked at the two statistical decontamination methods, both are related but one offered a simplified approach. Both have been used successfully (see references in the preceding text), but because of the variable morphology of the candidates it was important to determine if one method was able to produce reasonable results in all cases. A single method would simplify the automation of the decontamination process, although two could be supported if there was justification.

After testing it was clear that the Bonatto method produced clearer results with fewer spurious inclusions across all the different morphologies. The decontamination process is hard, especially in young, emerging, clusters with high and variable extinction, and it is well understood that any statistical, colour based algorithm will really only ever be able to remove the least convincing stars. This is usually sufficient to be able to decide if the high-probability candidate members display the expected features, and therefore if it looks real or not.

Each of the 403 output plots was examined with reference to the initially assigned class of candidate and notes made as to whether the plots support the visual assessment or not. Both 3 and 2-dimensional decontaminations and plots are used, in the case of ECC and EGr candidates the 2D decontamination is essential as many of these lack J-band detections because of the high extinction. The results of this analysis are available as part of the table in Appendix A and form a significant part of the assessment of potential new candidates. The full characterisation and analysis process is discussed in more detail in Chapters 5 and 8 respectively.

Chapter 4

Isochrone Analysis

In this chapter, I describe the methodology and results for the stellar population analysis of the cluster candidates. By fitting theoretical stellar evolution models to the observed near-infrared (NIR) photometry, I derive the primary physical parameters of the clusters, including age, visual extinction, and distance modulus. I have developed a 3D dust-constrained optimisation pipeline specifically to overcome the classical age-distance-extinction degeneracy that complicates infrared cluster studies in the Galactic Plane, resolving the parameter space where physical properties are otherwise degenerate.

4.1 Introduction

Near-infrared isochrone fitting is a fundamental tool for characterising star clusters in the Galactic Plane, where high interstellar extinction obscures clusters in the optical bands. However, NIR-only photometry (J, H, K) presents a significant challenge: a three-way degeneracy between age, distance, and extinction. A young, distant, and highly reddened cluster can often appear indistinguishable from a nearby, evolved cluster with lower extinction when viewed through only two or three IR filters.

To overcome this limitation, I have developed a multi-stage pipeline that combines the high-resolution photometry of the UKIDSS Galactic Plane Survey (UGPS) with 3D line-of-sight dust constraints from the Green et al. (2019) Bayestar19 map and the YSO

fraction (or Q index, as defined by Negueruela et al. (2007)). This approach allowed me to break the NIR degeneracy by forcing the optimiser to align with the physically expected dust distribution of the Milky Way, providing robust parameters for the 178 high-confidence new cluster candidates.

4.2 Methodological Framework

4.2.1 Theoretical Models and Metallicity

The stellar populations are modelled using the MIST (MESA Isochrones and Stellar Tracks) project isochrones (Choi et al. 2016). These models are based on the MESA stellar evolution code and provide a consistent set of tracks from the pre-main sequence (PMS) through to the red giant branch. I have assumed a Solar metallicity ($[Fe/H] = 0.0$) as a baseline for Galactic Plane clusters, consistent with the radial metallicity gradient meta-analysis by Spina, Magrini, and Cunha (2022) and the finalised Gaia-ESO Survey results (Magrini et al. 2023). This assumption is necessary due to the lack of spectroscopic data for the candidates.

4.2.2 Reddening and Extinction Law

Interstellar extinction is corrected using the updated NIR extinction law defined by Wang and Chen (2019). The coefficients relative to visual extinction (A_V) are adopted as: $A_J = 0.224 \times A_V$ and $A_K = 0.072 \times A_V$. These constants define the reddening vector in the $(J - K, K)$ colour-magnitude space. While the Green et al. (2019) maps provide the magnitude of visual extinction expected at a given distance, the Wang & Chen law determines the specific wavelength-dependent attenuation applied to the theoretical isochrone during the fitting process.

4.2.3 Population Diagnostics: The Q-Parameter and Multi-Wavelength Cross-Checks

Before performing a full isochrone fit, every cluster is subjected to a reddening-free diagnostic using the Q index, where $Q = (J - H) - 1.70 \times (H - K)$. The index was defined by Negueruela et al. (2007) for OB stars and subsequently applied to infrared cluster searches by Borissova, Bonatto, et al. (2011). This parameter is independent of interstellar reddening and allows for the identification of Young Stellar Objects (YSOs) with infrared excess. Stars with $Q < -0.05$ are flagged as potential YSOs. If more than 10% of the decontaminated cluster members exhibit a YSO signature, the cluster is flagged as a potential star-forming region.

However, in environments with extreme differential reddening, photometric scatter can artificially inflate the Q-parameter, leading to false-positive young classifications. To ensure physical accuracy, the photometric Q-parameter is strictly cross-referenced against external multi-wavelength data. Specifically, the candidates were cross-matched against diffuse mid-infrared nebulosity in WISE imagery and the SIMBAD database for known star-formation markers (e.g., documented HII regions or established YSOs). If a cluster exhibited a high Q-parameter but clearly lacked any external WISE nebulosity or SIMBAD markers, the external data acted as a ground-truth veto, and the cluster was classified as evolved. While this external veto effectively prevents false-positive young classifications caused by differential reddening in nearby fields, it represents a conservative approach. First, multi-wavelength databases like SIMBAD may simply lack catalogued YSOs for newly discovered or compact clusters that have not been included in published targeted surveys (e.g., Spitzer). Second, visual inspection of WISE imagery focused on diffuse nebulosity and HII emission; however, pre-main-sequence clusters at intermediate young ages ($\sim$ 6–10 Myr) will typically have dispersed both their natal HII regions and individual circumstellar YSO discs ($\leq$ 6 Myr). Furthermore, because WISE has significantly lower spatial resolution than UGPS ($\sim 6''$ vs $0.8''$), it can suffer from source confusion and blending in compact or distant cluster cores. Consequently, young clusters older than $\sim$ 6 Myr or distant compact systems whose emission falls below these

thresholds may be conservatively reclassified as evolved under these criteria. Conversely, the presence of explicit WISE HII nebulosity or SIMBAD star-formation markers flagged a cluster as young, overriding the photometric diagnostic if the core was too heavily obscured for the Q-parameter to detect. For clusters flagged as young by these combined multi-wavelength criteria, an upper boundary age constraint of $\log t \leq 8.0$ (~100 Myr) is strictly enforced during the optimisation to prevent the algorithm from converging on unphysical, degenerate old-age local minima.

4.2.4 2D Euclidean Fitting Metric

Traditional isochrone fitting often relies on vertical interpolation, which predicts a colour for a given magnitude. This method is numerically unstable for young clusters where the PMS tracks are often vertical in the CMD. To resolve this, I employed a 2D Euclidean distance metric, following Frayn and Gilmore (2002). For every observed star, the algorithm uses a KD-Tree search (Bentley 1975), similar to HDBSCAN (see §7.2.5), to find the closest point on the theoretical isochrone in $(J-K, K)$ space. The χ^2 objective function is defined as the weighted sum of these 2D distances, making the fit robust against both vertical and horizontal scatter. To account for the varying richness of the candidate sample, I employed the reduced chi-square (χ^2_ν) statistic as the primary metric for fit quality. By dividing the total weighted sum of squared distances by the degrees of freedom, this approach normalises the error and allows for a consistent comparison of fit robustness across clusters with vastly different membership counts, following the methodology established by Kharchenko et al. (2013).

4.2.5 3D Dust Map Integration and Reproducibility of the '3D Anchored Fit'

A critical aspect of the pipeline is the integration of the *Bayestar19* 3D dust map (Green et al. 2019) to perform what is termed the '3D Anchored Fit'. First, the line-of-sight extinction profile showing cumulative visual extinction (A_V) as a function of distance modulus (DM) is queried from Bayestar19 for the cluster candidate's Galactic coordinates

(l, b). Second, the physical distance modulus DM and foreground extinction A_V anchor the theoretical Zero-Age Main Sequence in parameter space, applying a steep Bayesian penalty to the objective function if the optimiser deviates from the 3D map's constraints. Third, with the vertical and horizontal positions anchored, the Nelder-Mead simplex optimiser (Nelder and Mead 1965) evaluates the age ($\log t$) across the MIST theoretical isochrone grid using the 2D Euclidean KD-Tree metric. The MIST theoretical isochrones represent a continuous stellar population across the full Initial Mass Function ($0.1\,M_\odot$ up to the age-appropriate turn-off mass). In observed infrared CMDs, survey saturation thresholds ($K \lesssim 11$) and magnitude limits ($K \gtrsim 18$), combined with the statistical rarity of massive stars in modest-sized clusters, mean that observed members often span only a subset of the mass spectrum. Because the 3D anchored fit locks the curve's distance and extinction to the physical line-of-sight dust distribution, displaying the full theoretical track naturally extends across parameter space beyond the observed stars, showing the complete physical model across all stellar masses. Consequently, empty sections along the upper and lower extents of the theoretical track are observationally expected: the bright end ($K \lesssim 11$) is depleted by UGPS detector saturation and the statistical rarity of massive stars in modest-sized clusters (stochastic IMF sampling), while the faint end ($K \gtrsim 18$) is truncated by survey sensitivity limits. Within the survey's sensitive dynamic range ($11 \lesssim K \lesssim 17$), the observed pre-main-sequence members tightly constrain the cluster's age and morphology without requiring a populated high-mass turn-off.

4.2.6 Systematic Uncertainties: Binary Fraction and Worst-Case Quantitative Envelope

A significant systematic consideration in near-infrared isochrone fitting is the presence of unresolved binary stars. Physically, an unresolved equal-mass binary ($q = m_2/m_1 \approx 1$) possesses twice the luminosity of a single star, shifting its CMD position vertically brighter by up to $\Delta m = -2.5 \log_{10}(2) \approx -0.75$ mag along the sequence. As a worst-case quantitative scenario, if a cluster possesses a high unresolved binary fraction ($\sim$ 30–50%), an unconstrained single-star isochrone fit attempting to minimise Euclidean

distance to a binary-broadened sequence can suffer a systematic bias: it may underestimate the true distance modulus by up to ΔDM $\approx$ 0.2–0.3 mag ($\sim$ 10–15% in distance) or artificially lower the age to accommodate brighter turnoff stars. However, in my pipeline, this worst-case distance degeneracy is strongly mitigated because the distance modulus is independently anchored by the Bayestar19 3D dust map rather than floated freely against the broadened sequence.

4.2.7 Limitations at Extreme Extinction: Differential Reddening

A critical limitation of near-infrared isochrone fitting arises when characterising deeply embedded clusters subjected to extreme visual extinction ($A_V > 10$). In such environments, the obscuring dust is rarely uniform. Instead, the cluster population suffers from significant differential reddening, where individual stars are attenuated by varying column densities of foreground and intra-cluster material. In the colour-magnitude diagram, this effect smears the intrinsic stellar sequence into a broad diagonal band parallel to the reddening vector.

It is important to emphasise that for these extreme cases, an isochrone fit cannot definitively confirm the physical reality of the cluster. The initial candidate status of these populations relies on visual identification and photometric decontamination (the Bonatto method) to isolate probable members from the field. However, when applied to a differentially reddened population, a theoretical isochrone curve cannot capture the 2D spread of the data; it simply attempts to find the mathematical mean of the distribution. Consequently, while the derived age and distance provide a useful first-order approximation of the population's mean properties, the resulting χ^2_ν statistic is often inflated. For these highly obscured discoveries, the derived parameters should be treated as mean estimates. Ultimately, true physical confirmation of these embedded candidates—proving they are bound systems rather than line-of-sight alignments—often requires dedicated infrared spectroscopic follow-up, as optical kinematic diagnostics (like Gaia) typically fail to penetrate these regions (see Chapter 6).

The magnitude of this smearing can be quantified directly. Figure 4.1 shows clus288,

an embedded group fitted at $A_V = 14.0$ mag, plotted first as an observed CMD and then in a plane in which both axes are invariant under extinction. The dispersion of the population contracts by a factor of 4.6 between the two, establishing that the observed colour spread is dominated by differential reddening rather than by any intrinsic property of the stars.

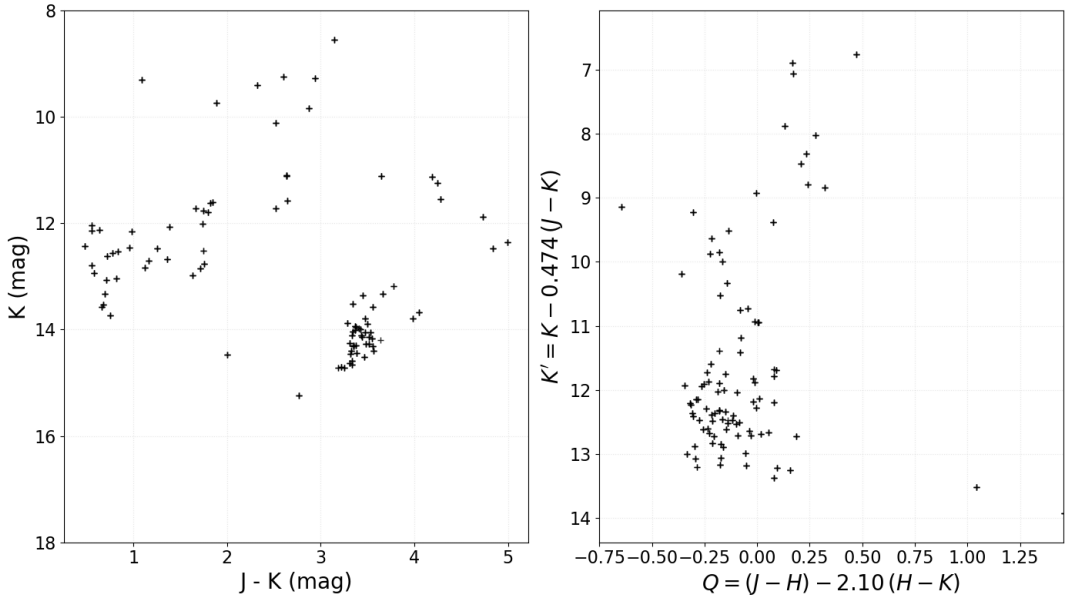


Figure 4.1 The effect of differential extinction on an embedded cluster, clus288 ($A_V = 14.0$ mag). *Left:* observed colour–magnitude diagram for the 286 decontaminated members, which span 4.5 mag in $J - K$. *Right:* the same members plotted in a reddening-free plane, in which both axes are invariant under extinction: $K' = K - 0.474(J-K)$ and $Q = (J-H) - 2.10(H-K)$, the coefficients following from the adopted extinction law. The dispersion contracts from $\sigma(J - K) = 1.21$ to $\sigma(Q) = 0.26$, a factor of 4.6. Since no amount of differential reddening can displace a star within the right-hand panel, the observed colour spread is dominated by extinction rather than by intrinsic stellar properties. It is this smearing that makes isochrone fitting substantially less secure for embedded systems than for the exposed clusters of Section 4.4.1.

While 1D isochrone fitting combined with Bonatto decontamination is highly effective, it has limitations in regions of extreme differential reddening. The power of combining mathematical fitting with rigorous visual scepticism is best demonstrated by comparing mathematical artefacts against true random scatter. In Figure 4.2, I contrast clus378 with clus170 (both now classified as Tier F) to highlight a specific failure mode: "False Positives via Differential Reddening."

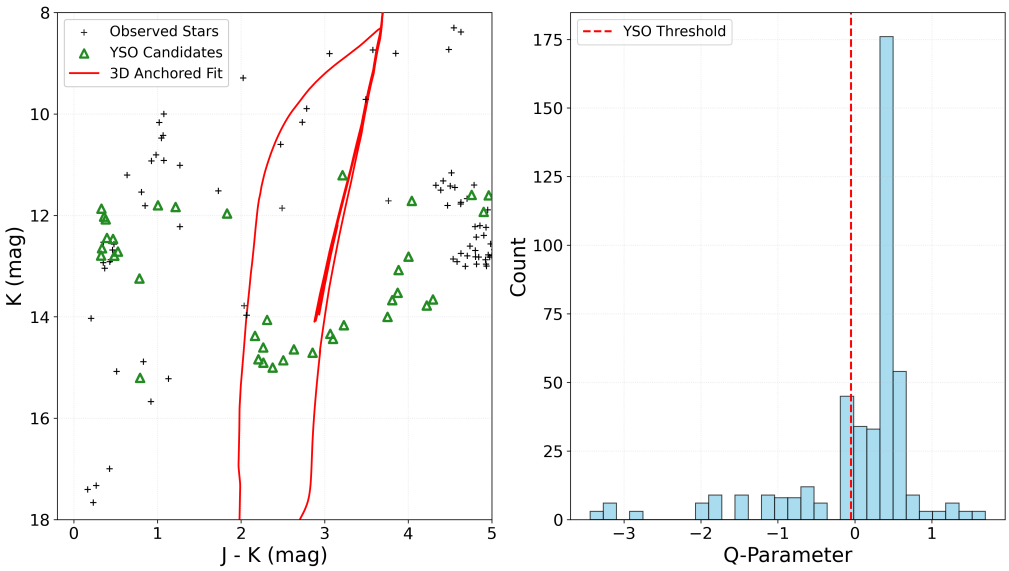
As seen in Figure 4.2(a), clus378 converges on an artificially low reduced χ^2 value

because the algorithm fits directly along the diagonal band of differential reddening ($A_V = 14.58$), even though visual inspection confirms that no coherent stellar sequence is present. A visual review of the sky images reveals that this is not a physical cluster core; instead, it is a dense background star field viewed through an irregular gap in heavy foreground dust lanes. In contrast, the "fit" for clus170 (Figure 4.2(b)) is a statistical failure, with stars scattered randomly across the colour-magnitude space and no coherent structure, resulting in a massive χ^2_ν. Identifying these differential reddening false positives (like clus378) necessitated a comprehensive visual review of all candidates to purge these mathematical artefacts from the final catalogue.

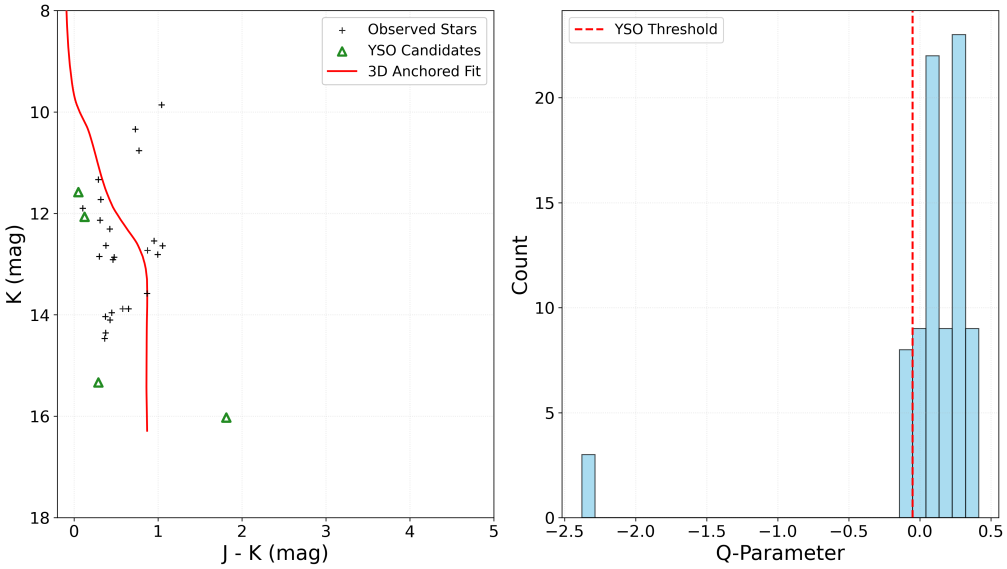
4.2.8 Limitations of Photometric Decontamination

The ultimate precision of the isochrone fitting pipeline is structurally bottlenecked by the inherent limitations of the Bonatto photometric decontamination method (described in Chapter 3). The Bonatto algorithm is fundamentally statistical; it operates by comparing the colour-magnitude density of the cluster core against a nearby control field. It does not definitively classify individual stars as physical members or non-members, but rather preserves stars in regions of the $(J - K, K)$ space that exhibit a mathematical overdensity relative to the background.

In the extremely dense environments of the Galactic Plane, the background and foreground field star distributions often perfectly overlap with the intrinsic cluster sequence, particularly along the highly populated red giant branch or the faint end of the main sequence. Because 2D photometric decontamination cannot perfectly separate members from field stars in these degenerate overlap regions, a residual population of non-members is inevitably retained in the final "cleaned" sample. When this decontaminated data is fed into the isochrone pipeline, the 2D Euclidean KD-Tree algorithm is forced to minimise the distance to all remaining points, including these residual field stars. This artificial noise unavoidably inflates the reduced chi-square (χ^2_ν) metric, establishing a structural limit on the mathematical precision of the fits. Consequently, while the photometric method successfully isolates the macroscopic



(a) Differential Reddening Artefact: clus378



(b) Random Scatter Artefact: clus170

Figure 4.2 Dual-panel comparison of fit quality for two Tier F field artefacts. Each panel shows the CMD fit (left) alongside the Q-parameter distribution (right). (a) clus378 yields an artificially low χ^2_ν by fitting along the diagonal band of differential reddening through dust lanes, despite lacking a physical core in the sky images. (b) clus170 is a random scatter failure showing no coherent population structure.

properties of the cluster, absolute member-by-member cleaning requires kinematic data (such as Gaia proper motions) or high-resolution spectroscopy.

4.3 Validation Against Known Clusters

To verify the robustness of the pipeline, it was applied to a control sample of six well-known star clusters. These clusters were selected to cover a broad range of evolutionary stages, from young star-forming regions (1–5 Myr) to more evolved open clusters (>300 Myr), and varying levels of interstellar extinction.

The results of the validation run are summarised in Table 4.1. The "Nature" classification is derived from the Q-parameter diagnostic and morphological definitions established in Section 1.5 (Table 1.1), where *Young* denotes embedded or pre-main-sequence systems ($\log t \lesssim 7.0$–8.0, or $\leq$ 10–100 Myr) exhibiting circumstellar dust or infrared excess, and *Evolved* denotes exposed open clusters ($\log t > 8.0/100$ Myr) where the natal cloud has dispersed. *Fit* $\log t$ is the derived age from the pipeline, compared against literature values. "Consistent" denotes an agreement within $\Delta \log t \leq 0.6$ dex, a threshold that accounts for systematic offsets between various isochrone families and multi-wavelength methodologies (Kharchenko et al. 2013). "Stage Confirmed" indicates a mathematical offset ($\Delta \log t > 0.6$ dex) due to extreme vertical degeneracies in the $(J - K, K)$ colour-magnitude space (Bell et al. 2007), but confirms the successful isolation of the fundamental evolutionary regime (e.g., embedded pre-main sequence vs. older exposed cluster).

Cluster	Lit. Name	Nature	YSO %	Fit $\log t$	Lit. $\log t$	Agreement
clus75	Be 55	Evolved	0.4%	9.09	7.93	Stage Confirmed
clus19	NGC 2194	Evolved	1.1%	9.10	8.52	Consistent
clus60	NGC 2254	Evolved	4.7%	8.77	8.31	Consistent
clus6	NGC 7129	Young	19.9%	6.58	6.48	Consistent
clus7	NGC 1893	Young	21.6%	6.04	6.18	Consistent
clus56	NGC 2175	Young	38.2%	5.64	6.95	Stage Confirmed

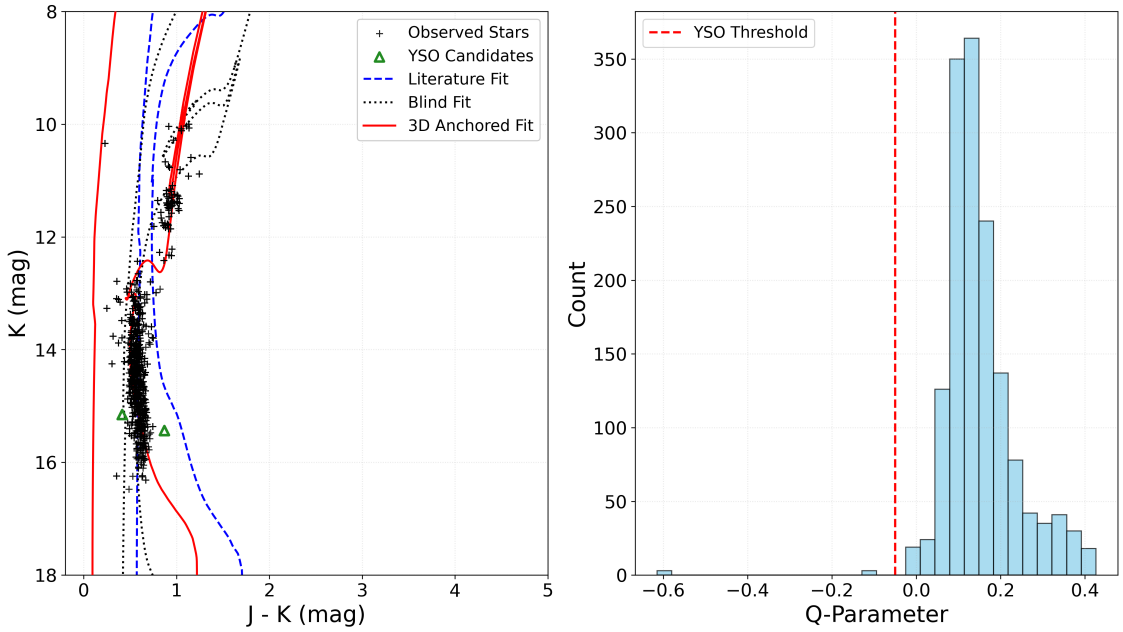
Table 4.1 Validation results for known clusters. Comparison of derived isochrone fit parameters against literature values for six control clusters.

Two of the six controls, Be 55 and NGC 7129, were additionally fitted blind, with distance left as a free parameter and no map constraint. These are presented first, since the comparison between the blind and the anchored fit establishes what the 3D anchor contributes; the remaining evolved and young controls follow. Where no blind fit was computed, only the anchored fit and the literature values are plotted.

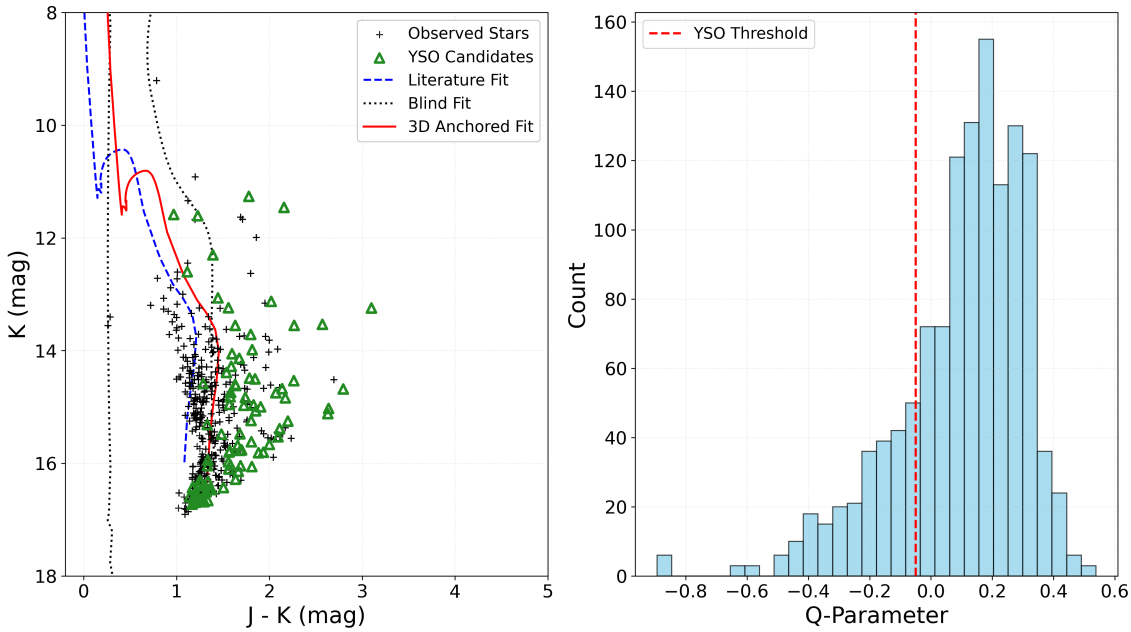
The same plotting conventions apply to Figures 4.3 to 4.5. In each left-hand panel the decontaminated members are drawn as black crosses whose arms span the typical UGPS photometric uncertainty (~ 0.08 mag), the 3D anchored fit as a red solid line and the literature isochrone as a blue dashed line; where a blind fit was computed it appears as a black dotted line. Each right-hand panel gives the distribution of the reddening-free Q index for the same members, with candidates satisfying $Q < -0.05$ marked as green open triangles.

Figure 4.3 shows the two controls for which a blind fit was also computed, one evolved and one young. Together they establish what the Bayestar19 anchor contributes. For Be 55 (clus75) the blind and the anchored fit converge on the same evolved solution at $\log t \approx 9.1$, and both recover the near-total absence of infrared excess, confirming a post-main-sequence stage consistent with the literature classification (Kharchenko et al. 2013); where the sequence is well populated and the turn-off clearly defined, the anchor is not needed. For NGC 7129 (clus6) the same comparison fails in the opposite direction: the unconstrained blind fit converges on an ancient age through the near-infrared age–distance degeneracy, whereas the anchored fit recovers the literature age (Kharchenko et al. 2013), and the Q-distribution independently confirms the young nature of the population. That degeneracy is most acute for pre-main-sequence populations, which is precisely where the anchor earns its place.

The remaining two evolved controls are shown in Figure 4.4. For NGC 2194 (clus19) the anchored fit follows the literature sequence closely and the Q-distribution is empty of YSO markers, which is what allows the pipeline to exclude young solutions outright. NGC 2254 (clus60) tests the stability of the 2D Euclidean metric against a well-populated red giant branch, which the fit reproduces in agreement with



(a) Be 55 (clus75), evolved



(b) NGC 7129 (clus6), young

Figure 4.3 Validation fits for the two control clusters for which an unconstrained blind fit was also computed, each shown as the colour–magnitude diagram with the fitted isochrone (left) and the Q-index distribution of the same members (right). For Be 55 the blind and anchored fits agree; for NGC 7129 the blind fit converges on an ancient age and only the anchored fit recovers the literature value. Symbols and line styles are as described in the text; fitted and literature ages are listed in Table 4.1.

Cantat-Gaudin and Anders (2020) and Kharchenko et al. (2013).

The remaining two young controls are shown in Figure 4.5. For NGC 1893 (clus7) the pipeline returns $\log t = 6.04$ against a literature value of 6.18, but the two isochrones are offset vertically by ~ 1 mag in K because the optical literature fit (Kharchenko et al. 2013) assumes an unconstrained distance modulus (DM ≈ 9.16), whereas the anchored fit adopts the Bayestar19 dust map constraint (DM = 11.6, Green et al. 2019), tying the theoretical sequence to the true infrared cluster distance. NGC 2175 (clus56) is the most heavily reddened of the six and has only a sparse high-mass sequence, yet anchoring the distance modulus at DM = 11.45 breaks the degeneracy and allows the age to be constrained from the pre-main-sequence morphology of the intermediate- and lower-mass members; nearly 40% of its members carry YSO signatures.

In all cases, the integration of the Bayestar19 constraint either correctly steered the fit away from degenerate local minima (as in clus6 and clus7) or provided consistent physical support for already stable solutions. The dual-panel view further reinforces the results by showing that the statistical fit is consistently supported by the independent, reddening-free population diagnostic provided by the Q-parameter.

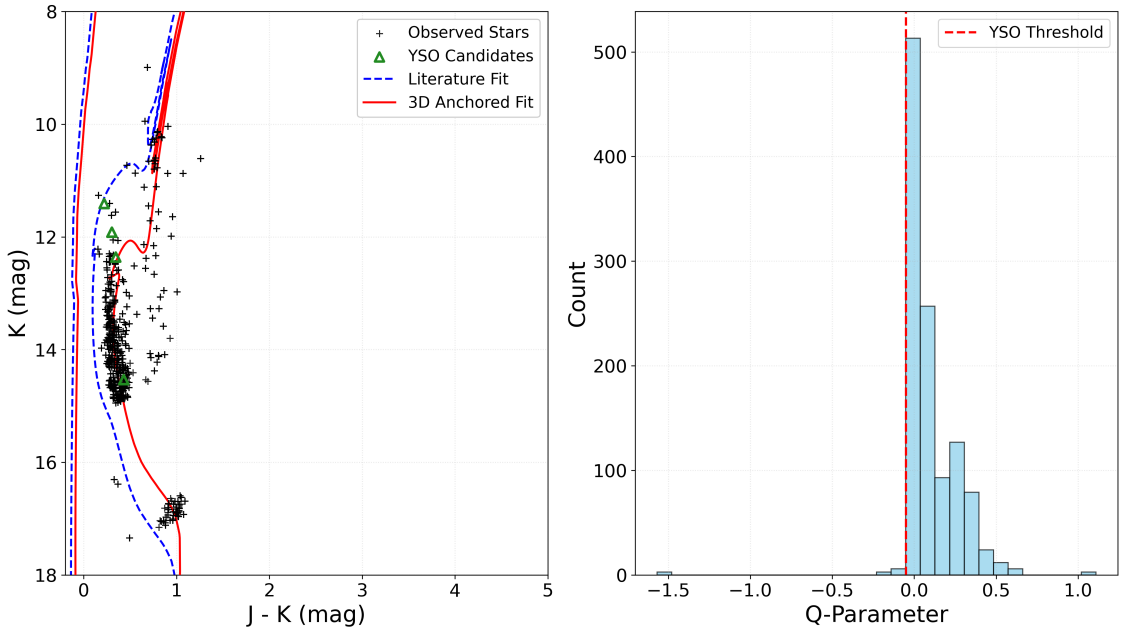
4.4 Results: Parameters of New Discovery Candidates

Following validation, the pipeline was applied to the 198 new discovery candidates. The results reveal a diverse population of star clusters throughout the Galactic disk, ranging from 100,000-year-old emerging clusters to ancient, highly obscured open clusters.

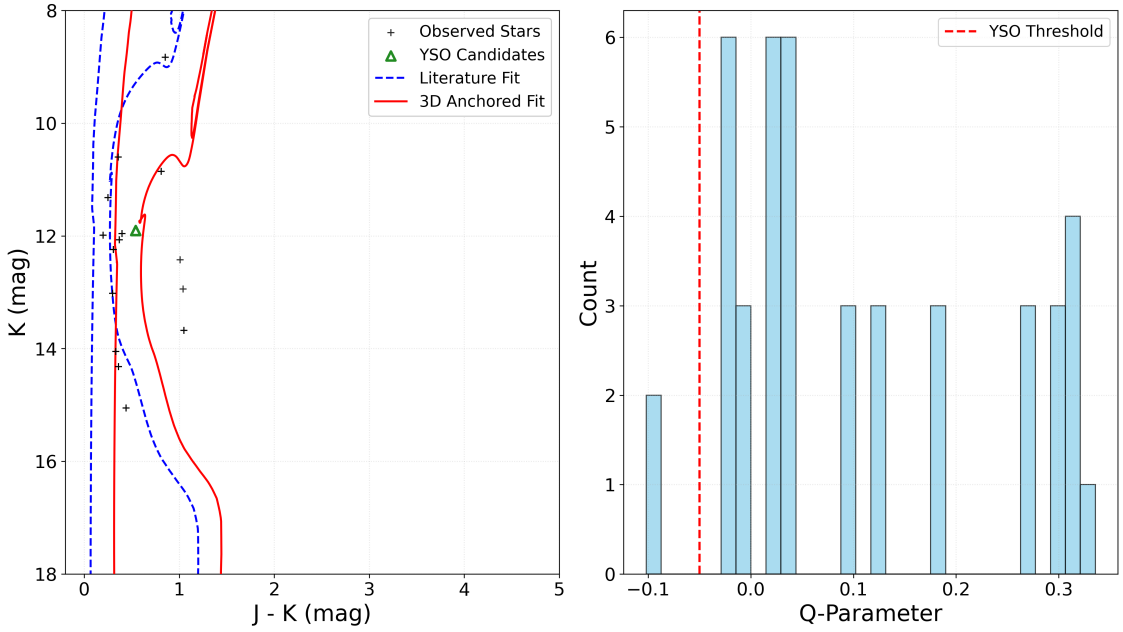
4.4.1 Quality Control and Validity Grading

To distinguish robust physical clusters from random line-of-sight field alignments, I have implemented a multi-criteria grading system based on population richness (N_{stars}) and the reduced chi-square (χ^2_ν) of the 2D Euclidean fit (Kharchenko et al. 2013). Candidates were categorised into three tiers:

- Tier A/B (Candidate): $N_{stars} \geq 10$ and $\chi^2_\nu \leq 5000$. These represent the

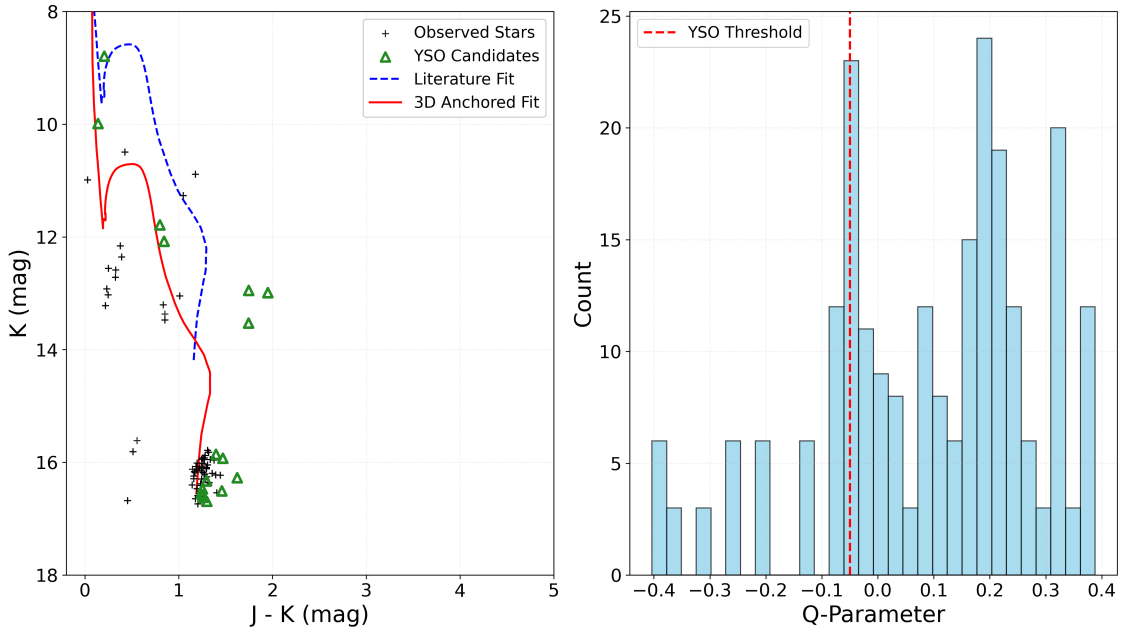


(a) NGC 2194 (clus19)

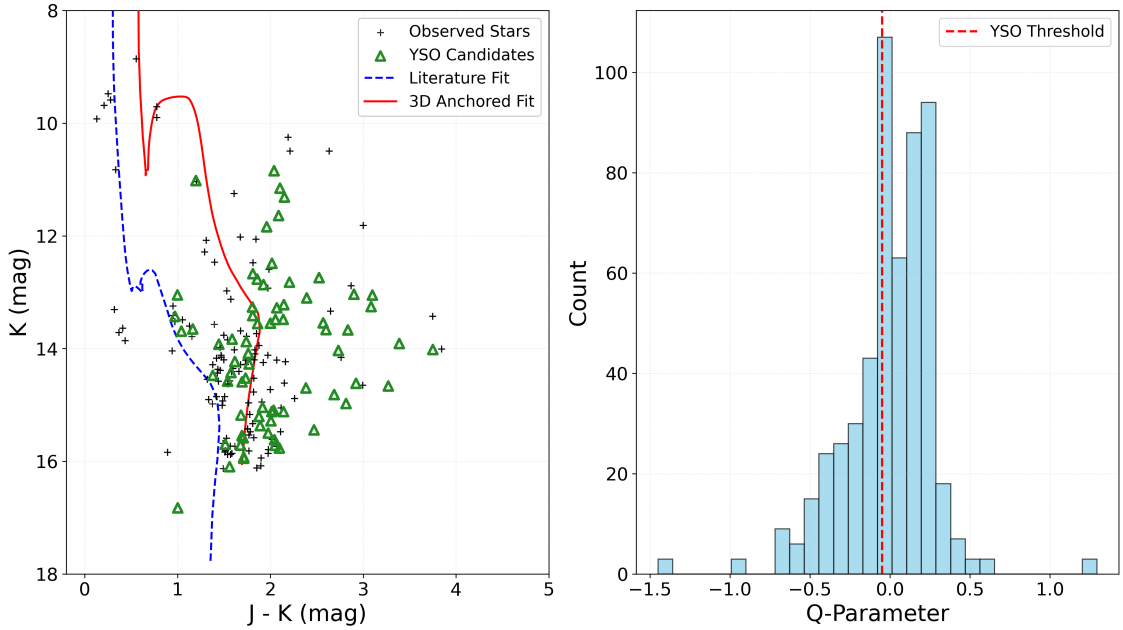


(b) NGC 2254 (clus60)

Figure 4.4 Validation fits for the two remaining evolved control clusters, plotted as in Figure 4.3. Neither was fitted blind, so only the anchored fit and the literature isochrone are shown. Fitted and literature ages are listed in Table 4.1.



(a) NGC 1893 (clus7)



(b) NGC 2175 (clus56)

Figure 4.5 Validation fits for the two remaining young control clusters, plotted as in Figure 4.3. Neither was fitted blind, so only the anchored fit and the literature isochrone are shown. Fitted and literature ages are listed in Table 4.1.

highest-confidence discoveries with statistically significant sequences.

- Tier C (Marginal): $N_{stars} < 10$, or a sequence exhibiting marginal scatter on visual review. These candidates either lack sufficient membership to be statistically distinguished from field noise, or retain a sequence too poorly defined for the fit to be considered secure.

- Tier F (Artefact): $\chi^2_\nu > 5000$. These cases represent catastrophic fit failures where the stars fail to form a coherent stellar sequence.

Out of 198 evaluated candidates, 178 (89.9%) satisfied the Tier A/B criteria and passed the final visual review. The remaining 20 were flagged as artefacts and excluded from the final discovery count. The high-confidence sample shows a median visual extinction of $A_V = 5.11$ and a median age of $\log t = 7.82$ (~ 66 Myr).

Within this high-confidence sample each candidate additionally carries a validity grade recorded in Appendix A, reflecting how well the membership constrains the fitted sequence: 170 are graded A, six B and two C. Figures 4.6 to 4.9 illustrate the progression, using the same pipeline and the same plotting conventions throughout so that the diagrams may be compared directly. The corresponding diagrams for all 178 high-confidence discoveries, generated identically, are provided in the data archive accompanying this thesis, so that the examples shown here may be checked against the full population.

4.4.2 Standout Discovery Populations

My analysis revealed several standout populations within the 178 high-confidence discoveries:

- Dusty: I identified a significant population of extremely obscured clusters, with visual extinctions exceeding $A_V > 10$. Examples include clus352 ($A_V = 14.99$) and clus360 ($A_V = 15.76$). When compared to the Global Survey of Star Clusters in the Milky Way (Kharchenko et al. 2013), where over 99% of statistically

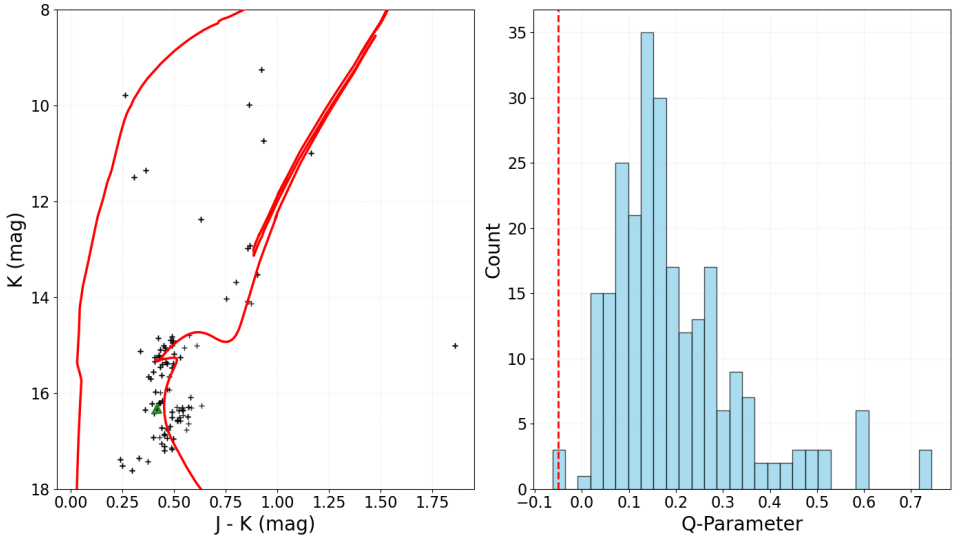


Figure 4.6 Isochrone fit for clus307, quality grade A. *Left:* colour–magnitude diagram of the 250 decontaminated members (black crosses) with the best-fitting MIST isochrone overplotted in red, for $\log t = 9.16$, $A_V = 1.83$ mag and $DM = 14.52$ ($d = 8.0$ kpc), reddened using the Wang & Chen (2019) law. The main sequence, the turn-off at $K \approx 15$ and the red giant branch are each traced by the model. *Right:* distribution of the reddening-free Q index for the same members, with the dashed line marking the $Q < -0.05$ threshold. Three of 250 members fall below it, consistent with an evolved population from which circumstellar material has dispersed.

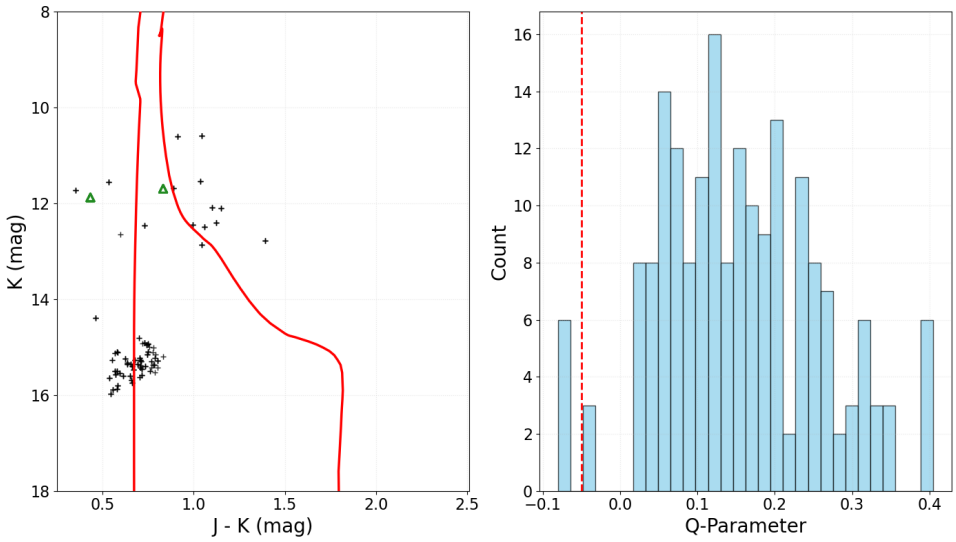


Figure 4.7 Isochrone fit for clus321, quality grade A, for $\log t = 7.77$ ($\sim$ 59 Myr), $A_V = 6.15$ mag and $DM = 10.10$ ($d = 1.05$ kpc). Symbols as in Figure 4.6. *Left:* the 189 members form a compact concentration at $K \approx 14.5$–16 which the model main sequence traverses; no main-sequence turn-off is present, as expected at this age. This candidate was recorded as 'No Cluster' during the initial visual screening and recovered only by the decontamination and fitting described in this chapter, illustrating the population of such recoveries discussed in Chapter 8. *Right:* five of 189 members show infrared excess, a low fraction consistent with an age well beyond the disc-bearing phase; the young classification here rests on the absence of a turn-off rather than on infrared excess.

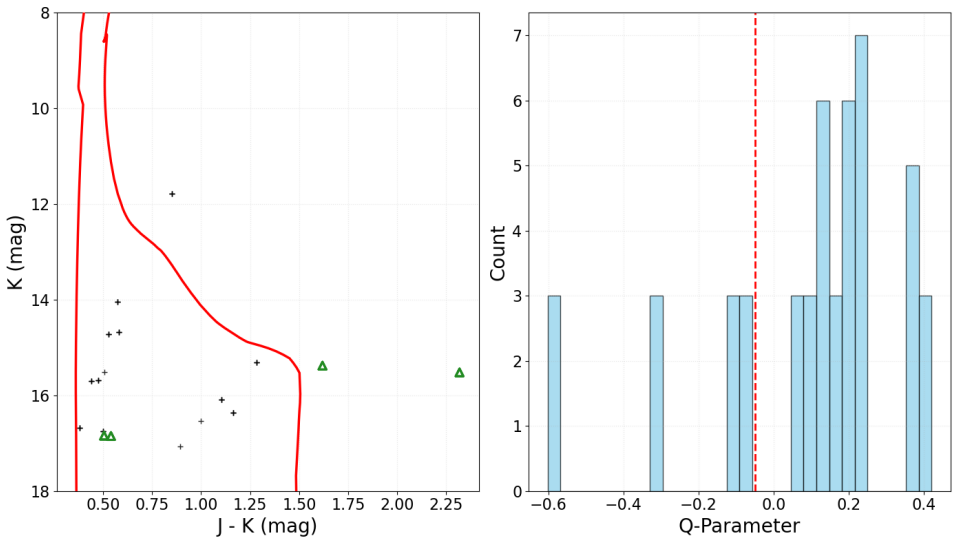


Figure 4.8 Isochrone fit for clus59, quality grade B, for $\log t = 7.79$ ($\sim$ 62 Myr), $A_V = 4.10$ mag and $DM = 10.31$ ($d = 1.15$ kpc). Symbols as in Figure 4.6. *Left:* the 48 members are consistent with the model but too few to constrain it. This is what distinguishes grade B from grade A — not a poorer fit, but less evidence supporting it. *Right:* 11 of 48 members lie below the $Q < -0.05$ threshold, a high infrared-excess fraction consistent with the emerging classification.

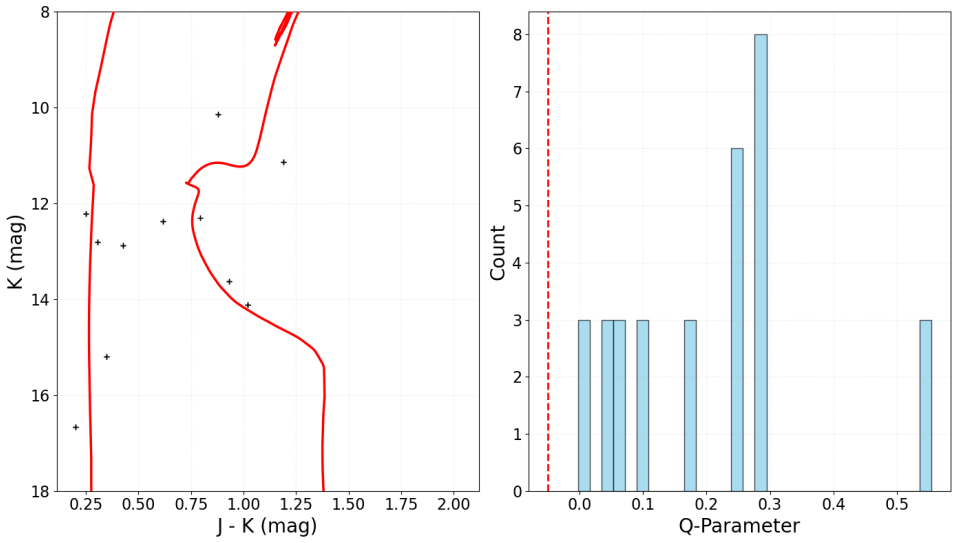


Figure 4.9 Isochrone fit for clus235, quality grade C, for $\log t = 9.40$, $A_V = 3.38$ mag and $DM = 10.05$ ($d = 1.02$ kpc). Symbols as in Figure 4.6. *Left:* with 32 members and no coherent locus, the fit is formally converged but the diagram provides no independent support for a cluster sequence. Grade C denotes exactly this case: the numerical membership threshold is met, but visual review finds the sequence too poorly defined for the fit to be considered secure. *Right:* no member falls below the $Q < -0.05$ threshold.

characterised clusters exhibit $A_V < 10$, these discoveries sit at the far end of the Galactic extinction distribution. These are likely among the most deeply embedded clusters currently characterised.

- Distant: Seven discoveries were fitted at distance moduli beyond $DM = 15.5$, placing them at the far edge of the Galactic disk. The most distant are clus242, clus324 and clus247, at $DM = 16.64$, 16.60 and 16.32 respectively.

- Rich: Several discoveries exhibit exceptionally high membership counts, making them primary targets for follow-up spectroscopic studies. clus373 ($N_{stars} = 1851$), clus272 ($N_{stars} = 1264$) and clus354 ($N_{stars} = 1152$) represent the most statistically robust sequences in the entire sample, with clear main-sequence and PMS populations.

- Young/Emerging: Several clusters show high YSO percentages consistent with very young ages. For example, clus31 and clus24 exhibit YSO fractions of 24.4% and 22.2% respectively.

In contrast, the Tier F artefact sample (e.g., clus170, clus151, clus378) serves as an important negative control. These candidates often satisfy the initial richness criteria ($N > 40$) but either fail to form any coherent sequence or are shown to be dust-lane artefacts upon visual inspection. By excluding these 20 cases, I have ensured that the final catalogue of 178 clusters represents a statistically sound and physically consistent set of real Galactic stellar populations.

4.5 Summary

In this chapter, I have presented a robust pipeline for the stellar population analysis of my cluster candidates. By combining MIST evolutionary models with the Q-parameter diagnostic and 3D line-of-sight constraints from the Bayestar19 dust maps, I have resolved the inherent NIR degeneracies that often hinder infrared survey work. The adoption of a 2D Euclidean fitting metric, supported by efficient KD-Tree geometry and

validated against a control sample of known clusters, has allowed for the derivation of precise physical parameters for 178 new star cluster discoveries. Categorised through a rigorous validity grading system based on global survey standards and manual visual review, these clusters form a high-confidence catalogue.

Chapter 5

Basic Characterisation

Following the decontamination process there were a number of characterisation steps applied to the candidates to gather basic properties, allow categorisation and provide environmental context. As well as this I made a number of searches against catalogues of known clusters, both visible and infrared, looking for existing references and other data such as distances, number of members, age and so forth. The data from existing well known clusters is a valuable tool to compare to the results obtained in this and following sections to identify any limitations, and this in turn will add confidence to the initial analysis of previously unknown candidates. The resulting data are tabulated in Appendix A and the data archive files accompanying this thesis.

The following steps were applied to each candidate post decontamination and each step is described in detail in the following subsections.

- Algorithmically determine the central coordinates
- Plot a radial density profile and determine the core and tidal radii where possible
- Estimate the angular diameter
- Record number of decontaminated members
- Identify star formation markers within 10 arcmin
- Investigate warm gas and dust using WISE All-WISE
- Record associated J, H or K-band nebulosity
- Perform literature & catalogue search

This data provides the basis for a systematic characterisation of the candidates and

identification as either a previously known cluster, or candidate cluster, a new candidate cluster or rejected. The section following this describes the way the data was used to make these determinations and how it compares to known data, where available.

5.1 Algorithmically Determined Central Coordinates

Whilst each of the candidates initially has a visually determined coordinate to enable data retrieval, it is important to determine accurate sky coordinates for the decontaminated candidates and allow for iteration where required. Some of the issues around determining a robust central coordinate have been discussed in previous sections but will again be summarised here. The basic issue is that the initial visual centre is subjective and based on a view that includes field stars as well as the clusters actual member stars. This can skew the perceived centre of density and lead to inaccuracy, particularly in the crowded fields towards the Galactic centre. The degree of contamination is also dependent on the optical depth of any accompanying dust and gas as this can obscure stars located behind the cluster for instance, although it can also redden field stars such that appear superficially to be cluster members. We need to have a method that can produce an algorithmically derived centre based on the actual cluster member stars (in so far as we can determine them with the data and tools available).

As discussed previously this requires a degree of iteration in order to settle on a robust centre. The initial visual centre is used to extract data in a 5 arcmin radius[1] before the algorithm is run to determine the first centre for decontamination. Following decontamination the algorithm is run again and a delta computed, if this is greater than 25% of the estimated apparent radius (see §5.2) then the decontamination is run again using the updated centre, and this iterates as needed.

The candidates exhibit several different morphologies and each of these presents a different algorithmic challenge so I investigated two different approaches in order to

[1]This is the upper limit for a GPS *crossId* (cone) search, but it is large enough to encompass all but the very largest candidates in this sample. As stated previously, I am primarily interested in the smaller candidates, with the larger ones all previously known.

determine whether one performed better across all morphologies, or a combination of both was required. This is discussed in the sections below. The source data is the WSA-2MASS cross-match that was extracted using basic quality cuts as described in section §2.2.1, with a 5 arcmin radius.

5.1.1 Kernel Density Estimation (KDE)

Open clusters often have a true central concentration and therefore a density peak method works well to identify the centre. Here I used *kernel density estimation* (KDE) as it is well suited to handling the highly variable distributions found in star fields, and has been used many times in the literature for similar problems (see Borissova, Bonatto, et al. 2011; Camargo, Bica, and Bonatto 2015; Froebrich, Scholz, and Raftery 2007 for examples). For 2-dimensional data such as sky coordinates (x_i, y_i), the KDE estimate at any point (x, y) is given by:

$$f(x, y) = \frac{1}{n h_x h_y} \sum_{i-1}^{n} K\left(\frac{x - x_i}{h_x}, \frac{y - y_i}{h_y}\right) \tag{5.1.1}$$

where n is the number of stars, K is the kernel function (usually a Gaussian) and h_x, h_y are the bandwidth (smoothing parameters). The kernel describes the shape of the function at each star and the bandwidth controls the maximum (size) of the function - in other words, too small a bandwidth and the result is a noisy, over-fitted, density map and too large a bandwidth can smooth over the actual structure of the distribution. The centre in this case is the point at which (h_x, h_y) is at its maximum. It is useful to iterate over different kernel functions and smoothing bandwidths to find that best suited to the data set.

Figure 5.1 shows a density plot for OC cluster candidate 88. There is a visibly strong central concentration of stars and the algorithm has picked a centre (shown by the red cross) only 5.9 arcsec from the visually selected one (shown by the blue cross). Whilst this works well for those candidates with a well defined central over-density and little variation in extinction across the field, for the majority of candidates for which this is not the case it can produce highly variable results. Indeed, many of the candidates in the

sample do not exhibit the strong central density that this method requires, the average delta from the visual coordinates is 73 arcsec, with a standard deviation of 45 arcsec, which is significant variation.

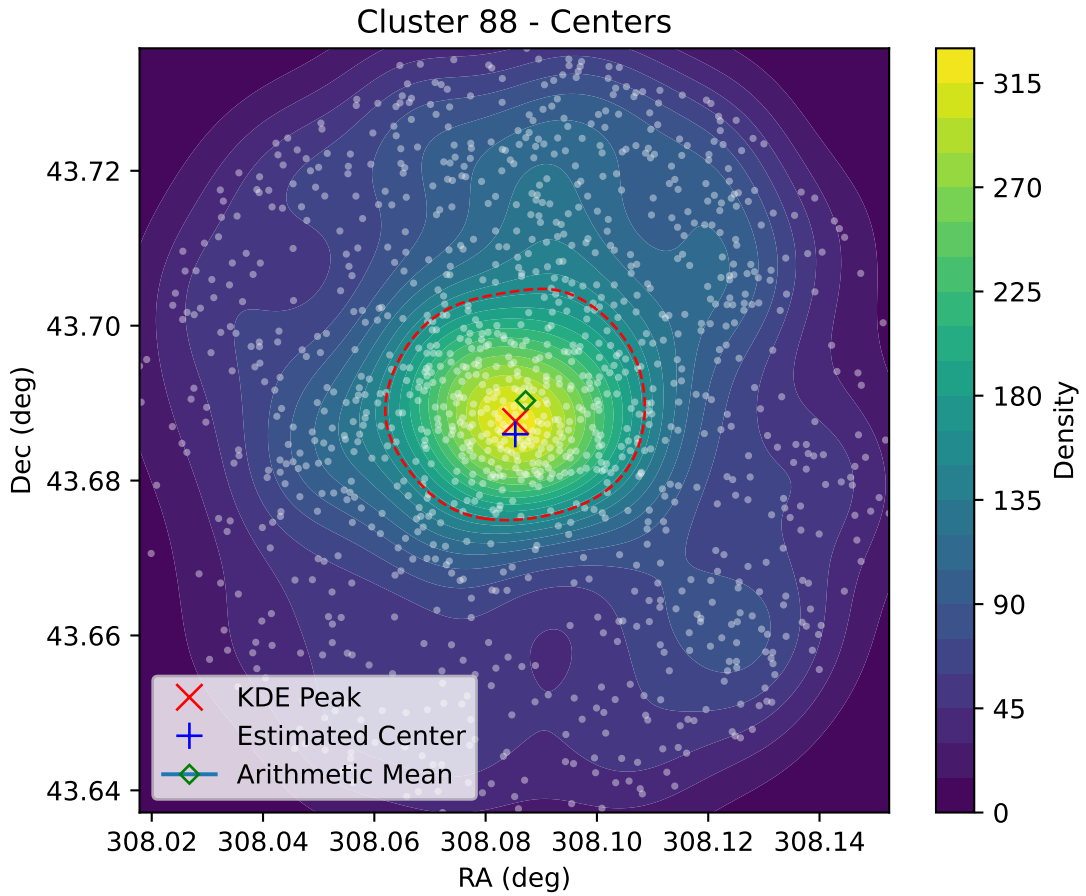


Figure 5.1 A KDE Density plot for cluster candidate 88. The image covers a 3 arcmin radius centred on the visual cluster centre (marked with a blue cross). Stars are shown as white points, there are 30 colour coded density contours and the density centre is marked with a red cross. The red dotted line marks the contour of the half the maximum density value and is included in this form to help gauge the form of the density gradient. The green diamond marks the centre as calculated using the arithmetic mean. For open clusters such as this, all methods return a good central estimate.

5.1.2 Arithmetic Mean Centre (AMC)

Whilst K-band can reveal embedded stars in young, emerging clusters, many can still be invisible as a result of the very high extinction of the associated dense clumps. These variations in extinction make a stellar density based approach, which relies on visible star

density, highly biased and therefore inaccurate and unreliable at best. Gutermuth, Megeath, et al. (2009), for example, used the arithmetic mean of the coordinates to work around this problem as it is extremely robust to patchy data and will still return a good centre even if $\sim 20-30\%$ of the core is invisible. The arithmetic mean of a series of coordinates of form $(x_i, y_i)...(x_n, y_n)$ is given by:

$$x_{mean}, y_{mean} = \left(\frac{1}{n} \sum_{i=1}^{n} x_i, \frac{1}{n} \sum_{i=1}^{n} y_i \right) \tag{5.1.2}$$

where n is the number of stars and (x_i, y_i) are the sky coordinates for each star in the target region. A good comparison between algorithms can be seen in Figures 5.2 and 5.3 which show candidate 380, a known star forming region, DR 21, in Cygnus. In this instance there is a dark, sparsely populated, central region surrounded by star forming regions. As can be seen in the figures, the KDE centre is located to the edge of the dark region where the most visible stars are located, and this gives a misleading coordinate for the whole region. The arithmetic mean centre however presents a much more convincing result, in this case being more closely aligned with the centre of the visible stars. The dense clump visible in the far-infrared WISE image (Figure 5.3, panel *a*). The UKIDSS GPS K-band image in panel *b* is shown for clear comparison of the field and the lack of visible central stars which skews the density based centre method. Across all the candidates the average difference between the visual coordinates and the AMC is just 10 arcsec, with a standard deviation of 15 arcsec - significantly smaller than for the KDE method. This clearly supports the suitability of the AMC for use with this sample and is the algorithm I have used for calculating the central coordinates throughout this work.

5.2 Radial Density Profiles (RDP) and Half-Maximum Radius (HMR)

The radial density profile of a star cluster is valuable for understanding its structure and evolution. The profile can confirm the presence of a cluster core and a allow us to determine a tidal radius (the outermost point where the cluster's gravitational influence is

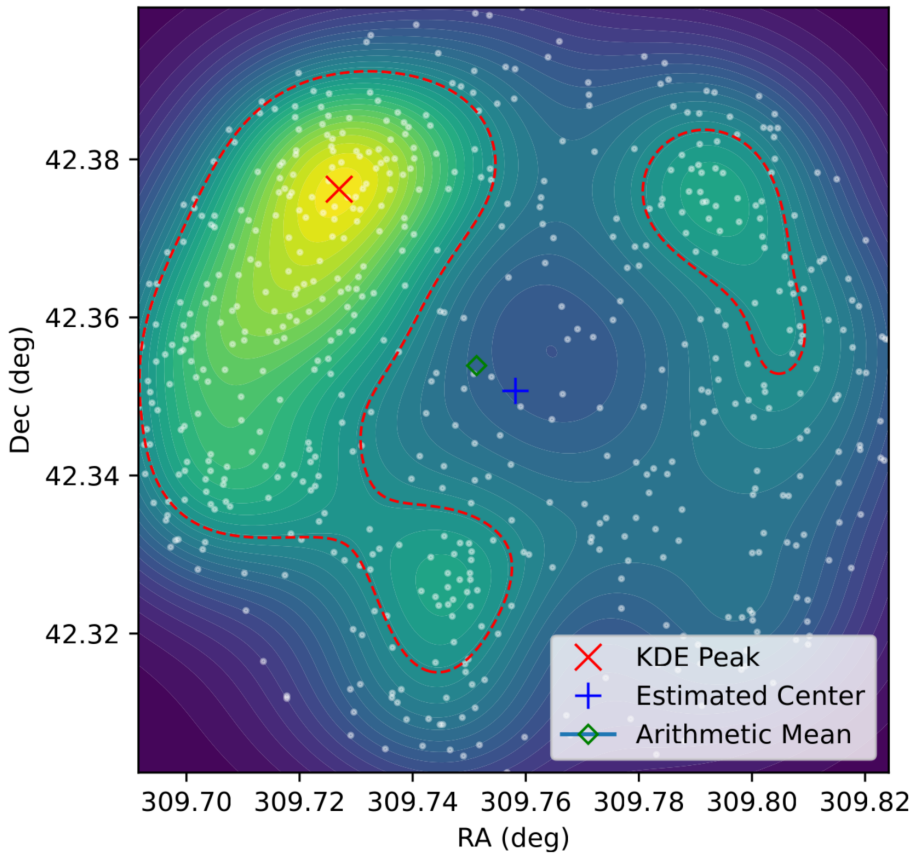
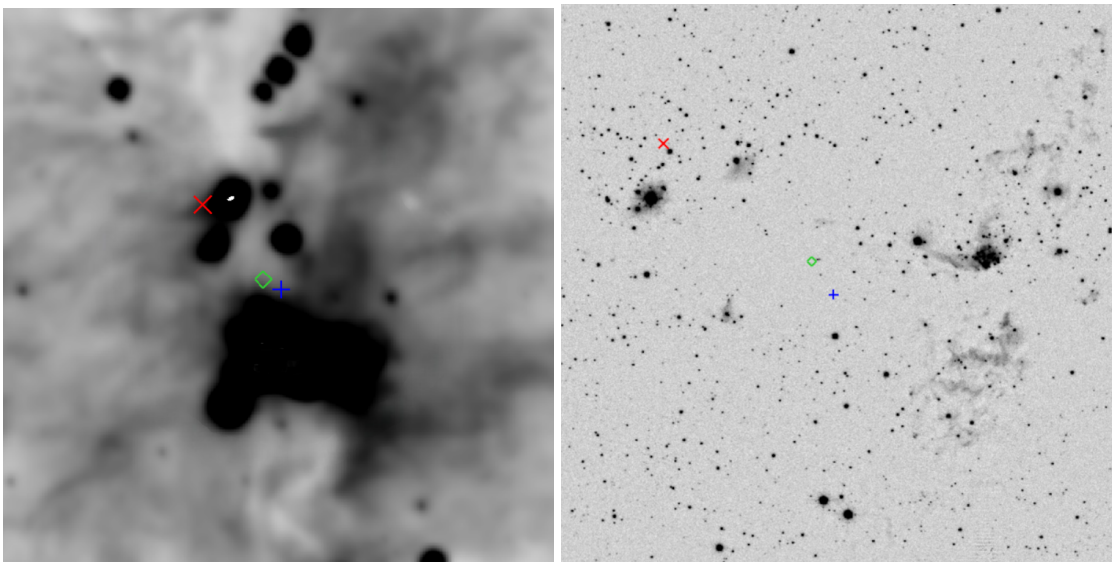


Figure 5.2 A KDE density plot and arithmetic mean centre for candidate 380, a known star forming region, DR 21, in Cygnus. The image covers a 3 arcmin radius centred on the original visually estimated cluster centre (marked with a blue cross). The white points are the star locations and there are 30 colour coded density contours and the density centre is marked with a red x. The red dotted line marks the contour of the half-maximum density value and is included as a notional boundary to help gauge the density gradient (see also §5.2). The green diamond marks the centre as calculated using the arithmetic mean. As can clearly be seen the dark central region is sparsely populated and the KDE identified a sub-group as its centre. The AMC however has still returned a believable centre. Figure 5.3 shows K-band and 12μm images of the same region showing that the cluster centre is indeed more in line with the AMC than the KDE.

dominant), as well as help distinguish between the types of cluster, giving us information regarding the state of relaxation, and hence age, of the cluster (although that is beyond the scope of this work).

There are three principle models that can be applied to this, the King, Plummer and Elson-Fall-Freeman (EFF), and these are described briefly below.



(a) WISE 12μm image of candidate 380.

(b) UKIDSS GPS K-band image for candidate 380.

Figure 5.3 Images of candidate 380, a known star forming region, DR 21, in Cygnus. In both panels the blue cross marks the visual centre of the cluster, the red 'x' marks the KDE centre and the green diamond marks the AMC centre. Panel *a* shows a WISE 12μm image which clearly reveals the dense clump of material from which the cluster is forming. Whilst the visual and AMC centres may not be the eventual true centre, they are clearly more aligned with the distribution of material than the KDE centre. Panel *b* shows the UKIDSS GPS K-band image for the same region. The sparse central area of the image biases the KDE algorithm; it sees two clumps of stars and incorrectly assigns the centre.

5.2.1 King Profile

The King profile is a classic empirical model used to describe the radial surface density of stars in a gravitationally bound star cluster, originally developed to model relaxed globular clusters and is given by:

$$\sum(r) = \Sigma_O \left(\frac{1}{\sqrt{1 + (r/r_c)^2}} - \frac{1}{\sqrt{1 + (r_t/r_c)^2}} \right)^2 \tag{5.2.3}$$

where $\sum(r)$ is the surface density at a projected radius r, Σ_O is the central surface density, r_c is the core radius, representing the radius at which the density drops to approximately half its central value, and r_t is the tidal radius, representing the approximate boundary of the cluster.

There are some limitations for this model and its usefulness in this situation, firstly it assumes a degree of relaxation of the cluster (i.e. mass segregation) and may not be

suitable for younger, irregular, clusters and it requires manual fitting of the parameters to ensure a good fit (see King 1962 and McLaughlin and Marel 2006, for example). While this would be acceptable for a small sample size it would be preferable to use an algorithm more suited to automation for the processing of several hundred candidates.

5.2.2 Plummer Profile

The Plummer profile offers an alternative, which is often used for younger or less dynamically evolved clusters and is a simpler model than the King profile. It does not have a sharp tidal cut-off, which is more appropriate for emerging clusters that haven't fully established their boundaries, however it might not accurately represent clusters with a well-defined edge. The Plummer profile is given by:

$$\sum(r) = \Sigma_O \left(1 + \left(\frac{r}{b}^2\right)\right)^{-2} \tag{5.2.4}$$

where $\sum(r)$ is the surface density at a projected radius r, Σ_O is the central surface density and b is the Plummer radius, which is related to the core size. The softer outer region might not accurately represent clusters with a well defined edge. Fitting a Plummer profile alongside the King profile can be informative. If the Plummer profile provides a comparable fit, it might suggest that the cluster is not yet significantly affected by tidal forces or that its outer regions are more diffuse. Differences in the best-fit parameters between the two profiles can offer insights into the cluster's dynamical state. See Plummer, Plummer, and C. (1911) as the original paper and Dejonghe (1987) for an example application.

5.2.3 Elson-Fall-Freeman (EFF) profile

Finally, there is the Elson-Fall-Freeman (EFF) profile. This profile is often found to provide better fits to young and massive clusters. It is a power-law profile with a core radius and does not have the tidal cut-off of the King profile which again makes it more appropriate for systems that haven't yet established clear tidal boundaries. Importantly for young, emerging clusters it makes no assumption about the state of relaxation. The

2D EFF profile is given by:

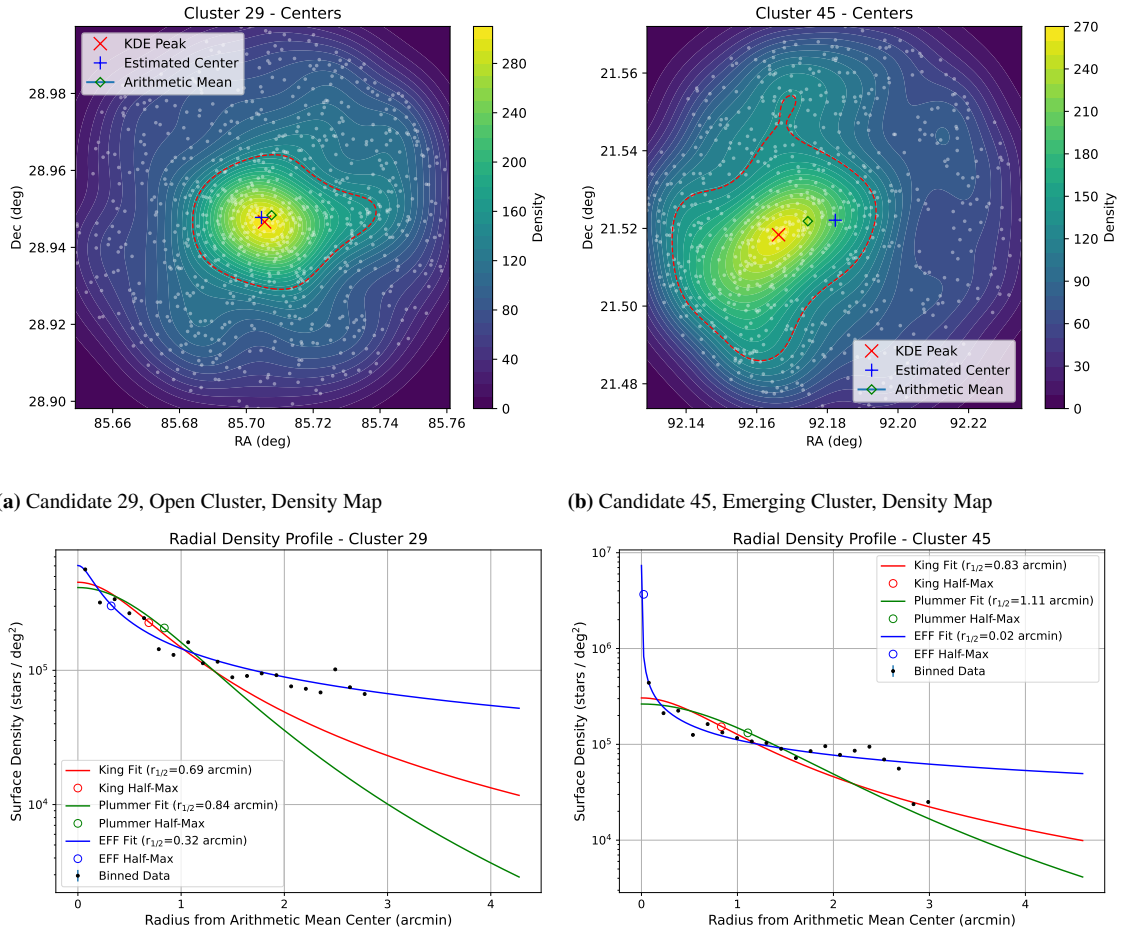
$$\sum(r) = \Sigma_O \left(1 + \left(\frac{r}{r_c}\right)^2\right)^{-\frac{\gamma}{2}} \tag{5.2.5}$$

where $\sum(r)$ is the surface density at a projected radius r, Σ_O is the central surface density, r_c is the core radius and γ is the power law exponent for the outer regions. In this case the power-law exponent, γ, allows for a wider range of outer density declines than the King profile and there is no assumption of tidal truncation. This can make the EFF profile more suitable for young clusters that may not be tidally limited, however if the cluster is starting to feel tidal effects, this will not be captured. See Elson et al. (1987) and Borissova, Alegría, et al. (2016) for example. The EFF profile sometimes struggled to assign a useful core-radius vs the other profiles but I found that a density-based, average half-maximum radius worked well for those candidates where the EFF estimate was poor, so this has also been included. The EFF profile does not give a value for the tidal radius, so in order to make a useful estimate of the angular diameter of a candidate, I have also measured and recorded the average maximum radius for the decontaminated cluster members, using the post-decontamination AMC coordinates as the centre.

Figure 5.4 shows examples of radial density plots from two candidates, one open and one emerging. Panels *a* and *c* show candidate 29, known open cluster [FSR2007] 0826. In this case the curves of the three models are similar to a radius of about 1 arcmin, although the EFF models slope is inverted. After this point they diverge significantly and the EFF model most closely fits the binned density data out to 3 arcmin. The HMRs of the King and Plummer functions are similar at about 0.75 arcmin, but the EFF is significantly less at about 0.25 arcmin - all indicating a tightly concentrated central over-density as we see in the density plot. Panels *b* and *d* show candidate 45, known emerging cluster [FSR2007] 0889. Here we see a different picture, the EFF model shows a very steep initial drop in density, with the HMR just an arcsec from the centre, but then closely follows the density values as radial distance increases. The sharp initial drop is due to the AMC being just at the edge of a dense region in this emerging, amorphous cluster, and indeed there is a sharp drop in average density. The King and Plummer profiles are much shallower and slightly overestimate the density in the region just

outside the centre before tailing off to underestimate in all but the final two bins. The EFF curve again provides the best overall fit to the data, albeit more sensitive to an amorphous distribution.

In both cases it is clear that the EFF profile provides the best fit to the data, albeit more sensitive to an amorphous distribution of stars in emerging candidates. Across many examples the EFF provided similarly good fits and so is the profile I have used to produce the RDPs for the candidates.



(a) Candidate 29, Open Cluster, Density Map

(b) Candidate 45, Emerging Cluster, Density Map

(c) Candidate 29, Open Cluster, Radial Density Profile

(d) Candidate 45, Emerging Cluster, Radial Density Profile

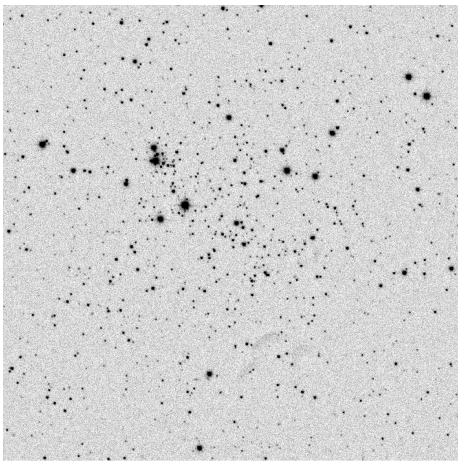
Figure 5.4 Surface and radial density plots showing the output of three different models; King in red, Plummer in green and EFF in blue. The radius of the functions half-maximum value is marked as an approximation of the clusters radial size.

5.3 WISE Far-Infrared Images

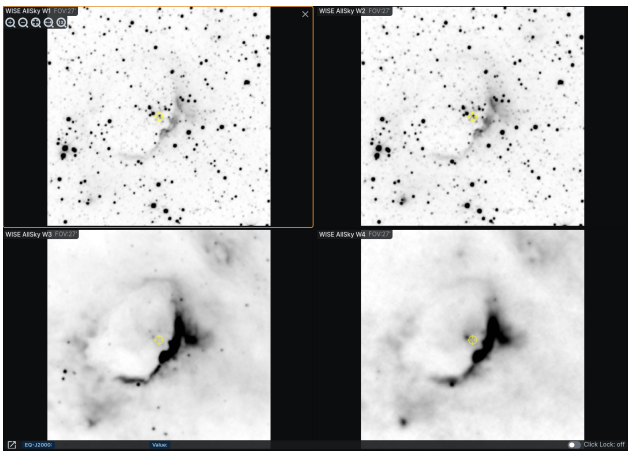
As part of the initial search I had noted whether candidates have any apparent K-band nebulosity as a rough guide to separating the candidates by age, young clusters would typically still have more of the progenitor material in close proximity as the brighter stars have not yet completed the process of photoevaporation (see Chapter 3 for more details). This material is typically warm and radiates at the mid-infrared wavelengths detected by WISE. In order to further investigate this and gain a deeper understanding of the context for each of the candidates I made a cross-reference with the IRSA WISE AllSky image catalogue as described previously in Chapter 2, §2.3.

The IRSA viewer provides easy access to all four of the wise bands, 3.4, 4.6, 12 and 22μm and these wavelengths, longer than those visible to the GPS, can provide a wide-field, but lower angular resolution, view of the candidates environment. This is particularly important for any younger clusters as the warm temperature region is where we see a lot of early star formation markers, HII regions, IR dark clouds, cores and YSOs. Seeing this kind of activity will provide supporting evidence not only for a new cluster but also to help begin characterisation in terms of its relative age (see Peretto and Fuller 2009; Anderson et al. 2013; Allen et al. 2004; Gutermuth, Myers, et al. 2008 among others).

Figure 5.5 shows a faint candidate in K-band with a hint of nebulosity in close proximity (panel *a)*. Using WISE images for a slightly larger region we can see in panel *b* that there is indeed a band of warm material associated with the over density and this becomes very apparent in the W3 and W4 bands (12 and 22μm respectively). The presence of warm material like this would suggest that we do indeed have a relatively young cluster in which the more massive, and luminous, members have begun the process of photoevaporation, forming and HII region.



(a) Candidate 9, GPS K-Band image, 5x5 arcmin



(b) WISE AllSky images for candidate 9, 10x10 arcmin tiles)

Figure 5.5 Candidate 9, panel *a* shows a 5x5 arcmin GPS K-Band image, - A known faint open cluster with hint of nebulosity. WISE follow-up will show the distribution of lower temperature material in the region, a clump of dense cooler material could indicate a site of star formation activity, in conjunction with star formation markers. Panel *b* shows WISE AllSky Atlas images, each 10 x 10 arcmin, showing all four of the WISE wave-bands, W1 - 3.4μm, W2 - 4.6μm, W3 - 12μm and W4 - 22μm. The hint of nebulosity seen in panel *a* is indeed a larger clump of warm material with the characteristic circular shape of an HII ionised region. This suggests the cluster has recently formed massive stars which have begun clearing the gas and dust through a combination of strong stellar winds and photoionisation.

5.4 Literature and Catalogue Search

For this work it was important to understand which of the objects identified during the visual search were previously known, misidentified, or unknown in order that I can focus on those unknown or candidate objects. As described in Chapter 2, SIMBAD and Visier are the CDS data services that provide curated databases of astronomical objects including references, catalogues and characteristics, and include all major catalogues.

For each candidate I performed a 10 arcmin cone-search, search similar to that described in looking for matching clusters and useful characteristics to crossmatch. Identifying whether a candidate is already known is not just a case of finding references to it in one of the databases but also following up the likely looking literature references to see if a cluster has actually been identified correctly, whether it has been confirmed or

perhaps misidentified. It is not uncommon to find that a smaller, dense cluster has been previously mis-identified as another type of object, a galaxy, galaxy cluster or YSO, for instance. The higher resolution and depth of the GPS survey allows us to see the true nature of many of these objects. Where this has occurred a note has been made in Appendix A.

The environment of each candidate, also within a 10 arcmin radius was investigated for any associated star formation markers (see Table 5.1). This is important for several reasons, these markers act as a diagnostic toolkit providing evidence of ongoing or recent star-forming activity and form part of the assessment of a candidates age and local environment, whether it is an isolated cluster or part of a larger star forming region. The search for previously known clusters and parameters included:

- SIMBAD catalogue (Wenger et al. 2000) (looking for object types):
  - Cl*, Cluster of Stars
  - Cl?, Cluster Candidate
  - OpC, Open Cluster
  - GlC, Globular Cluster
  - Gl?, Globular Candidate
- Hunt and Reffert (2023) (HNR) catalogue
- Milky Way Star Clusters (MWSC) catalogue (Kharchenko et al. 2013)
- DAML02 Open Cluster catalogue (Dias et al. 2002)
- WISE (wisecat) Catalogue of Galactic HII Regions (Anderson et al. 2013)
- FSR (2007) Catalogue (Froebrich, Scholz, and Raftery 2007)
- Unified Cluster Catalogue (UCC) (Perren et al. 2023) [2]

[2]UCC is here replacing WEBDA whilst it is offline, by suggestion of maintainer, Ernst Paunzen, Department of Theoretical Physics and Astrophysics of the Masaryk University, Brno (Czechia)

Object Type	Description
HII	HII Region
Mas	Maser
IR	Infrared Source
Y*O	Young Stellar Object
Y*?	Candidate Young Stellar Object
cor	Dense Core
SFR	Star Forming Region
Cld	Cloud
MoC	Molecular Cloud
DNe	Dark Cloud / Nebula
glb	Globule (low-mass dark cloud)

Table 5.1 SIMBAD star formation related object types ('otypes')

The output data from these searches has been collated in Appendix A and wherever possible I have included the different values for common properties, identified by source catalogue and method of determination, to allow for a comparison with my findings. This is especially useful for distances, for example, where I have been able to estimate a distance from the Gaia data crossmatch (see chapter 6) it is useful to compare to a kinematically derived distance.

Chapter 6

Gaia Proper Motion and Photogeometric Distance

The Gaia DR3 catalogue (Vallenari et al. 2023) provides measurements of proper motion in both RA and DEC for 1.5 billion stars down to G ~ 21 mag (see chapter 2 for a description of the Gaia project). The CMDs produced in chapter 4 are useful to help determine the reality of a given candidate and some basic properties. The decontamination is a statistical process and by its very nature is making assumptions about the membership of a given star based on its position in colour-space relative to a control field. The ESA Gaia project provides us with the ability to look at the proper motion (PM) and parallax of a large number of stars with a distance accuracy of 20% at 30 kpc (Luri et al. (2018)). The Gaia data have their limitations in the context of this thesis, with its focus on young, emerging clusters there are still a significant number of Open Cluster Candidates and known candidate open clusters that can benefit from its inclusion.

Chapter 2, Figure 2.18, shows a cluster of co-moving stars with very similar PM (where the extinction is minimal and uniform). Cross-matching the Bonatto algorithm high-probability cluster members with the Gaia stars for the target region of a cluster we should see a fairly tight grouping in PM-space and this would provide support to the cluster hypothesis. In some instances, such as Figure 2.18 the group may even have a PM

quite different form the field, although this depends on the position of the cluster relative to the field. A local cluster will have a motion in line with that of the rest of the visible field, however a larger cluster at a much greater distance would have a very different PM to the field as a result of the differential rotation of the Galaxy and this would become more exaggerated at longitudes toward or away from the Galactic centre. It is therefore unlikely that we would see many clusters as clearly separated in PM as that in Figure 2.18, in the majority of cases we would expect to see members of a cluster candidate grouped tightly together in a PM diagram. What we would not want to see is a large scatter as this would tend to indicate the group was not co-moving, and in fact the stars were unrelated and this is similar to the approaches taken in Wright, Bouy, et al. 2016 and Baumgardt et al. 2019.

Crucially, within the multi-wavelength methodology of this thesis, Gaia proper motion and photogeometric distance analysis does not constitute the primary diagnostic proof of cluster reality. Optical G-band flux is severely attenuated by interstellar dust ($A_G \approx 0.85 A_V$), so where extinction reaches $A_V \gtrsim 10$ mag the cluster members fall below the $G \approx 21$ mag detection limit (Vallenari et al. 2023) and many genuinely embedded or distant young clusters in my sample yield sparse or null Gaia cross-matches. Parallax precision also degrades, and distance inference becomes strongly biased, beyond $\sim$ 3–4 kpc (Luri et al. 2018; Cantat-Gaudin, Anders, et al. 2020). Therefore, my primary physical evidence remains the deep UGPS near-infrared Bonatto decontamination and MIST isochrone fitting presented in Chapters 3 and 4. Gaia astrometry acts strictly as secondary, independent corroboration for the less-reddened fraction of cluster members (principally evolved open clusters and bright foreground stars of emerging clusters). It follows that no distance beyond this $\sim$ 3–4 kpc limit is quoted anywhere in this thesis on the basis of Gaia astrometry alone. Where larger distances are reported, most notably the 7.7 to 12.6 kpc range discussed in Chapter 8, they derive from WISE HII kinematic measurements or from the isochrone fits of Chapter 4, neither of which depends on parallax.

6.1 Proper Motions

In order to make the proper motion plots I extracted Gaia data using a cone search centred on the clusters arithmetic coordinates as determined from the decontamination and characterisation process. This gives the field of stars in the cluster apparent radius. In proper motion most clusters tend to have a PM spread in RA and Dec of about 1 mas yr^{-1} at a distance of $\sim$ 1 kpc (Hunt and Reffert 2021). Generalising this we see that;

$$\begin{aligned} \sigma_{\mu} &= \frac{\sigma_{\upsilon}}{0.00474 \times D} \\ \sigma_{\upsilon} &= \sigma_{\mu,known} \times 0.00474 \times D_{known} \\ &= 1 \text{ mas yr}^{-1} \times 0.00474 \times 1000 \text{ pc} \\ &= 4.74 \text{ km s}^{-1} \end{aligned} \tag{6.1.1}$$

where σ_{μ} is the proper motion dispersion in mas yr^{-}1, σ_{υ} is the is the intrinsic one-dimensional velocity dispersion of the stars within the cluster, measured in km s^{-}1, D is the distance to the cluster in pc and the constant 0.00474 converts units from km s^{-}1 to mas yr^{-}1 for the given distance in pc. Taking the value of σ_{υ} from above, this means we can say that the dispersion at any distance is given by;

$$\begin{aligned} \sigma_{\mu,new} &= \frac{\sigma_{\upsilon}}{0.00474 \times D_{new}} \\ &= \frac{4.74 \text{ km s}^{-1}}{0.00474 \times D_{new}} \\ &= \frac{1000}{D_{new}} \end{aligned} \tag{6.1.2}$$

where $\sigma_{\mu,new}$ is given in mas yr^{-}1 and D_{new} is given in pc. From this we would hope to see clumps of high-probability Bonatto stars within this $\sigma_{\mu,new}$ limit. For each candidate the median photogeometric distance and a figure for $\sigma_{\mu,new}$ is calculated. The results of this test are included in the table in appendix A in columns *gaia_dist_med, exp_pm_disp, act_pm_disp, diff_pct* & *gaia_real*.

It is important to note here that, as alluded to in the introduction above, not all the Bonatto stars appear in the Gaia catalogue as they either fall below the detection limit or have other quality issues. This particularly affects faint and emerging clusters where many members are only visible in the infrared, meaning that the Gaia cross-match has

potentially limited usefulness for emerging candidates in the parameter space of interest to this thesis. However, the high number of open clusters and candidates identified here could benefit from this additional data, particularly in terms of helping confirm previously identified Open Cluster Candidates.

I initially applied this technique across a sample of known clusters from the visual search to gain confidence with the process across a range of morphologies. The process takes the output of the Bonatto decontamination and cross-matches to the Gaia data using a 0.5 arcsec matching criterion. The results from these matches are plotted in §6.3, where the top left panel shows the basic RA/DEC plot with Gaia field stars as black circles, and Bonatto selected stars as red points. The top right tile shows the mean photogeometric distance (see section 6.2) against Dec. This gives us a view of the distance range of the field vs the cluster. We would like to see a well defined vertical line of red points to indicate a strong distance correlation in the Bonatto selected stars, which would support the assertion that they are related. The bottom left tile shows the proper motion plot while the bottom right tile shows the proper motion plot with error bars for each point. The photogeometric distance panels in these figures show the full range of matched sources, extending well beyond the $\sim$ 3–4 kpc within which Gaia distance inference is reliable. The range is retained rather than truncated so that the distribution of matched sources is shown in full; no distance claim in this thesis rests on individual sources beyond that range, and the median values quoted for each candidate lie within or close to it. The wide scatter at large distances is itself the signature of the foreground and background contamination discussed above.

6.2 Parallax Derived Distances

In addition to the proper motion Gaia provides parallax measurements which can be used to infer the distance to a star (see Luri et al. 2018, Bailer-Jones, Rybizki, Fouesneau, Mantelet, et al. 2018 and Bailer-Jones, Rybizki, Fouesneau, Demleitner, et al. 2021), although the accuracy of the parallax, ω, degrades rapidly with distance and magnitude it can still be useful for those stars with good parallax measurements. The uncertainties in

the parallax, σ_ω, from the second data release (DR2) are typically 0.04 mas for sources with G-band brightness greater than ~14 mag, about 0.1 mas for sources with G-band brightness ~17 mag and around 0.7 mas around the detection limit of ~20 mag although all of these uncertainties are known to be underestimated by between 8-12% for faint sources $G \geq 16$ mag outside the Galactic plane, and up to 30% for bright stars with $G \leq 12$ mag Luri et al. 2018). The Gaia early data release 3 (EDR3) reduces the random and systematic errors in EDR3 by 30% vs DR2 but even a small uncertainty in the parallax can represent a large range of distances, and this becomes an increasing problem as ω becomes smaller. Indeed, it is possible to have a negative parallax in some cases (24% of EDR3 - Bailer-Jones, Rybizki, Fouesneau, Demleitner, et al. 2021), although obviously this does not equate to a negative distance.

The accuracy of the parallax and how this is translated into a distance estimate is therefore of great importance if we are to be able to make use of this immensely valuable source of data. As described in Luri et al. 2018, the problem becomes clear when looking at the naive translation of parallax to distance by merely inverting the measured parallax, ω, where distance, r, would be given by $r = \frac{1}{\omega}$). In practice as ω becomes small this relationship does not hold because of the non-linear effect of uncertainty in the value of ω (σ_ω). The only time where $r = \frac{1}{\omega}$ is true is where we know the true value of the parallax, ω_{True}, and using the observed value of ω instead gives us $\rho = \frac{1}{\omega}$ where the value of ρ varies as a probability distribution of the value of ω around the value of ω_{True}:

$$P\left(\omega \mid \omega_{True}\right) = \frac{1}{\sigma_\omega \sqrt{2\pi}} \exp\left(-\frac{\left(\omega - \omega_{True}\right)^2}{2\sigma_\omega^2}\right) \tag{6.2.3}$$

$$\begin{aligned} P\left(\rho \mid \omega_{True}\right) &= P\left(\omega = 1/\rho \mid \omega_{True}\right) \cdot \left|\frac{\partial \omega}{\partial \rho}\right| \\ &= \frac{1}{\rho^2 \sigma_\omega \sqrt{2\pi}} \exp\left(-\frac{\left(1/\rho - \omega_{True}\right)^2}{2\sigma_\omega^2}\right) \end{aligned} \tag{6.2.4}$$

If we then try to apply $\rho = \frac{1}{\omega}$, again making the naive assumption that ω is free of systematic error and is normally distributed around ω_{True}, we can see that for large values of ω the shape of the probability distribution for the value of ρ looks normal and

symmetrical around the value of r (as shown in the diagram to the left of Figure 6.1). However, as ω decreases for objects at greater distance the value of ρ can differ significantly from the value of r (right of Figure 6.1). The distribution here is no longer symmetrical and has a strong bias to shorter distances as well as a significant tail to much longer distances than the true value of r. It is not possible to correct for this bias without prior knowledge of ω_{True}.

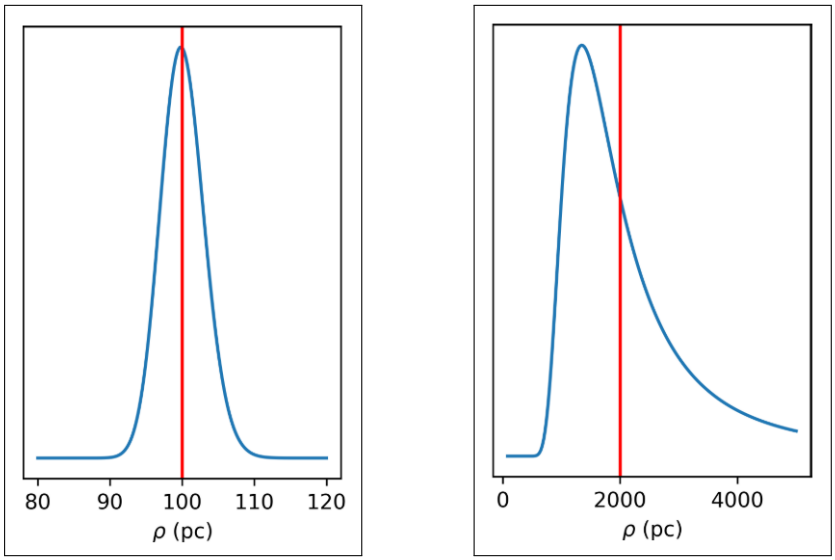


Figure 6.1 The two figures show the probability distribution functions for $\rho = 1/\omega$ in two cases. The left image is for an object at $r = 100$ pc with $\sigma_\omega = 0.3$ mas. The image on the right is for an object at $r = 2000$ pc with $\sigma_\omega = 0.3$ mas. Image credits Luri et al. 2018.

To improve the distance estimates Bailer-Jones et al have written a number of papers iterating on the problem of getting accurate distances from the Gaia parallaxes by introducing different direction-dependent priors that reflected Gaia's ability to select stars (see Bailer-Jones 2015; Astraatmadja and Bailer-Jones 2016a; Astraatmadja and Bailer-Jones 2016b; Bailer-Jones, Rybizki, Fouesneau, Mantelet, et al. 2018; Bailer-Jones, Rybizki, Fouesneau, Demleitner, et al. 2021). The latest version of the distance estimate described in Bailer-Jones, Rybizki, Fouesneau, Demleitner, et al. 2021 and known as the photogeometric distance, adds another dimension to the distance prior based on parallaxes by using the source's G-band magnitude and BP-RP colour along with a model of the direction-dependent distribution of stellar extincted absolute magnitudes. Both the geometric (parallax only) and photogeometric components use

priors based on a Gaia Early-DR3 (EDR-3) mock catalogue developed by Rybizki et al. 2020 which provides the positions, distances, magnitudes, colours, and extinctions of 1.5 billion stars as an example of what was expected to appear in EDR-3. This catalogue is based on the Besançon Galactic model and PARSEC stellar evolutionary tracks and provides a way to construct the priors that are used in the probability calculations.

In summary the photogeometric distance is defined as $P_{pg}\left(r \mid \omega, \sigma_\omega, G, c, p\right)$ where r is the real distance, ω is the parallax, σ_ω is the uncertainty in the parallax, G is the apparent magnitude in G-band, c is the $BP - RP$ colour and p is the HEALpixel number (which depends on Galactic latitude and longitude). In summary this becomes:

$$P_{pg}\left(r \mid \omega, \sigma_\omega, G, c, p\right) = P\left(\omega \mid r, \sigma_\omega\right) P\left(r \mid p\right) P\left(Q_G = G - 5\log_{10} r + 5 \mid c, p\right) \tag{6.2.5}$$

where Q_G is the true magnitude of the star and is equivalent to the apparent magnitude plus the attenuation:

$$Q_G \equiv M_G + A_G = G - 5\log_{10} r + 5 \tag{6.2.6}$$

The photogeometric distances are made available by the Astronomisches Rechen-Institut (ARI) in Heidelberg as a complements table for Gaia E-DR3 and DR3 data. For this work I used an ADQL JOIN statement to extract the photogeometric mean value for each source when making the cone search for each cluster.

6.3 Gaia Control Plots

NGC-6705

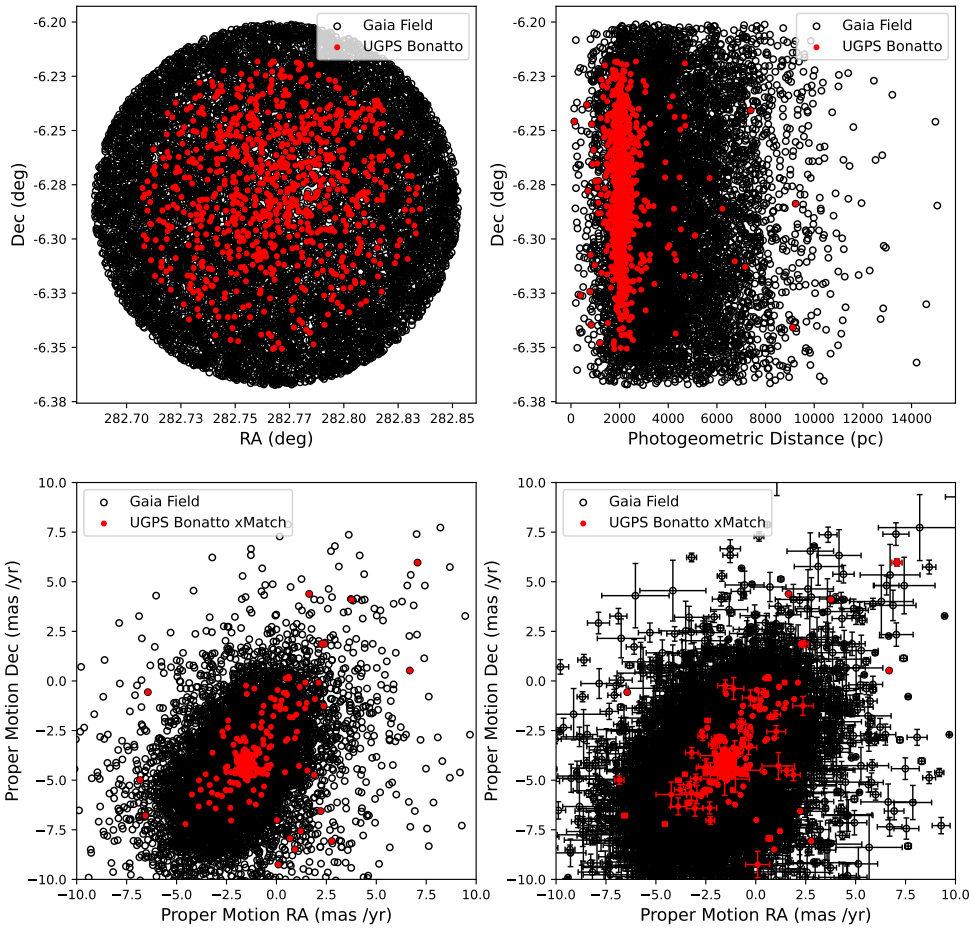


Figure 6.2 NGC-6705, an open cluster in a rich field with a literature distance of 1.9 kpc (Wenger et al. 2000). A crowded field but the Bonatto members show a clear increase in density towards the centre. The proper motion plots show a tight clump in the centre as one would expect with a small amount of scatter from the field, error bars for the Bonatto stars are also tight indicating reliable measurements. The distance tile, top right, shows a strong vertical grouping (the photogeometric axis) of Bonatto stars around the 2 kpc distance with minimal scatter. This agrees well with the literature estimate and would provide strong support for a cluster. The photogeometric distance axis is shown in full; values beyond ~ 3–4 kpc are not reliable (Section 6.2).

NGC-2401

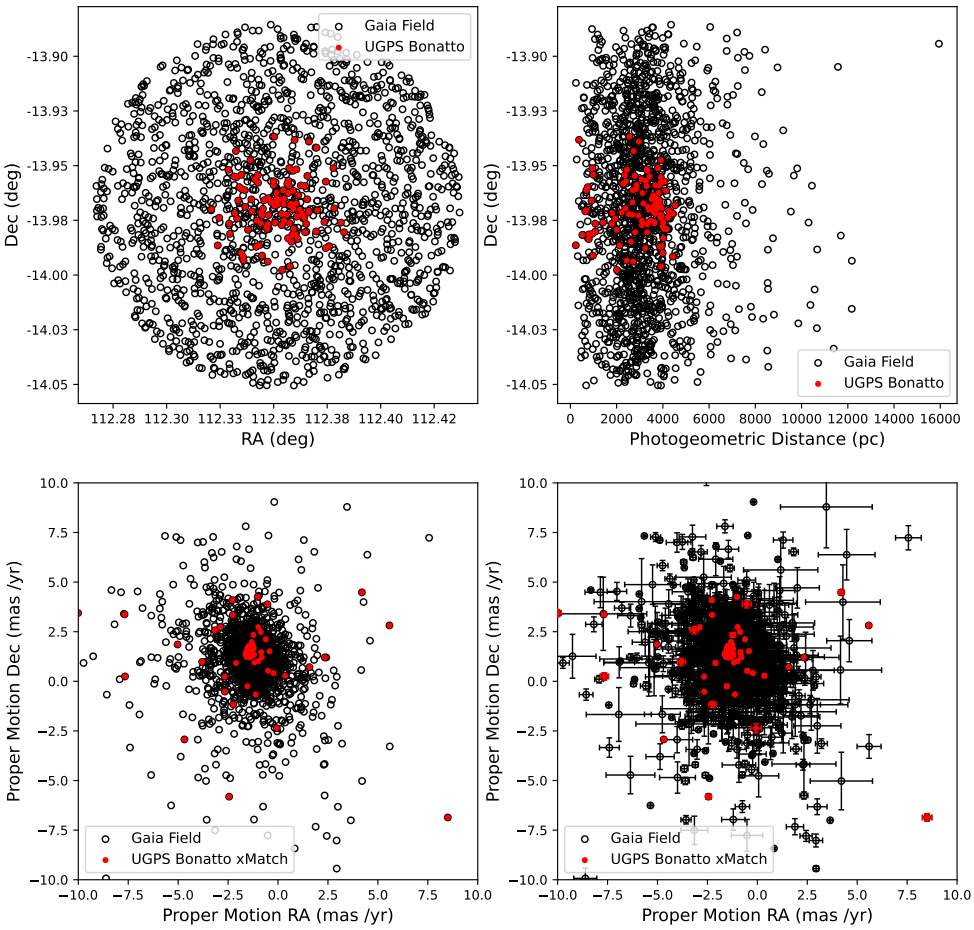


Figure 6.3 NGC-2401, Open cluster towards the outer Galaxy with an apparent diameter of 1.8 arcmin and literature distance of between 3.8 and 5.8 kpc (MWSC and DAML02 respectively). The top right tile shows a smaller central over-density which has a high number of Bonatto stars selected. The top right tile shows the photogeometric distance plot where we see a clump corresponding to the central over-density at a mean Gaia distance of approximately 3.9 kpc (agreeing with the MWSC value), although there is a significant amount of scatter throughout the plot likely caused by inadvertent selection of field stars in the target region. The proper motion plot looks convincing with a tight central group, although there is some scatter. The distance and proper motion result here provide a convincing case for a cluster as we would expect. The photogeometric distance axis is shown in full; values beyond $\sim$ 3–4 kpc are not reliable (Section 6.2).

6.4 Example Plots - Unknown Candidates

Candidate - 122

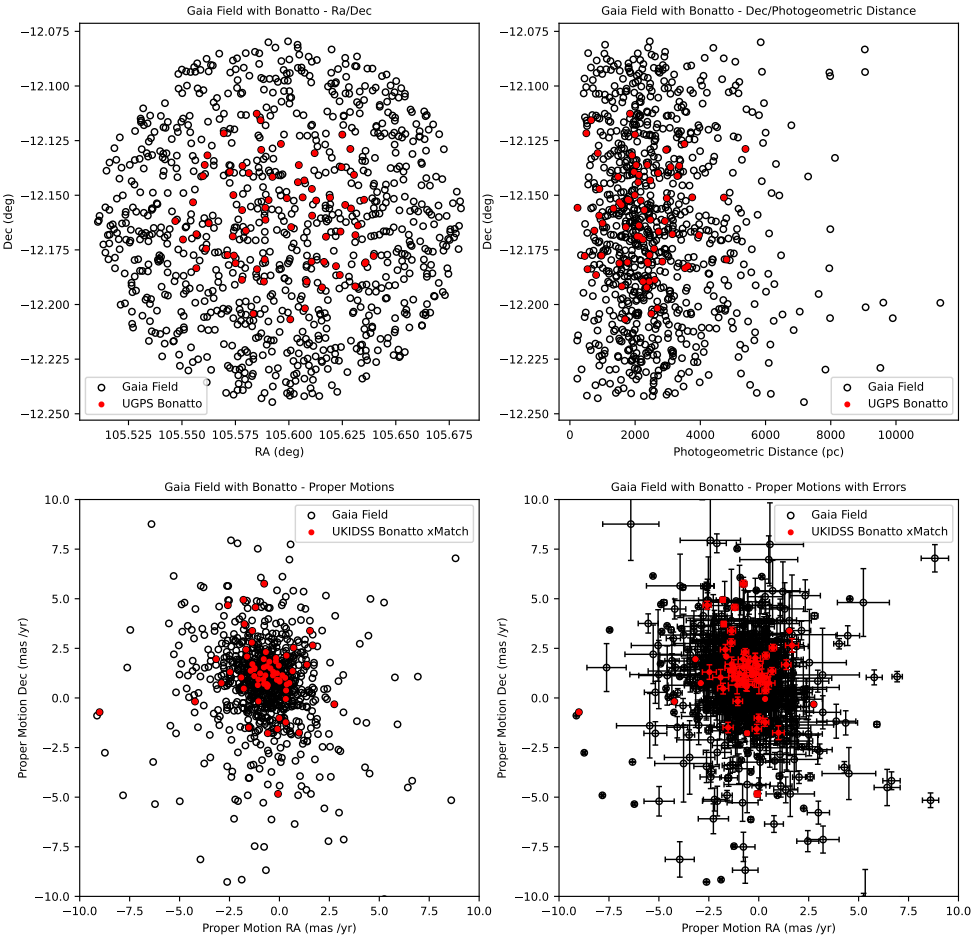


Figure 6.4 Candidate 122 - Open Cluster Candidate. Gaia matched 68 of the 159 Bonatto members which are loosely grouped in PM space. The error bars are small for the members in the centre of the group which gives confidence that the measurements are reliable. The photometric distance displays much scatter although there is a higher density around the 3 kpc distance, and indeed the median of the distance measurement is 2.9 kpc. The grouping of the matched points in PM space is supportive of a cluster and while there is the hint of a grouping in distance there is much scatter. This plot certainly shows a group of stars co-moving with the field, however the Gaia data alone would not be sufficient to claim a new cluster. It does however provide a degree of support to a hypothesis built using many strands of evidence. The photogeometric distance axis is shown in full; values beyond ~ 3–4 kpc are not reliable (Section 6.2).

Candidate - 211

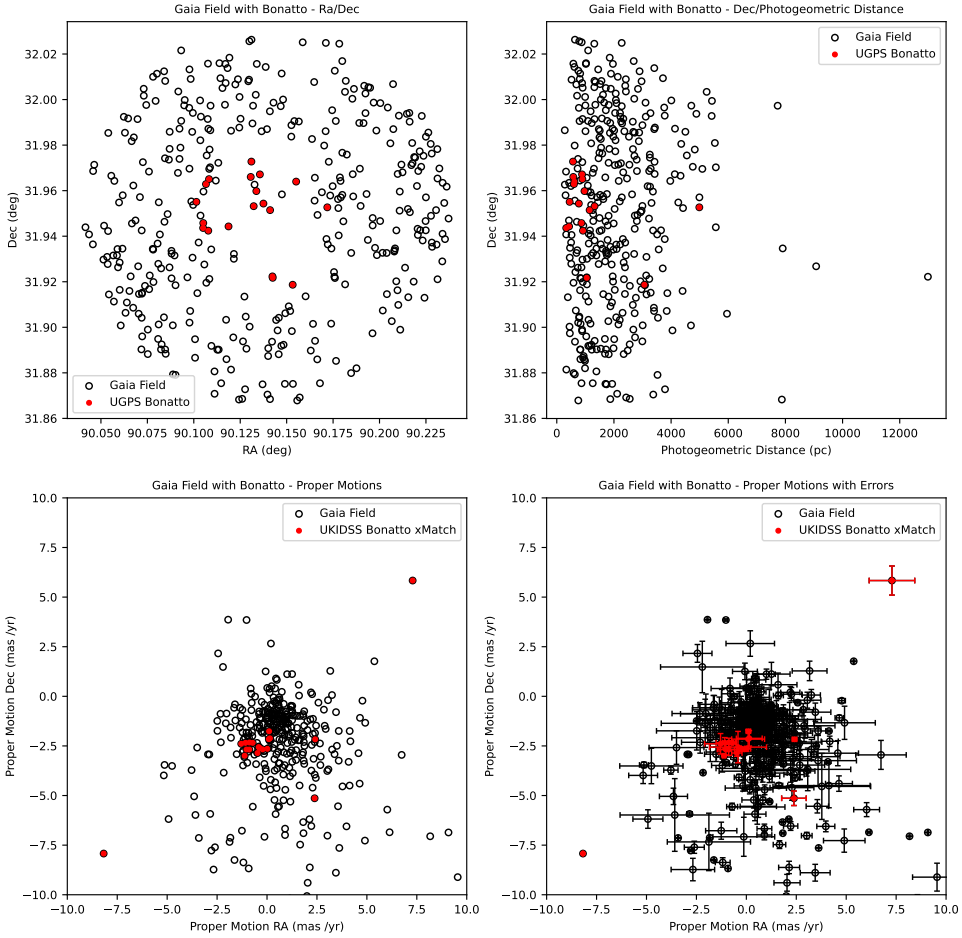


Figure 6.5 Candidate 211 - Emerging Cluster Candidate. This is a small, emerging cluster in a highly extincted field. Gaia has only matched 19 of the 88 Bonatto members however they form a tight grouping in PM space, slightly out of step with the bulk motion of the field. The photogeometric distance shows a distinct grouping at 1 kpc, which agrees with the median distance. Together, these features support a cluster hypothesis - we have a group of related stars with a proper motion slightly different to the rest of the field, with similar photogeometric distances. The photogeometric distance axis is shown in full; values beyond $\sim$ 3–4 kpc are not reliable (Section 6.2).

6.5 Gaia Cross-match Summary

This technique is extremely useful for adding the support of a co-moving group, and perhaps a photogeometric distance estimate, especially across the open clusters in the UGPS sample, as shown in §6.3. It was expected that in the cases of high extinction candidates that we would get significantly reduced numbers of stars cross-matching, and this is indeed the case for the majority, however even in some of these cases the cross match can still produce useful data that can lend support to a cluster hypothesis and even provide a distance estimate (see §6.4). Hunt and Reffert (2023) found that an open cluster candidate in the Gaia data at a distance of $\sim$ 1 kpc typically have a proper motion dispersion of $\leq$ 1 mas yr^{-1} (Hunt and Reffert 2021) so the dispersion figure for each candidate is calculated for the median photometric distance of its cross-matched members and used to flag any candidates with a PM dispersion $\geq \sigma_{\mu,new}$. This along with the cross-matches and plots for all candidates is included in the summary data in appendix A and is discussed in more detail in chapter 8. This will be a powerful technique to add to the analysis of infrared candidates once we have instruments such as the Nancy Roman Telescope and ESA's Extremely Large Telescope which should be able to produce proper motion data in the near infrared with precision near to that of Gaia, (see Sosey et al. 2021 and Matteo et al. 2019 for details).

As emphasised at the outset of this chapter, these Gaia cross-matches serve here as purely secondary astrometric corroboration for accessible stars and are not be used as a primary criterion to accept or reject any candidates.

Chapter 7

HDBSCAN Clustering Algorithm

An alternative approach to identifying clusters in the Gaia data uses a density-based clustering algorithm with the proper motions and parallax (or distance) estimates. Several attempts have been made with this approach and have had some notable success, see Kounkel and Covey (2019) and Kounkel, Covey, and Stassun (2020), and Hunt and Reffert (2021) and Hunt and Reffert (2023) for example. This approach offers an alternative to the Bonatto method of colour-based member selection, and provides an independent way to search for clusters in the regions of interest. This approach would still likely struggle with the smaller, fainter candidates or those with high extinction as they will lack detections in the Gaia data. It may however prove to be successful for the large number of potential open clusters in the sample, and with infrared instruments such as the Nancy Roman Telescope and Extremely Large Telescope soon to be producing proper motion data with precision near to that of Gaia, (see Sosey et al. 2021 and Matteo et al. 2019 for details), the technique can be applied to fainter, smaller targets in due course.

HDBSCAN is a clustering algorithm developed by Campello, Moulavi, and Sander (Campello, Moulavi, and Sander 2013) and stands for Hierarchical Density-Based Spatial Clustering of Applications with Noise. It extends the previous algorithm, DBSCAN, by converting it into a hierarchical clustering algorithm, and then using a technique to extract a flat clustering based in the stability of clusters. HDBSCAN uses an approach which makes few implicit assumptions about the clusters. It is a non-parametric method that looks for a cluster hierarchy shaped by the multivariate

modes of the underlying distribution. Rather than looking for clusters with a particular shape, it looks for regions of the data that are denser than the surrounding space.

7.1 Methodological Role and Avoidance of Circularity

When interpreting density-based clustering algorithms such as HDBSCAN across a candidate sample initially selected via visual spatial overdensities, extreme caution must be exercised to avoid circular reasoning. If a candidate is identified because it represents a spatial overdensity on a sky image, demonstrating that an automated spatial clustering algorithm also groups those stars cannot be claimed as independent proof that the cluster is physically real. Furthermore, like my Bonatto analysis, HDBSCAN applied to Gaia astrometric coordinates is inherently restricted by optical extinction limits.

Consequently, throughout this chapter, HDBSCAN is treated strictly as a secondary, corroborative tool rather than a primary diagnostic. My specific objective is to test whether an independent, non-parametric clustering algorithm applied to 5D astrometric and spatial space ($RA, Dec, \mu_\alpha, \mu_\delta, \varpi$) recovers the same stellar membership cleanly isolated by my near-infrared Bonatto colour-magnitude decontamination. Where significant cross-match overlap occurs ($\geq 25\%$), HDBSCAN provides valuable kinematic corroboration to my primary MIST isochrone fits. However, the absence of an HDBSCAN cluster detection in deeply embedded or distant regions is treated as an expected limitation of optical astrometry rather than evidence against the cluster hypothesis.

7.2 Overview of the Algorithm

The HDBSCAN algorithm uses five stages to analyse the data as described in McInnes, Healy, and Astels 2017 and McInnes 2024. These are summarised below.

7.2.1 Transform the Data Space Using Density

The data set we are trying to analyse comprises dense regions (where the separation between data points is small relative to the distribution as a whole) and sparse regions, that is to say that regions where the separation between data points is large (see Figure 7.1). HDBSCAN uses the density-based probability distribution of the data and it calculates this by looking at the radial distance from each point at which it can reach a specific number of neighbouring points, this is known as the kth neighbour distance. This is stated more formally as the core distance for the value of the parameter k, of a point x, i.e. $\text{core}_k(x)$. A point which has a small value of $\text{core}_k(x)$ lies in a denser region and conversely a point with a high value would lie in a sparse region.

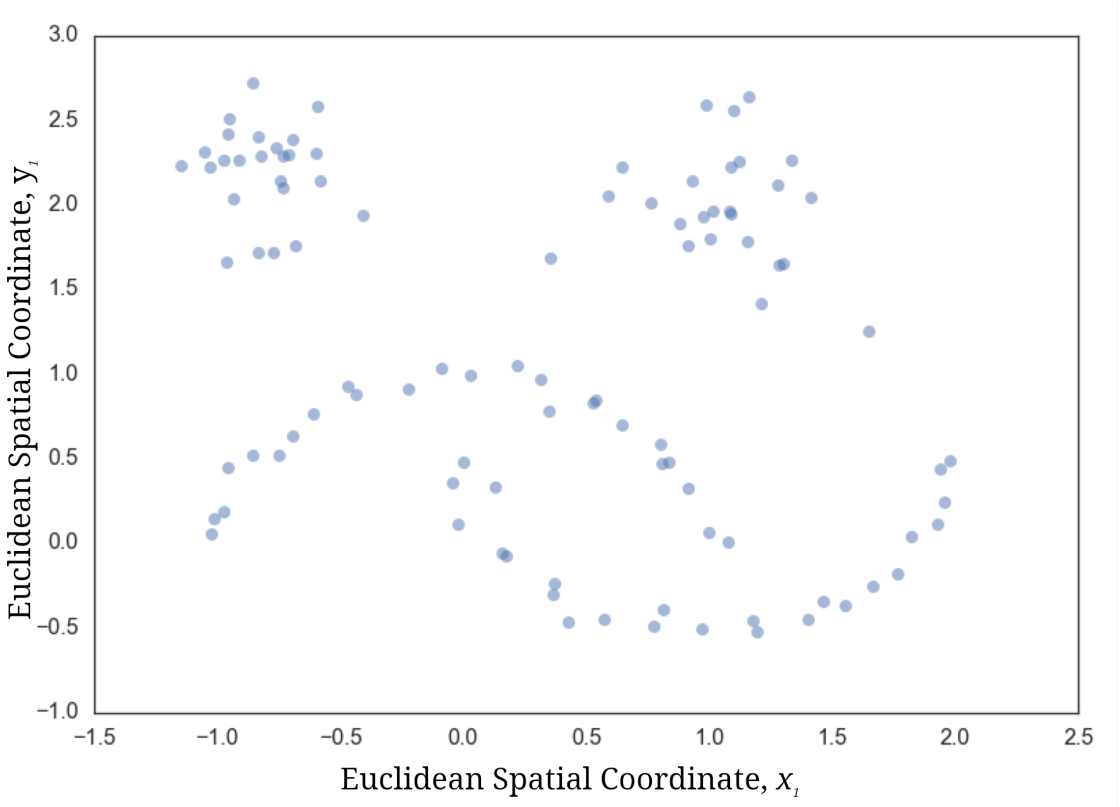


Figure 7.1 The data set used in this image is the basis for demonstrating the HDBSCAN algorithm. It has four clusters of different shapes and densities are not all spatially distinct from one another in order to demonstrate the capabilities of the algorithm. In this schematic plot, the x- and y-axes represent dimensionless Euclidean spatial coordinates (x_1, x_2) of the synthetic dataset. Image credit McInnes 2024

The algorithm uses an externally set parameter, `min_samples`, to set the value of k

neighbours when calculating the core-distance and this parameter controls how sensitive the clustering will be (see fig. 7.2). Setting `min_samples` to a higher value results in the algorithm selecting denser regions and declaring more of the data as noise. The optimal value for `min_samples` is entirely dependent on the data set being used and requires some testing to determine and in practice the automation I developed for this work uses a range of input parameters for each cluster region enabling the selection of the best fit for the data.

By computing the core distance for every data point and inverting the values of $\text{core}_k(x)$, HDBSCAN can estimate the true shape of the underlying probability density function, $f(P_d)$.

$$f(P_d) = \frac{1}{\text{core}_k(x)} \tag{7.2.1}$$

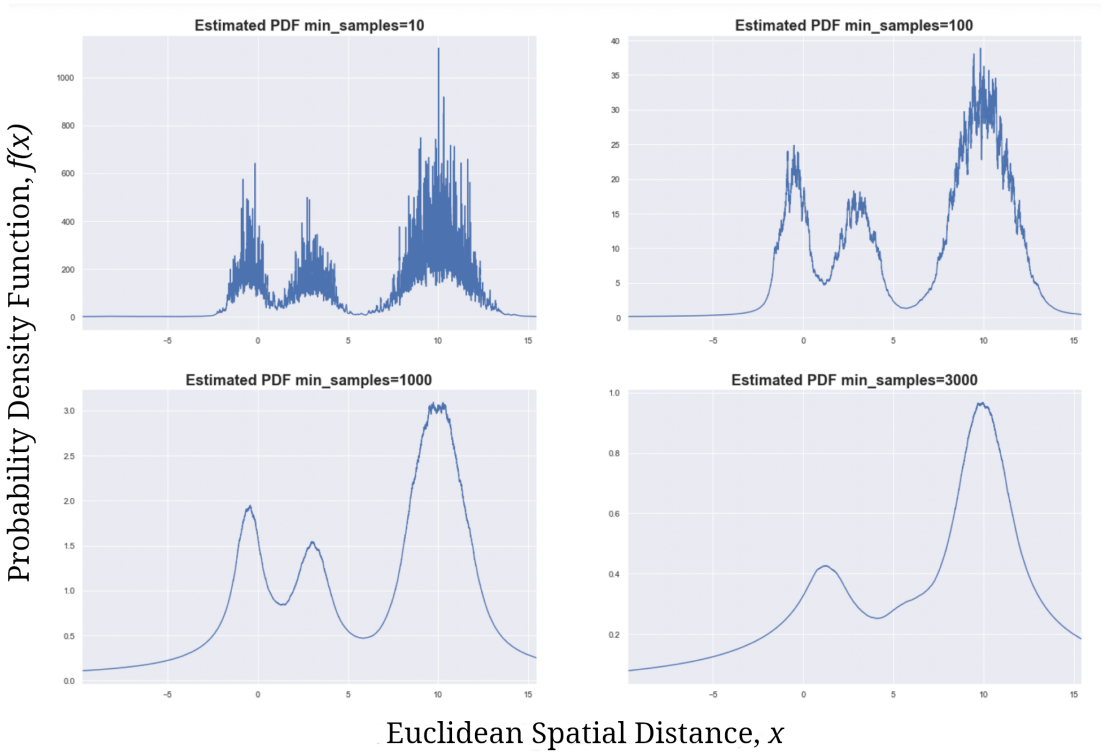


Figure 7.2 The figure shows four values for `min_samples` and their effect on the form of the recovered PDF. The greater the value of `min_samples` the smoother the PDF but the less sensitive it is to smaller clusters. Across all four panels, the horizontal axis represents Euclidean spatial distance, (x), while the vertical axis denotes probability density function $(f(x))$ derived from the inverted k^{th}-neighbour core distance. Image credit Berba 2020

Once the algorithm has calculated the core distances it then moves on to estimate the clusters and the hierarchy of clusters using the *mutual reachability distance*, $d_{mreach-k}$. This distance, as described in Eldridge, Belkin, and Wang 2015, uses the core distance and is defined as:

$$d_{mreach-k}(a, b) = max\{\text{core}_k(a), \text{core}_k(b), d(a, b)\} \tag{7.2.2}$$

where $d(a, b)$ is the Euclidean distance between points a and b. The effect of this metric is that the points in dense regions, with low core distance, remain the same distance from each other but sparser points are effectively pushed away from each other to be at least their core distance away from any other point. This effectively highlights the clusters as they become more defined relative to the surrounding, sparse, 'noise' and gives the algorithm its density sensitivity. The caveat here is that this is dependent upon the choice of `min_samples` (k); larger `min_samples` values interpret more points as being noise and reduce the number of points considered to be in the clusters.

7.2.2 Build the Minimum Spanning Tree

Once the algorithm has calculated the $d_{mreach-k}$ across all points it can build the minimum spanning tree very efficiently via Prim's algorithm (Prim 1957), building the tree one edge at a time and adding the lowest weight edge that connects the current tree to a vertex not yet in the tree. Figure 7.3 shows the raw dataset we want to analyse. Figure 7.3 shows the minimum spanning tree for mutual reachability distance which is different from the pure distance used in fig. 7.1.

7.2.3 Build the Cluster Hierarchy

The next step is to convert the minimum spanning tree into a hierarchy of connected components from which we can form the clusters. The algorithm does this in reverse order, i.e. it sorts the edges of the tree in increasing order by distance and then iterates through them creating a new merged cluster for each edge. The only difficult part here is to identify the two clusters each edge will join together, but it uses a disjoint-set data

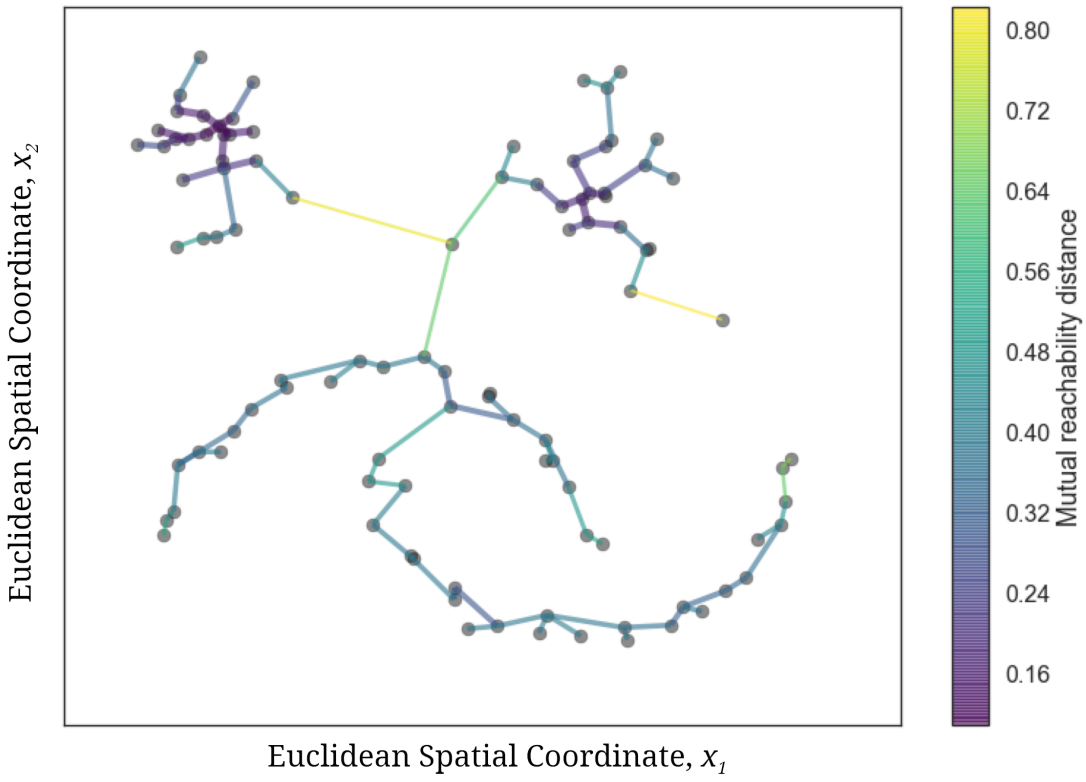


Figure 7.3 The minimum spanning tree for the sample data set. The darker the lines connecting the points the lower the $d_{mreach-k}$ and hence the denser the cluster. Spans between the clusters are at the very high end of the $d_{mreach-k}$ and can be easily filtered. The x- and y-axes denote dimensionless Euclidean spatial coordinates (x_1, x_2), while edge colour represents the mutual reachability distance ($d_{\text{mreach}-k}$). Image credit McInnes 2024

structure in order to achieve this. We can view the result as a dendrogram as shown in Figure 7.4.

This brings us to the point where robust single linkage clustering stops working. The dendrogram in Figure 7.4 shows the beginnings of a hierarchy but the splitting of the nodes is really just showing where points fallout of the cluster rather than a true separation of clusters. DBSCAN uses a cut based on a single mutual reachability distance in order to determine the clusters and this takes no account of the variability in density we see in the actual data. This is the point at which the HDBSCAN algorithm makes a significant advance.

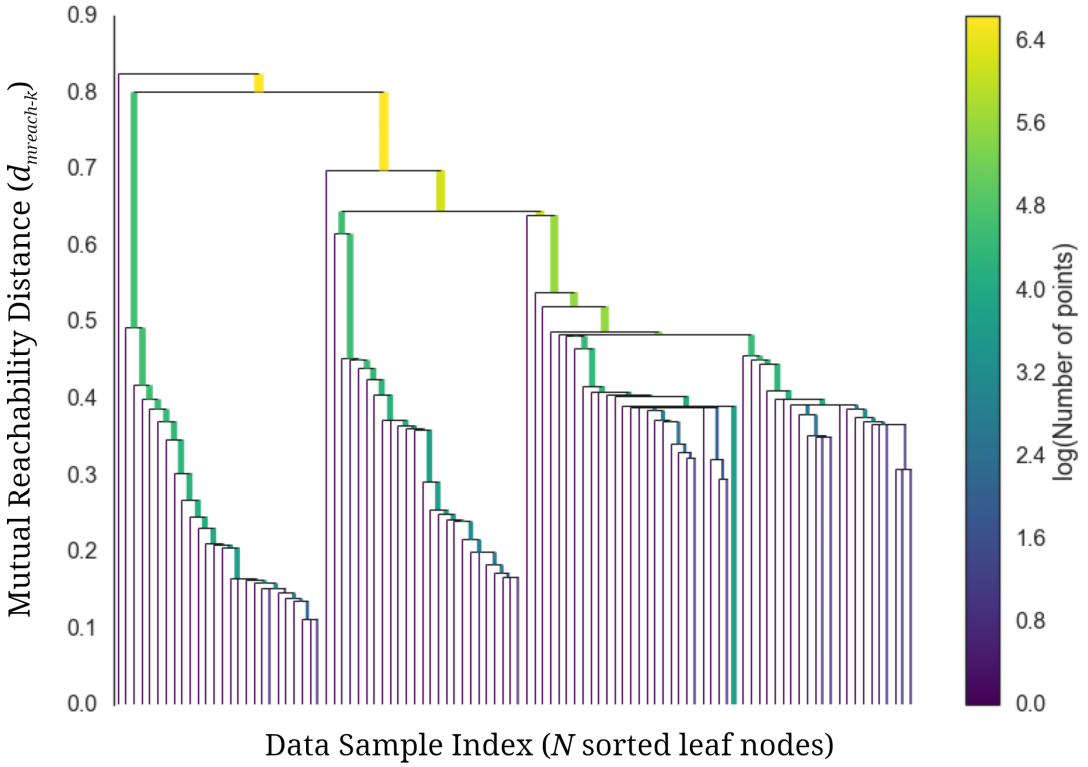


Figure 7.4 Dendrogram showing the single-linkage cluster tree built from the minimum spanning tree. This is beginning to show the cluster hierarchy but the splits are essentially points becoming detached from a cluster rather than the true separation of real clusters. We need to move to a more condensed view to be able to recover the true clusters. The horizontal axis represents the sorted sample index (N individual data points across the base), while the vertical axis denotes the mutual reachability distance ($d_{\text{mreach}-k}$) threshold at which components split or merge. Image credit McInnes 2024

7.2.4 Condense the Cluster Tree

In order to form stronger clusters and stop the fragmentation seen in the single linkage tree (fig. 7.4) the algorithm introduces an additional parameter, `min_cluster_size`, which gives control over the way that the clusters dissolve. Using this parameter the algorithm walks through the hierarchy and at each split checks whether the split has fewer members than the `min_cluster_size`, if so then it is treated as points falling out of the cluster and the remaining points retain the original cluster identity. If both sides of the split contain more points than the `min_cluster_size` then it is treated as a genuine split and both clusters are retained in the tree. Overall this results in a much smaller tree with a small number of nodes that retain information about the way the cluster dissolves

as λ increases. This is shown in Figure 7.5.

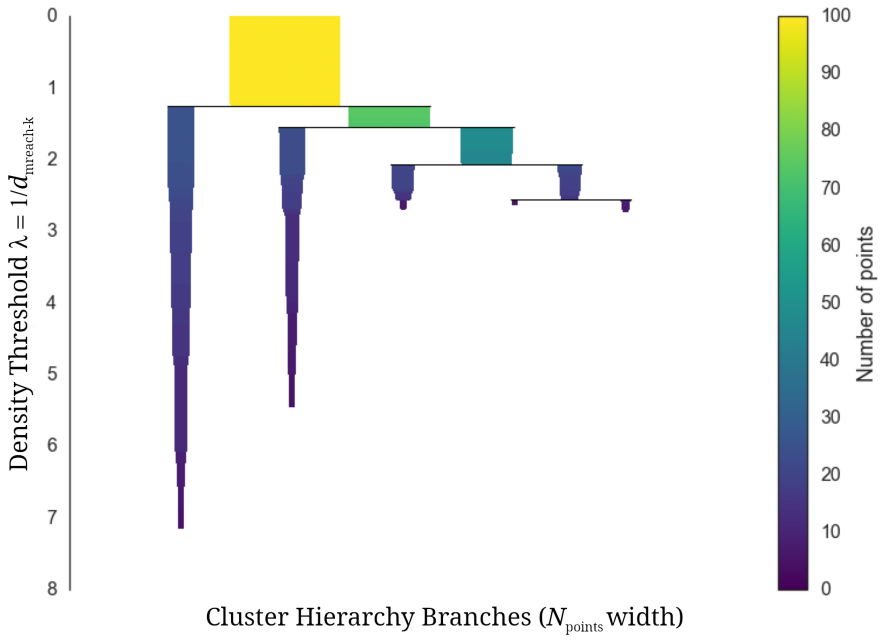


Figure 7.5 Condensed tree plot showing the effect of the `min_cluster_size` parameter in retaining the clusters. The width of the line in this diagram represents the number of points in the cluster and shows how the population diminishes and it eventually disappears as the density requirement, λ is relaxed (i.e. its value increases). he vertical axis represents the density threshold ($\lambda = 1/d_{\text{mreach}-k}$), which increases downward as clusters dissolve. The horizontal axis separates distinct hierarchical branches, with branch width proportional to the number of points (N_{points}) remaining in each cluster. Image credit McInnes 2024

7.2.5 Extract the Clusters

By default clusters are selected on the basis of the greatest area, with a restriction that no decedents of a cluster can be selected. This is known as the *Excess of Mass* (EOM) method and while this is suitable for most applications there is an alternative in the *leaf* method that selects only the leaves of the tree resulting in a greater number of smaller clusters. For this work I have used the EOM method as we are really looking for single clusters, perhaps in some cases fragmented, but we are not looking for many small clusters (although this could be appropriate in a large SFR, for instance).

In order to be able to the persistence of the clusters the algorithm uses the inverse of

$d_{mreach-k}$, i.e.

$$\lambda = \frac{1}{d_{mreach-k}} \tag{7.2.3}$$

.

From this it is then possible to define values for λ_{birth} and λ_{death} to define where a cluster splits and becomes a new cluster and the value at which the cluster either splits into smaller clusters or dissolves. For a given cluster, the algorithm notes for each point, p in that cluster a value λ_p as the value at which that point became detached from the cluster. This value will be between λ_{birth} and λ_{death} since the point either falls out of the cluster at some point in the cluster's lifetime, or leaves the cluster when the cluster splits into two smaller clusters.

Now, for each cluster it can compute the *cluster stability* as

$$\sum_{p \in \text{cluster}} \left(\lambda_p - \lambda_{\text{birth}}\right) \tag{7.2.4}$$

To make the final cluster selections the algorithm first assumes all leaves are clusters, then works up through the tree towards the root using the cluster stability metric calculated in Equation 7.2.4 above. If the sum of the stabilities of a clusters children is greater than the stability of the cluster then we use the sum of the child stabilities as its stability value, conversely if the clusters stability value is greater then the cluster is selected and all children deselected. Once this is done for all leaves and clusters and it has reached the root then the cluster selection is finalised and returned. An example of this is shown in Figure 7.6.

Points not in a selected cluster are classed as noise and given a label of -1. For points in the clusters the value λ_p at which the points separate from the cluster can be normalised to give a membership-strength or probability value. This value is then attached to each point as the `probabilities_` attribute of the `clusterer` object. This value can be used to give an additional dimension to the cluster plots, for instance by desaturating the member points colour as shown in Figure 7.7. In the figure noise points are shown as grey.

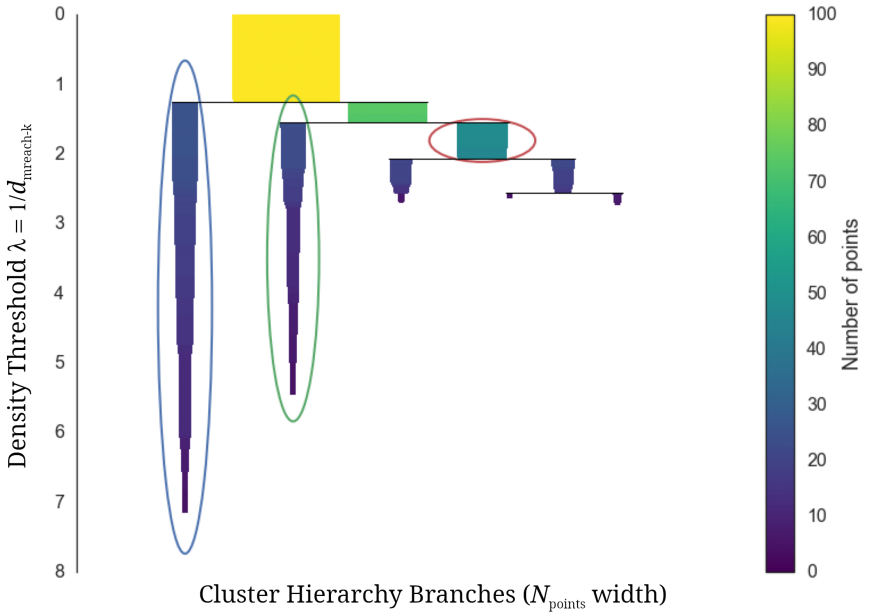


Figure 7.6 The clusterer makes selection of clusters based on the hierarchical stability of the leaves and clusters. This image shows the condensed-tree view with the selected clusters highlighted. The vertical axis plots the density threshold ($\lambda = 1/d_{\text{mreach}-k}$) across distinct cluster hierarchy branches along the horizontal axis, where branch width represents remaining cluster size (N_{points}). Highlighted ovals indicate the persistent clusters selected by maximising the Excess of Mass stability metric defined in Equation 7.2.4. Image credit McInnes 2024

7.3 Applying HDBSCAN to the Sample Candidates

Applying HDBSCAN to the sample candidates requires setting the input parameters of `min_samples` and `min_cluster_size` to values that are suited to the specifics of the dataset under consideration (in all cases I used the *Excess of Mass* selection method as described in §7.2.5). Individually setting these parameters across hundreds of candidates was not realistic so I chose 10 combinations of parameters (based on manual testing across a sub-sample of candidates) that give the best chance of sensible clusters for a range of morphologies and field densities. For each candidate the HDBSCAN clusterer iterates over this input parameter list and produces a corresponding range of output clusters, HDCLs from here on. As stated previously, clusters found in the Gaia data at a distance of $\sim$ 1 kpc typically have a proper motion dispersion of $\leq$ 1 mas yr^{-1} (Hunt and Reffert 2021) so the HDCLs are filtered to remove any clusters with a proper motion

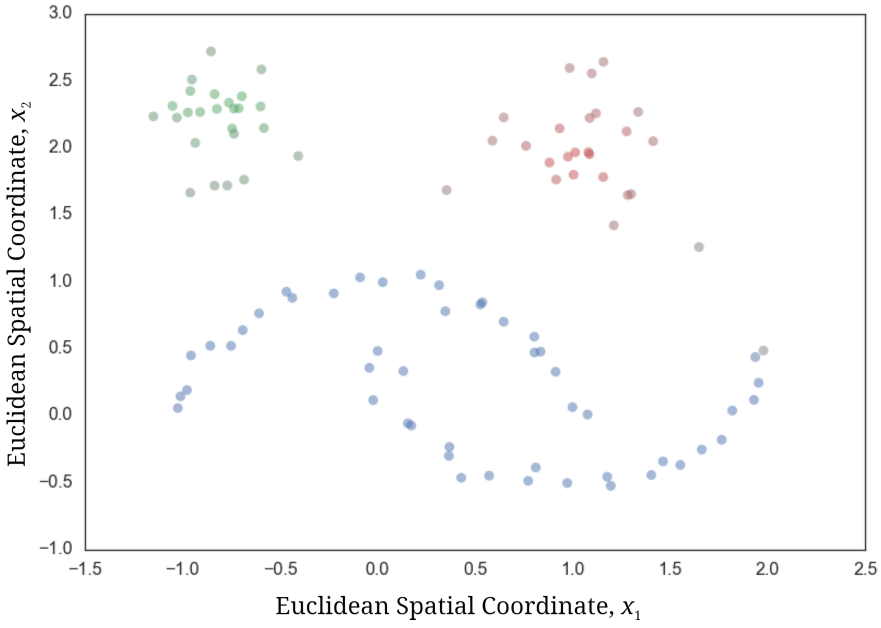


Figure 7.7 This image shows the final selected clusters, each coloured individually. Noise points are shown as grey and the saturation level of the cluster members reflects the membership probability given by the `probabilities_` attribute. It is also possible to separate the two arced clusters at the bottom of the image using a slightly different parameter mix, see Berba 2020. In this schematic plot, the x- and y-axes represent dimensionless Euclidean spatial coordinates (x_1, x_2) of the synthetic dataset. Image credit McInnes 2024

vector standard deviation > 1, for the $\sigma_\mu \leq 1$ mas yr^{-1} criteria.

Some examples of the clusters identified by the HDBSCAN algorithm are shown in §7.3.1. The Gaia field stars are shown as black dots and the HDBSCAN selected stars as blue circles. Each plot shows the results from a different set of input parameters, these are shown in the plot title, for example a plot titled `15_mcs_100_ms_5_csm_eom_subClus_1_fullCluster_rawPlot` has parameters `minimum_cluster_size` (mcs) of 100 stars, `minimum_samples` (ms) of 5, and the `cluster_selection_mechanism` (csm) is excess of mass (eom). Across all three figures HDBSCAN is able to identify a central over-density in some of the iterations, however the parameter selection results in a variation in the points selected. This makes HDBSCANs ability to select real cluster members highly unreliable and it relies on external evaluation to assess the quality of the output in each case. The purpose of using HDBSCAN in this context is not to try and identify all real cluster members, but rather to

add confidence to an existing hypothesis for each candidate. The principal sources are the UGPS CMD & CCDs as these benefit from the higher resolution and sensitivity of WFCAM and are most suited therefore to compact or young clusters. Each of the multiple HDBSCAN results for a candidate were cross-matched to the Bonatto high probability members to understand the relative completeness of the HDBSCAN clusters. If the match percentage is high it suggests there is a cluster in the field that satisfies the $\sigma_\mu \leq 1$ mas yr^{-1} criteria.

To further investigate this some CMD and CCDs were plotted from the HDBSCAN results and compared to the Bonatto results. Some examples are given below in §7.3.3 and it is clear that in the case of brighter, open clusters a good match is made.

This result in conjunction with the other strands of evidence is then used in the final judgement on the reality of a candidate. The results for some of the candidates are shown in section 7.3.2 and a table with the cross-match statistics is shown below, for example Table 7.1 where it gives the HDBSCAN input parameters, *hdbscanParams*, the proper motion standard deviation, σ_μ, the number of HDBSCAN stars selected as the cluster, *HDBSCANStars*, the number of Bonatto High probability cluster members, *BonattoHPStars*, the number of matches between the HDBSCAN and Bonatto outputs, *Matches* and the percentage match, *%matchToBonatto*.

7.3.1 HDBSCAN Example Output

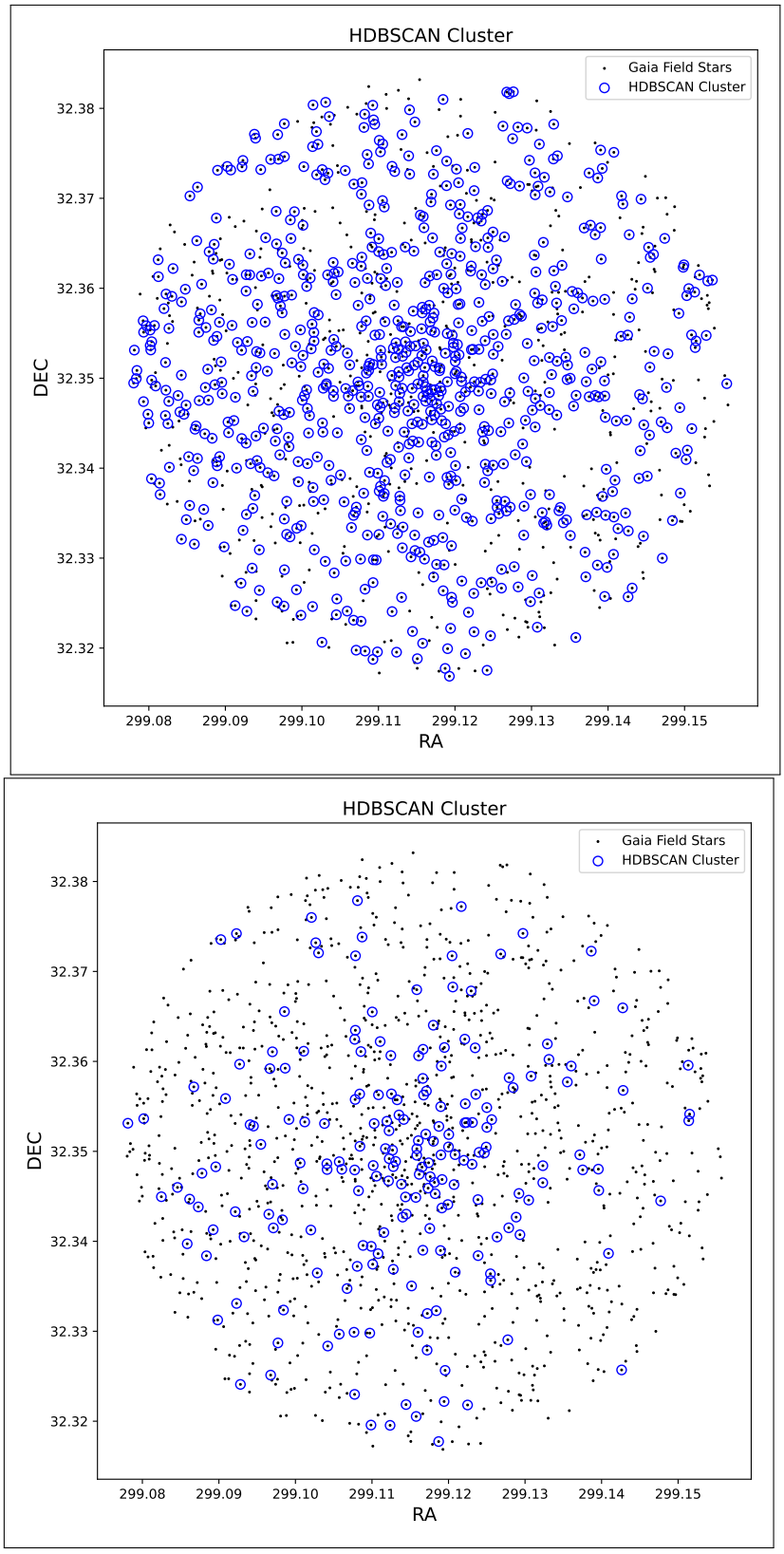


Figure 7.8 NGC-6846 - Example HDBSCAN Output Clusters.

7.3.2 HDBSCAN UGPS Cross-match

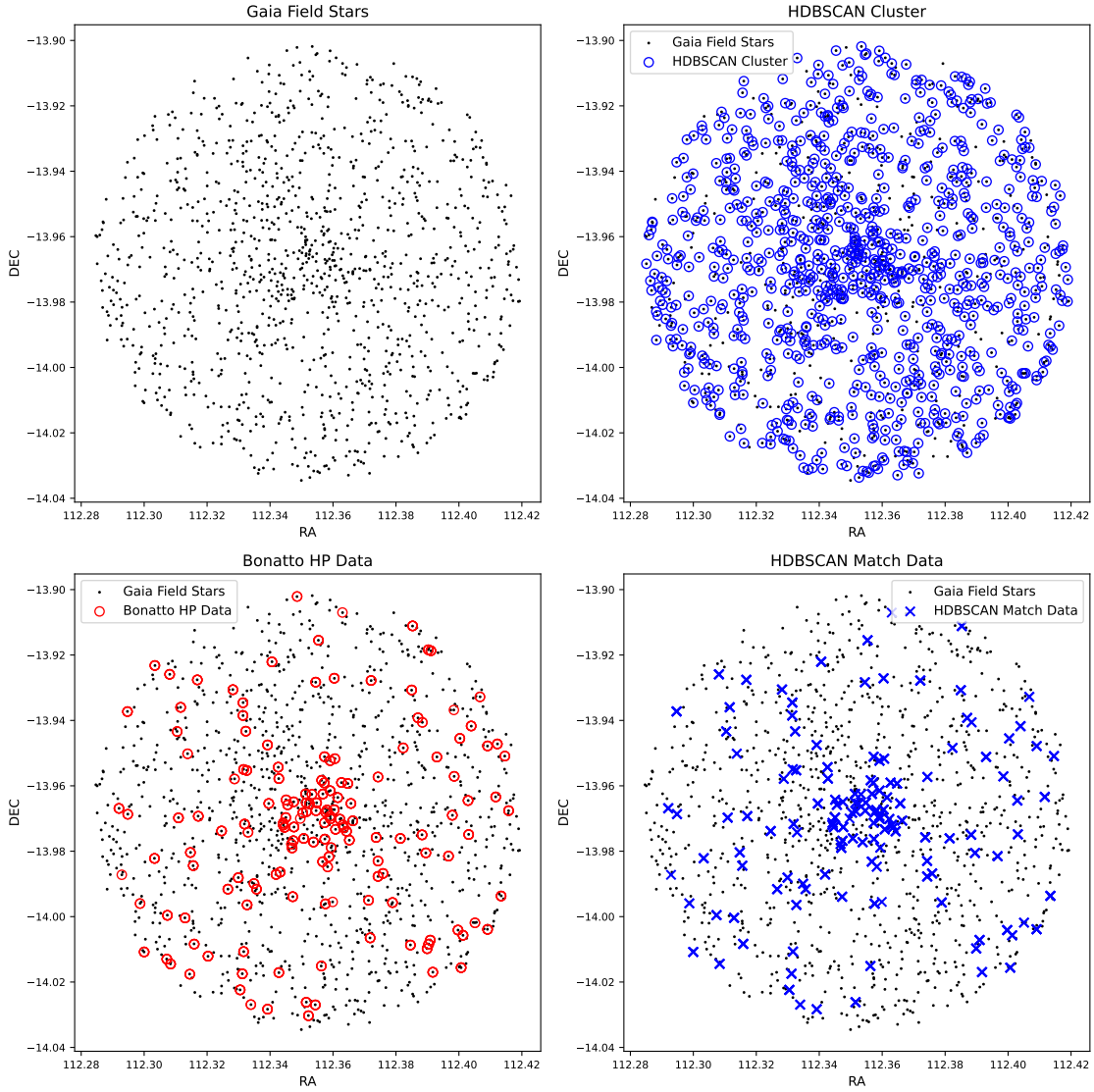


Figure 7.9 NGC-2401 - HDBSCAN-Bonatto Cross-match.

hdbscanParams	σ_μ	HDBSCANStars	BonattoHPStars	Matches	%matchToBonatto
896_mcs_20_ms_2_csm_eom	0.86	918	584	507	87
896_mcs_20_ms_4_csm_eom	0.88	912	584	503	86
896_mcs_20_ms_10_csm_eom	0.46	366	584	351	60
896_mcs_50_ms_5_csm_eom	0.33	272	584	329	56
896_mcs_200_ms_20_csm_eom	0.22	210	584	290	50
896_mcs_200_ms_10_csm_eom	0.29	238	584	314	54

Table 7.1 NGC-2401 - HDBSCAN-Bonatto Cross-match Statistics

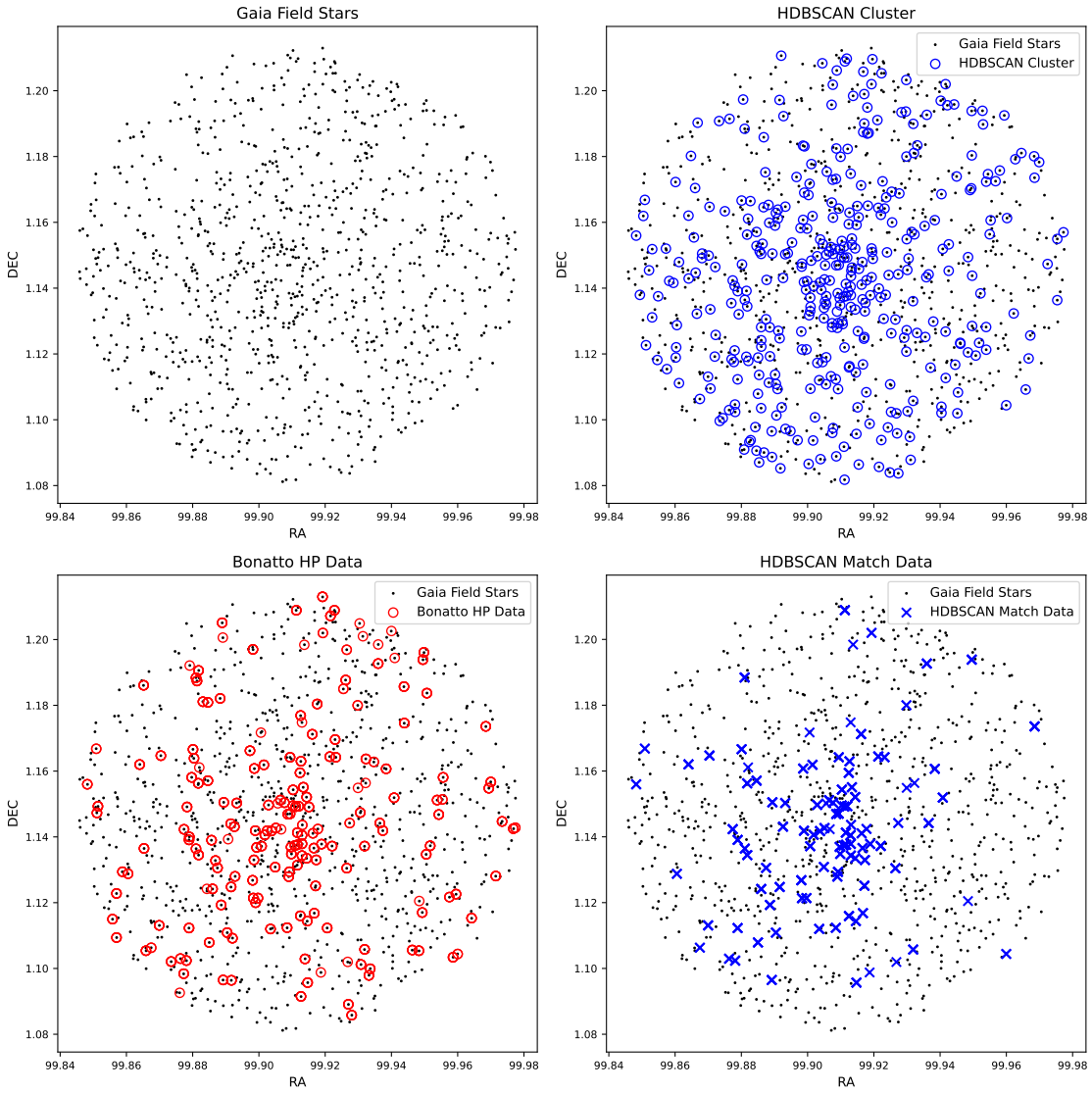


Figure 7.10 NGC-2262 - HDBSCAN-Bonatto Cross-match.

hdbscanParams	σ_μ	HDBSCANStars	BonattoHPStars	Matches	%matchToBonatto
899_mcs_20_ms_2_csm_eom	0.13	22	683	4	1
899_mcs_20_ms_2_csm_eom	0.14	21	683	17	2
899_mcs_20_ms_2_csm_eom	0.16	386	683	365	53
899_mcs_20_ms_4_csm_eom	0.04	93	683	189	28
899_mcs_20_ms_10_csm_eom	0.04	26	683	67	10
899_mcs_50_ms_10_csm_eom	0.04	51	683	126	18
899_mcs_50_ms_5_csm_eom	0.05	51	683	107	16
899_mcs_100_ms_5_csm_eom	0.05	101	683	206	30

Table 7.2 NGC-2262 - HDBSCAN-Bonatto Cross-match Statistics

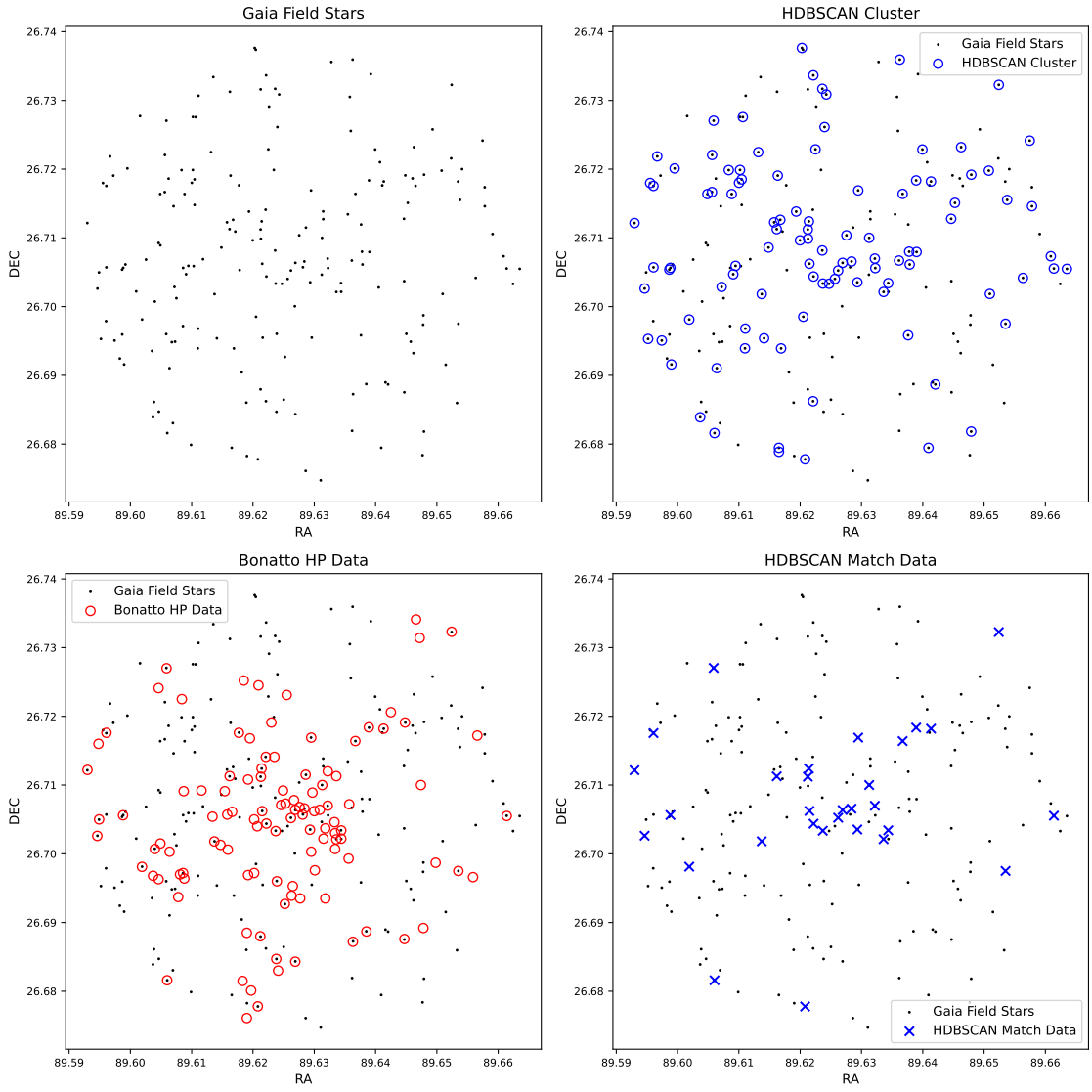


Figure 7.11 Candidate 15 - HDBSCAN-Bonatto Cross-match.

hdbscanParams	σ_μ	HDBSCANStars	BonattoHPStars	Matches	%matchToBonatto
15_mcs_20_ms_2_csm_eom	0.14	22	112	7	6
15_mcs_20_ms_4_csm_eom	0.10	21	112	7	6
15_mcs_50_ms_10_csm_eom	0.26	52	112	19	17
15_mcs_100_ms_5_csm_eom	0.43	101	112	30	27

Table 7.3 Candidate 15 - HDBSCAN-Bonatto Cross-match Statistics

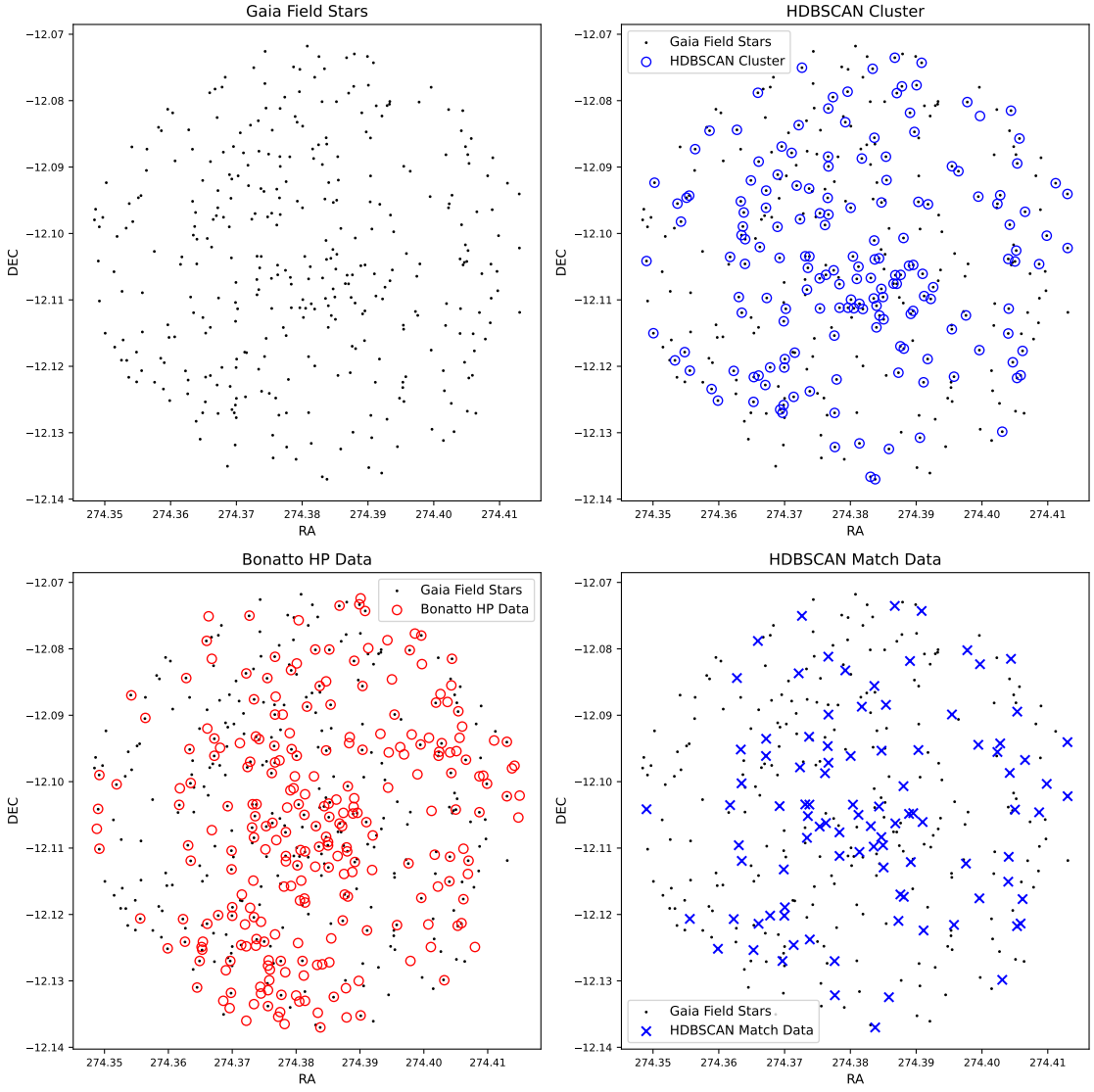


Figure 7.12 Candidate 29 - HDBSCAN-Bonatto Cross-match. In this plot we can see that HDBSCAN has produced a cluster that is visually similar to the Bonatto one. Bonatto has also selected a number of members that do not feature in the Gaia data (the empty red circles). Despite this, the cross match has highlighted the central over-density and lends support to the assertion that this is a real cluster.

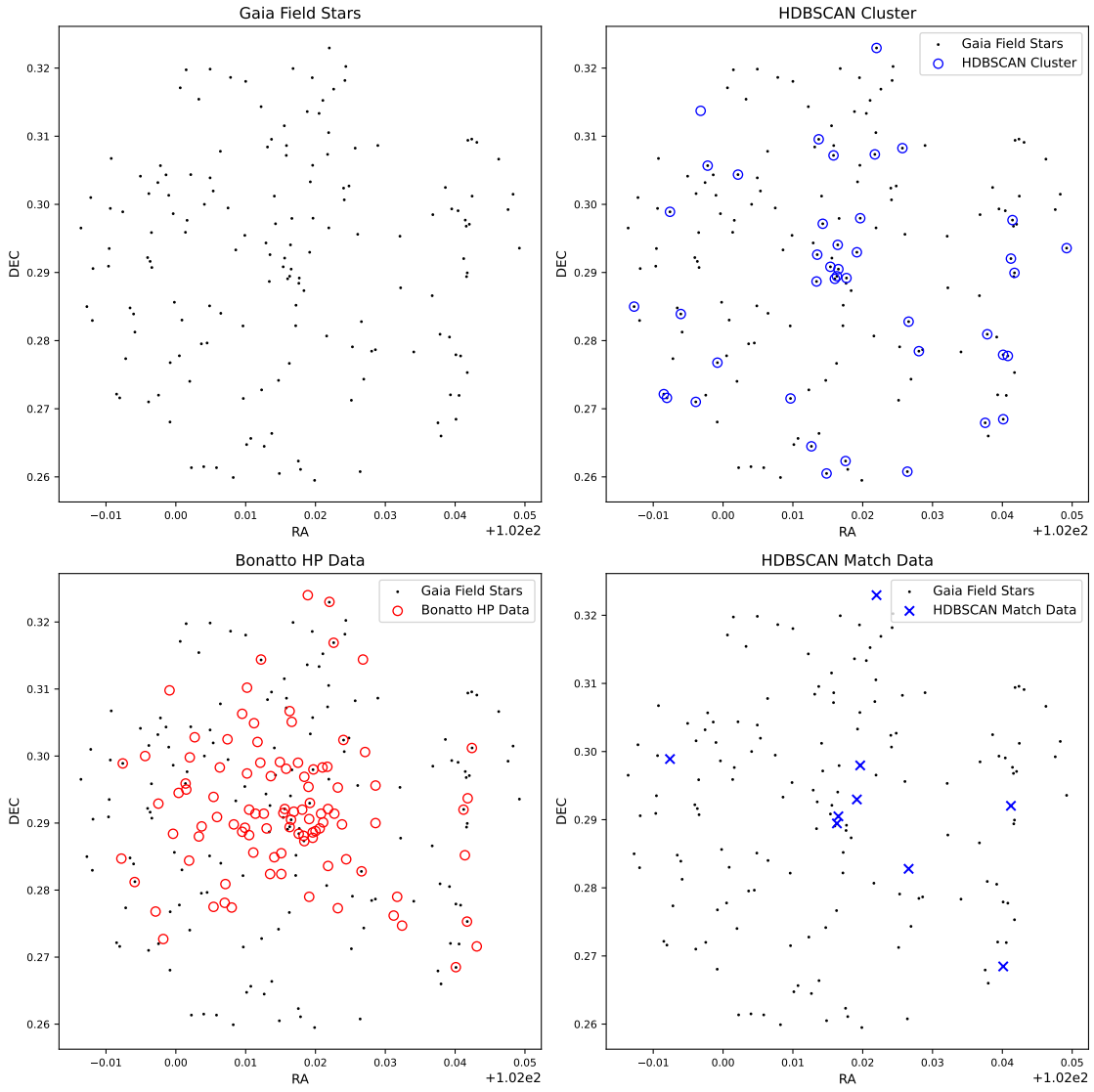


Figure 7.13 Candidate 86 - HDBSCAN-Bonatto Cross-match. A smaller number of stars in the Gaia field, Bonatto has selected a cluster with central over-density and while HDBSCAN has a smaller data set to work and this has resulted in a low cross-match percentage, it has also selected the central region and again would support the assertion that this is a real cluster.

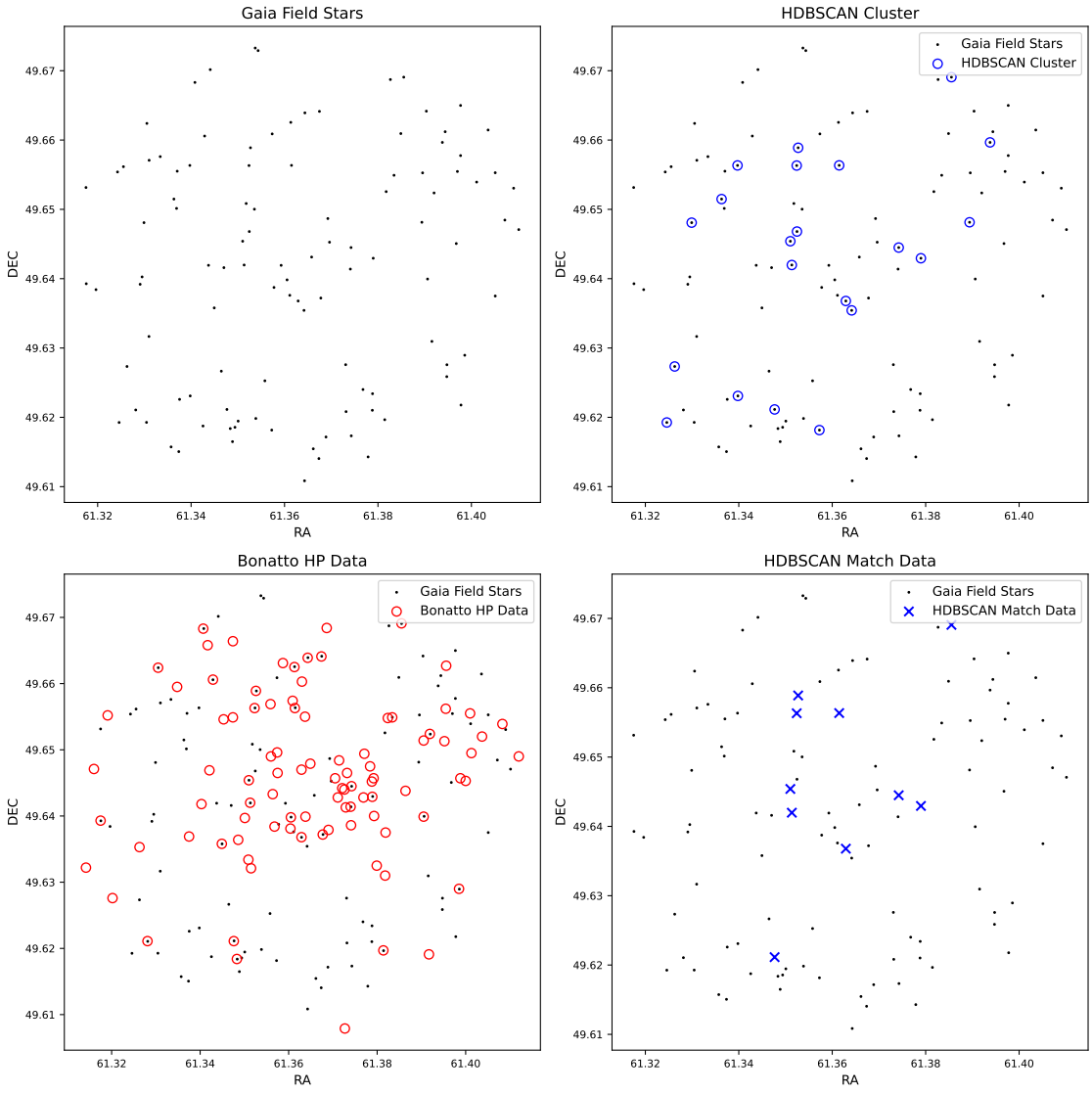


Figure 7.14 Candidate 105 - HDBSCAN-Bonatto Cross-match. Bonatto has selected the central over-density and a couple of surrounding groups although again the Gaia data is limited in this field. HDBSCAN has failed to produce a reasonable cluster and in this case is unable to provide any support to the reality of the candidate.

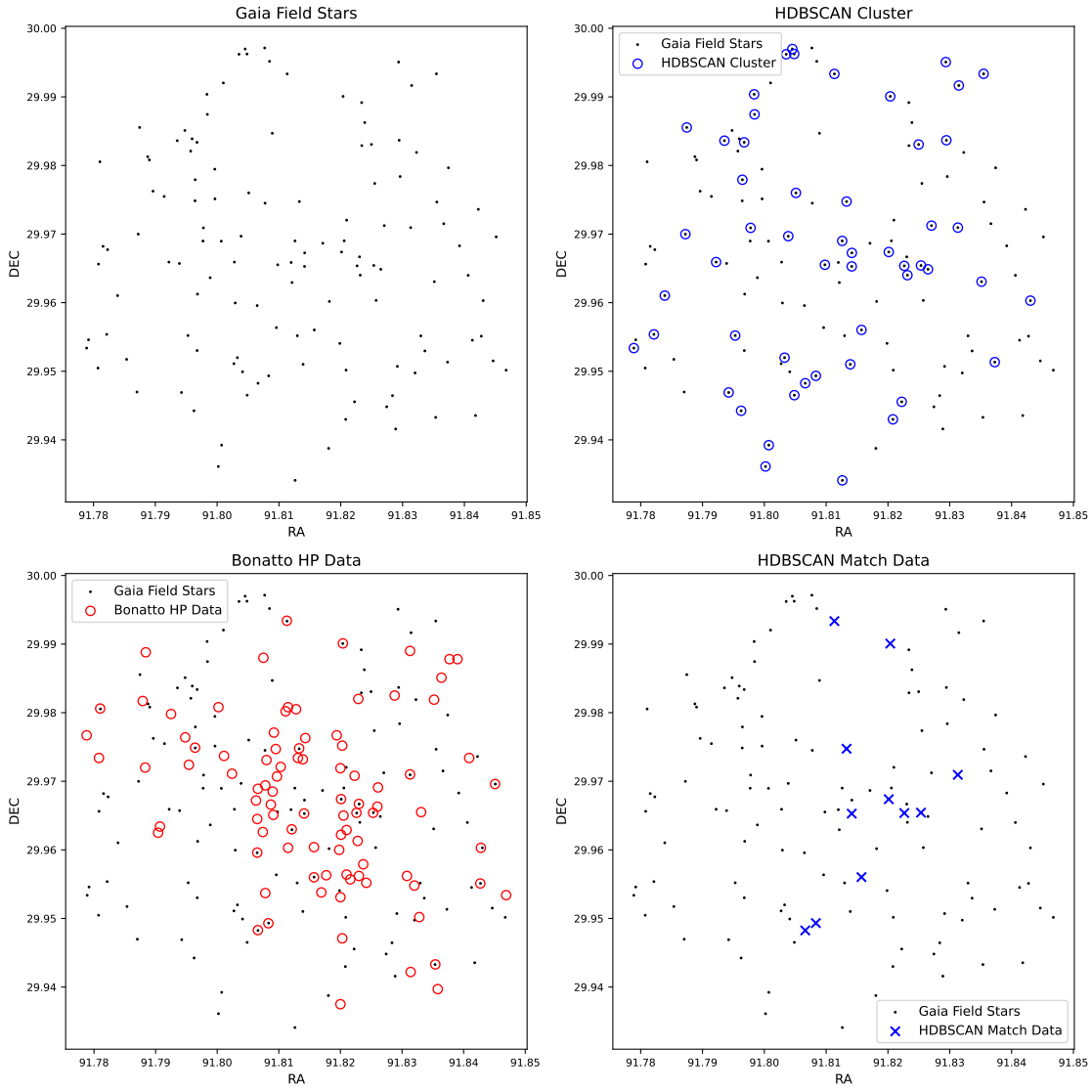


Figure 7.15 Candidate 132 - HDBSCAN-Bonatto Cross-match. Again we find Bonatto selecting a central over-density with many members not appearing in the Gaia data. HDBSCAN has produced a cluster but this has limited cross-match and the morphology is different to the Bonatto Cluster. In this case it does not bring much support to the argument of whether this candidate is real.

7.3.3 Comparison of HDBSCAN and Bonatto CMDs

Candidate 30 - Open Cluster

Candidate 30 is previously known Open Cluster - Teutsch-58 (DSH_J0606.6+1557, FSR_926, SAI_60)

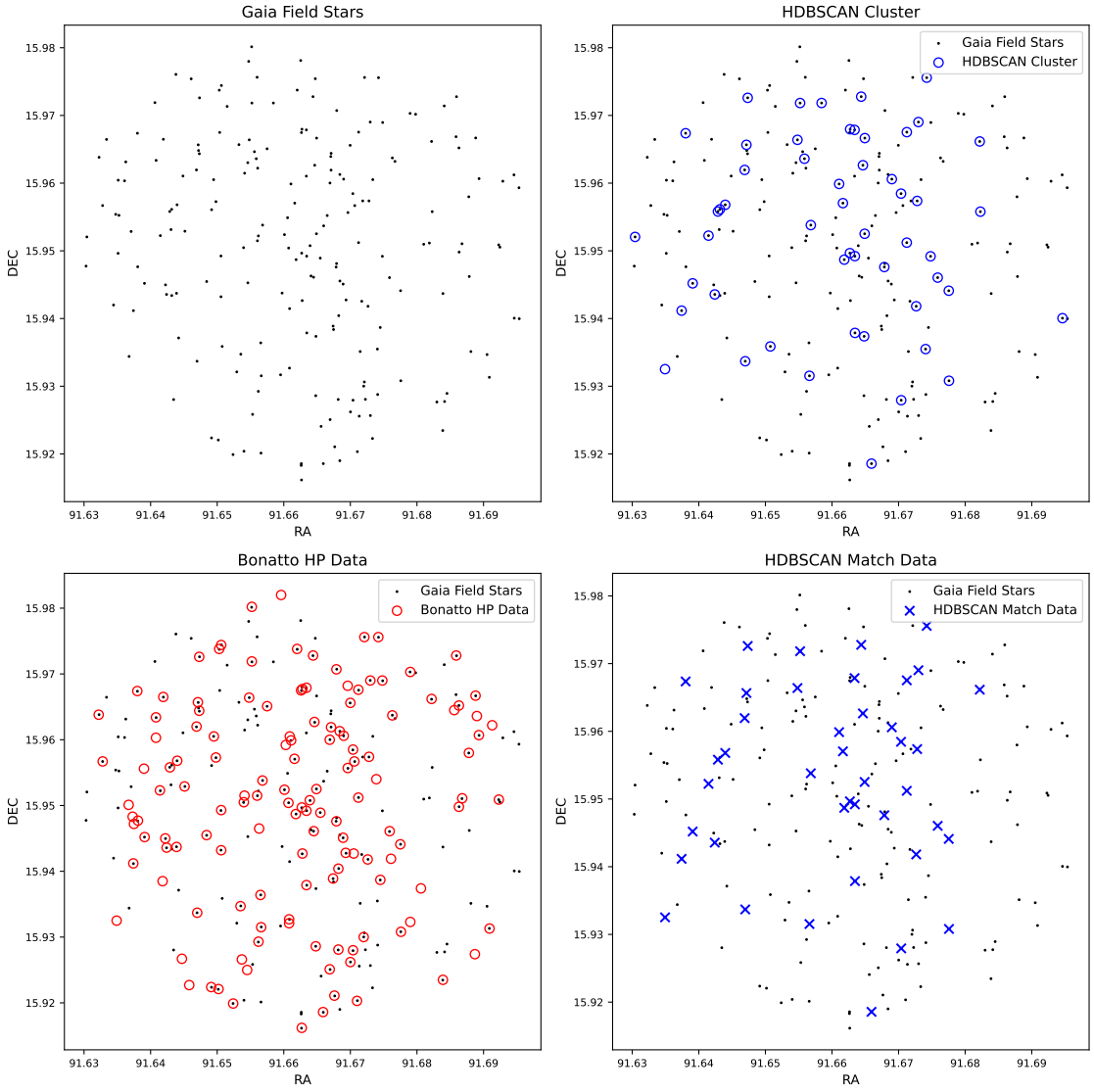


Figure 7.16 Candidate 30 - HDBSCAN-Bonatto Cross-match. In this case there is a 30% correlation between Bonatto and HDBSCAN selected members.

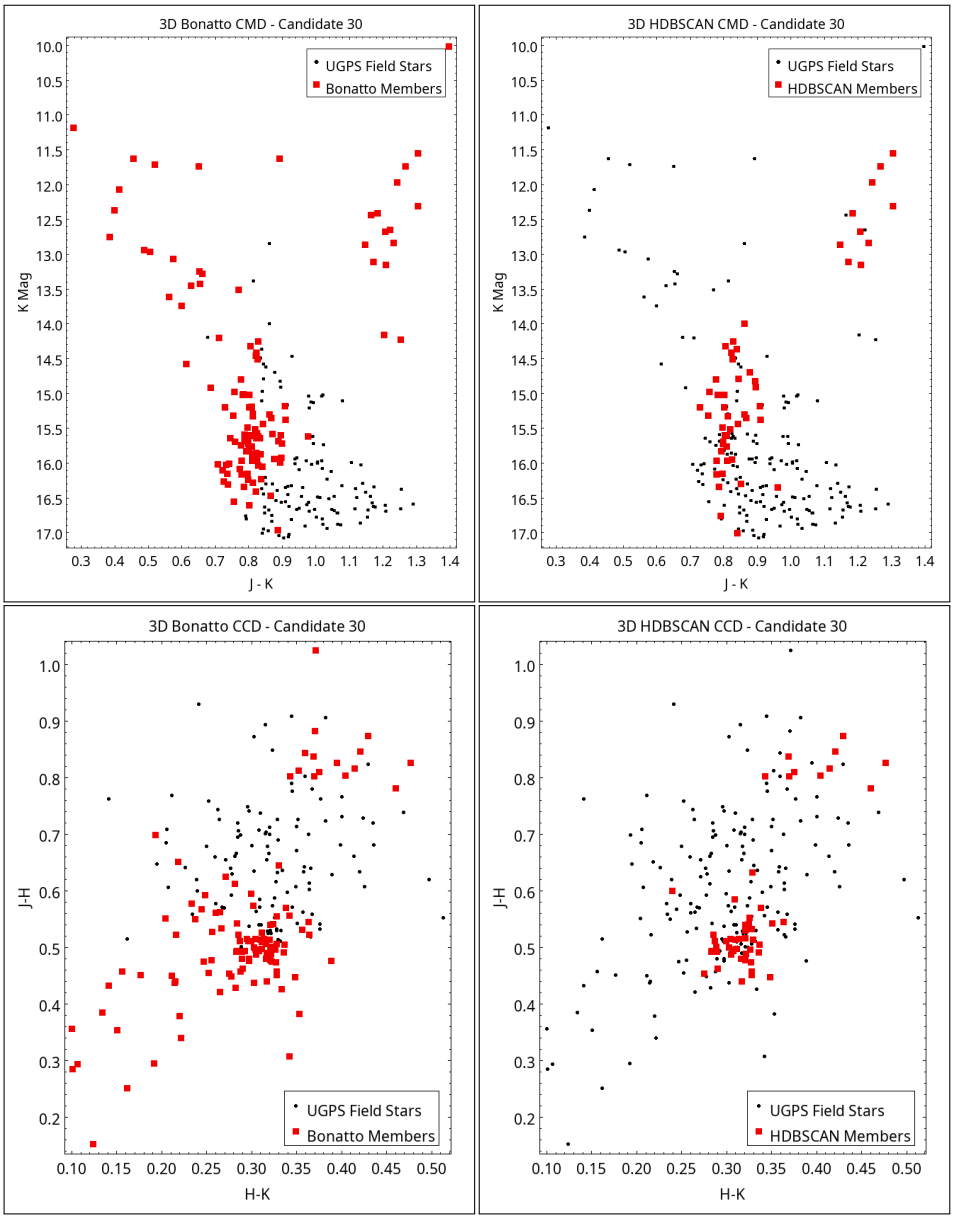


Figure 7.17 Candidate 30 CMDs and CCDs - The upper left-hand tile shows the Bonatto 3D K vs J-K CMD, whilst the upper right-hand tile shows the equivalent HDBSCAN CMD. The Bonatto plot shows a main sequence with turn-off, Red Giant Branch (RGB) and a Red Giant Clump (RGC). However, it appears that although HDBSCAN has selected members that appear to be very closely related to the Bonatto group, it has not selected members in the brighter section of the main sequence or the RGB and has omitted a clump of members in the fainter part of the MS. The HDBSCAN selection however makes a significant match to Bonatto and would certainly support a cluster hypothesis. The lower left-hand tile shows the Bonatto 3D CCD, whilst the lower right-hand tile shows the equivalent HDBSCAN CCD. In these plots we can see that HDBSCAN has selected two groups similarly to the Bonatto code, a main-sequence clump and a reddened clump. It is clear in this plot that it has not selected the less-reddened main-sequence stars and whilst follow up observation would be required to confirm membership of each member, the fact that there is significant overlap between the plots would certainly support the cluster hypothesis in this case.

Candidate 5 - Misidentified Emerging Cluster Candidate

Candidate 5 is currently misidentified as galaxy AFGL 5142, also known as IRAS 05274+3345. The UGPS JHK-band images show a clear, small emerging cluster with the majority of members only visible at longer IR wavelengths.

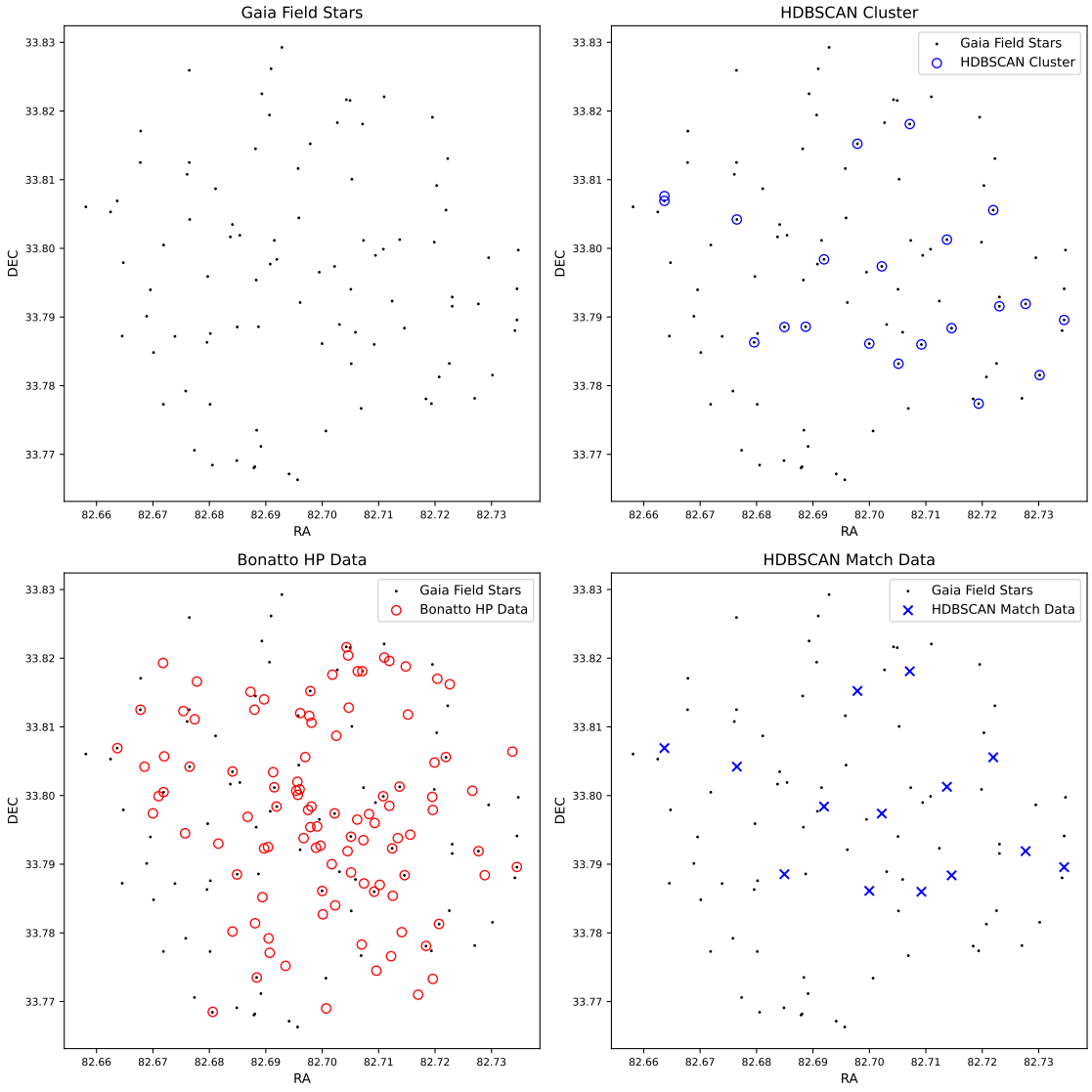


Figure 7.18 Candidate 5 - HDBSCAN-Bonatto Cross-match. In this case there was a 14% correlation between Bonatto and HDBSCAN selected members.

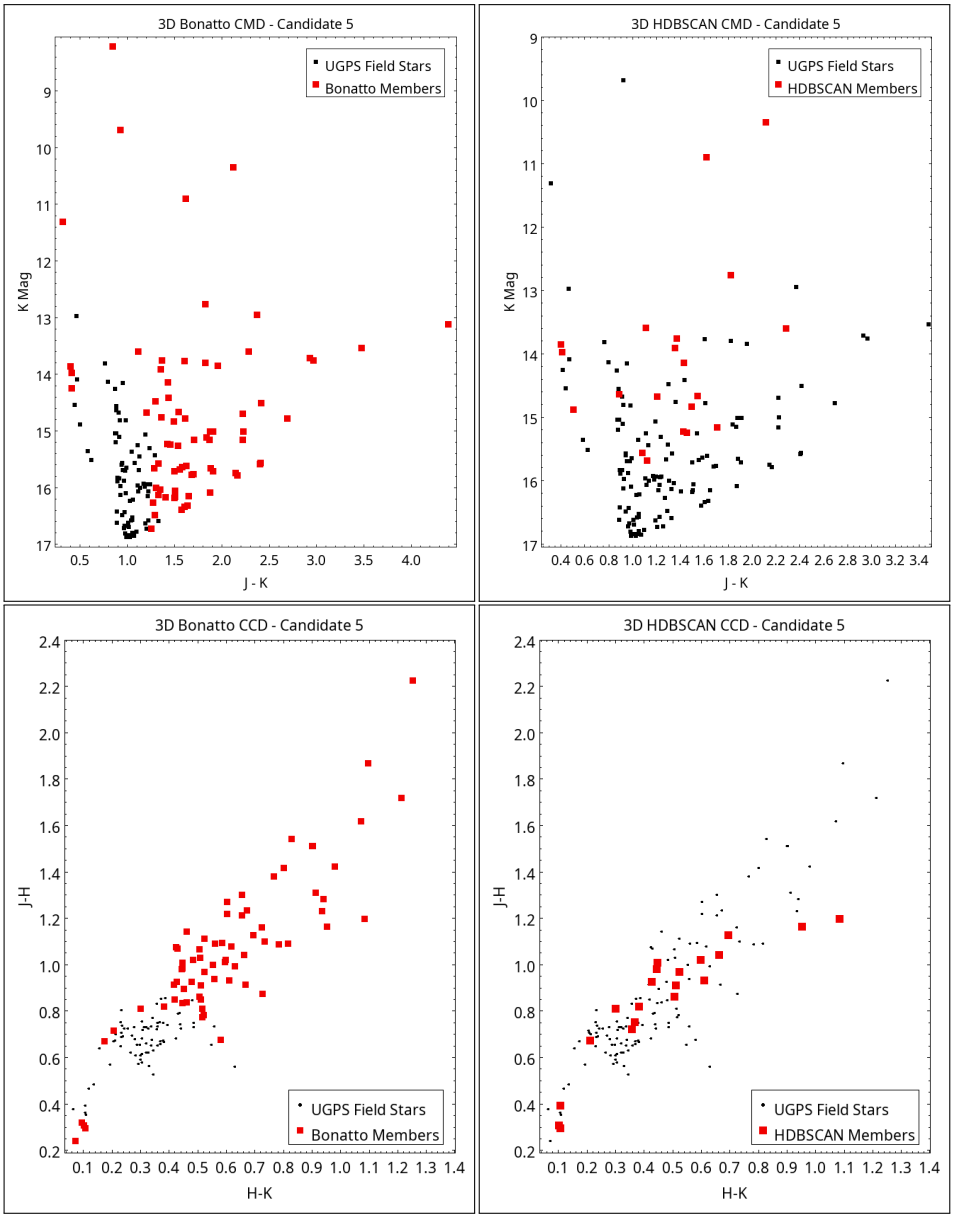


Figure 7.19 Candidate 5 - The upper left-hand tile shows the Bonatto 3D K vs J-K CMD, whilst the upper right-hand tile shows the equivalent HDBSCAN CMD. The Bonatto algorithm has selected principally reddened stars, although with a few brighter stars included. HDBSCAN by contrast has selected a small number of stars of variable brightness and extinction, the selection does offer any support for a cluster hypothesis in this case. The lower left-hand tile shows the Bonatto 3D CCD, whilst the lower right-hand tile shows the equivalent HDBSCAN CCD. The CCDs show a similar pattern to the CMDs, with Bonatto selections being primarily reddened with a tiny selection of bright stars whilst HDBSCAN has selected far fewer stars across the reddened range with no grouping. Given this cluster is an Emerging Cluster Candidate, this is the expected result as Gaia is unable to detect cluster members only visible in the infrared.

Candidate 108 - Open Cluster Candidate

Candidate 108 is a previously identified Open Cluster Candidate [FSR2007] 1171.

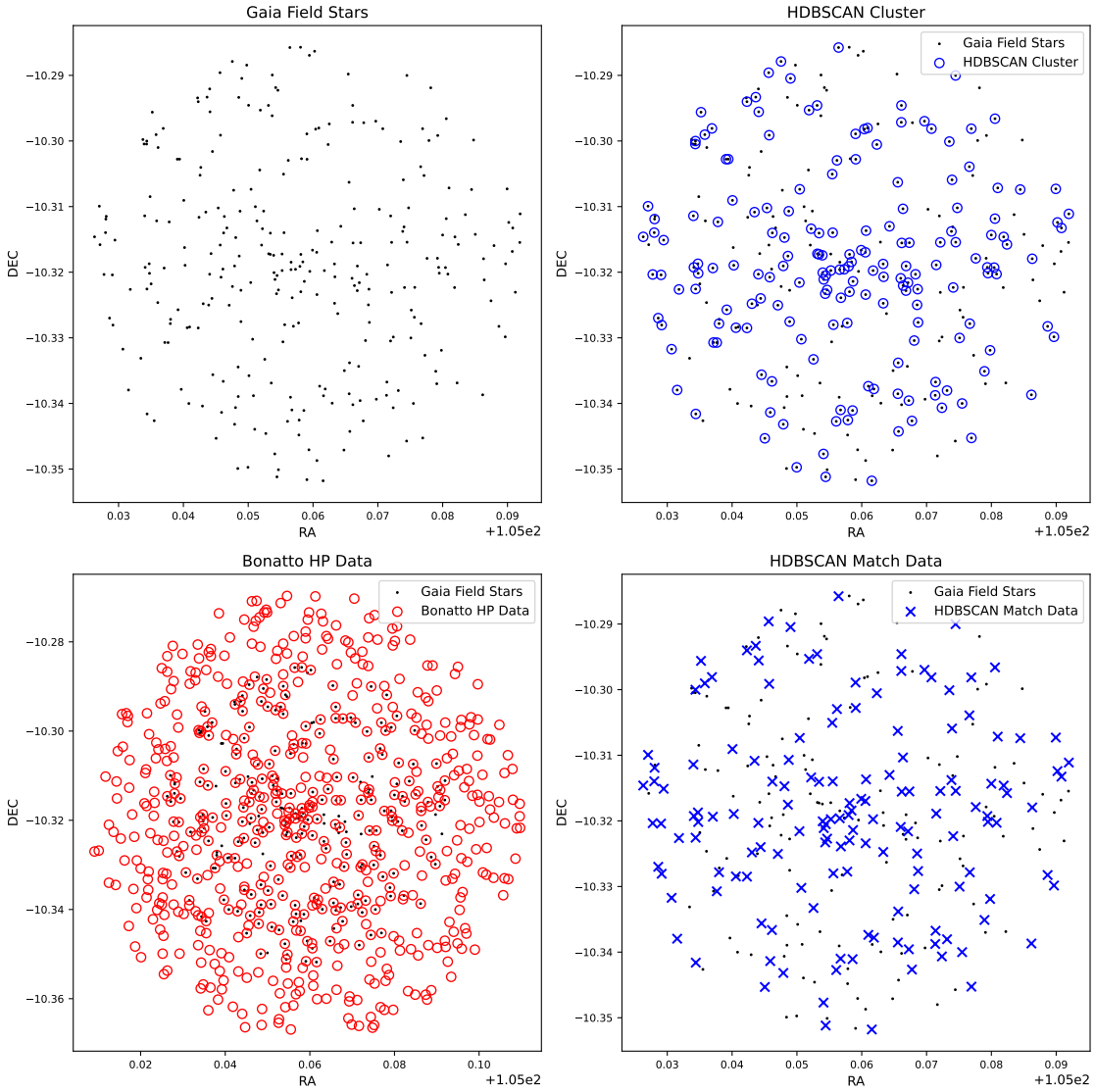


Figure 7.20 Candidate 108 - HDBSCAN-Bonatto Cross-match. There was a 24% correlation between Bonatto and HDBSCAN selected members.

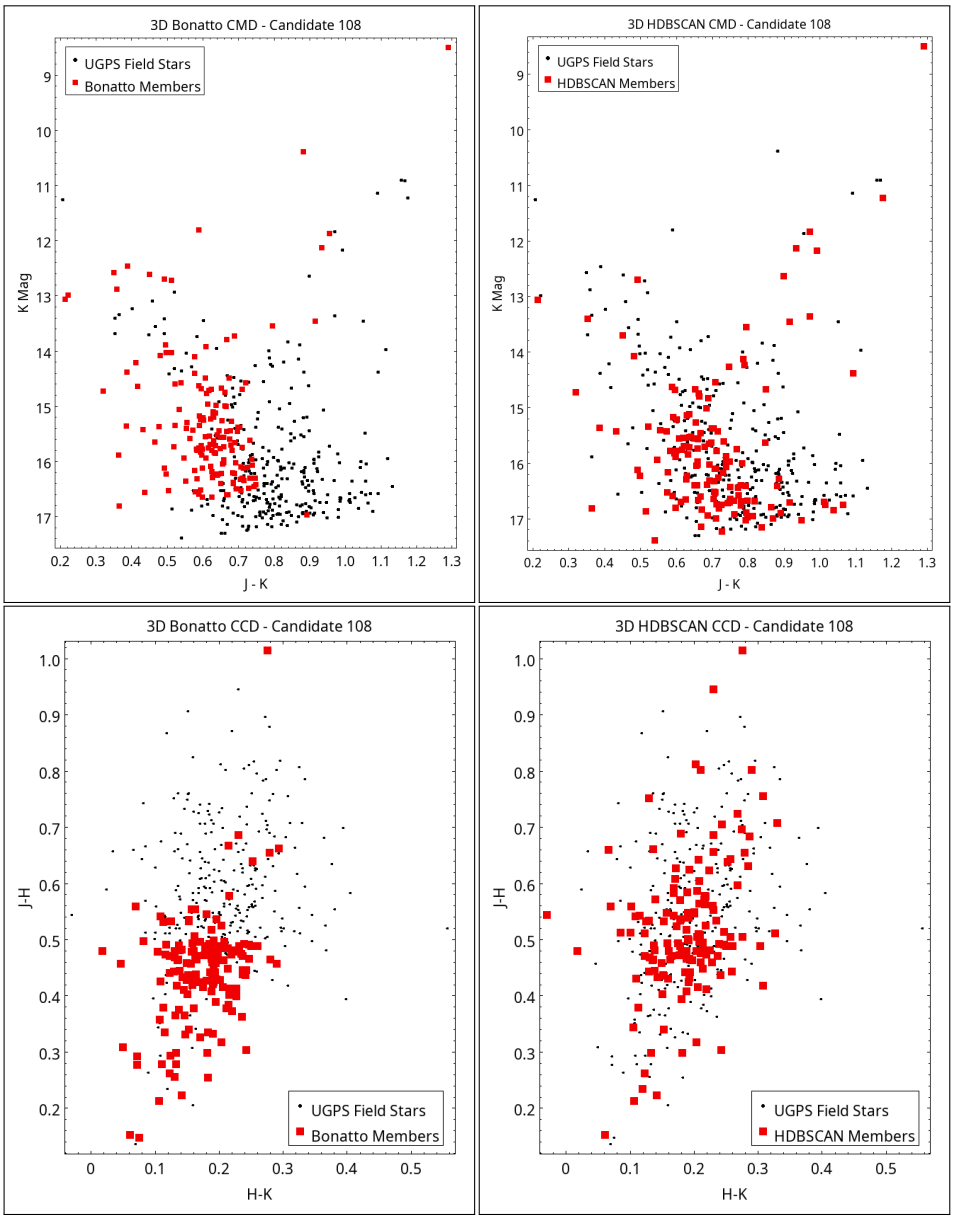


Figure 7.21 Candidate 108 - The upper left-hand tile shows the Bonatto 3D K vs J-K CMD, whilst the upper right-hand tile shows the equivalent HDBSCAN CMD. In this case the HDBSCAN selection tends to be at lower brightness and includes more members from the RGC region. The lower left-hand tile shows the Bonatto 3D CCD, whilst the lower right-hand tile shows the equivalent HDBSCAN CCD. In colour space the Bonatto algorithm members are in a tighter group and in the MS region of the diagram vs the HDBSCAN members which are more widely distributed across colour space. In this instance Bonatto supports the open cluster hypothesis but it would appear that HDBSCAN has struggled with the lack of detections vs Bonatto resulting in an inconclusive result which would not be able to support a cluster hypothesis.

7.4 HDBSCAN Summary

Given a suitable range of input parameters, HDBSCAN can produce useful cluster selections as shown in the preceding figures of this section. It is however important to appreciate that with the limited data available in Gaia for faint, infrared clusters it can be less reliable in terms of cluster support for emerging clusters rather than something like the Bonatto algorithm which has a significantly larger, NIR dataset in the WSA. As established in Section 7.1, these cross-matches serve strictly as secondary kinematic corroboration where optical depth permits, eliminating any circular dependence on spatial clustering for my primary discovery claims.

A number of studies have been made with HDBSCAN and other clustering algorithms across the whole Gaia DR2 and DR3 datasets with great success for brighter, open clusters (see Hunt and Reffert 2023, Chi et al. 2023 & Cánovas et al. 2019). In all cases HDBSCAN was found to be the most sensitive algorithm as it is able to detect clusters of varying density and handle noise. I also found it valuable in adding a degree of confidence to the cluster hypothesis in those cases where the cluster can be clearly distinguished from the field, as in the case of open clusters, and it is able to produce CMDs and CCDs that closely resemble those produced by the Bonatto Algorithm. Indeed, the similarities between the results suggests that given an equivalent data set HDBSCAN might be able to select the same members as the Bonatto algorithm, but using independent parameters. If this can be substantiated it would make a valuable complementary method for candidate verification. This will sadly have to wait for future instruments such as the Nancy Roman Telescope and ESA's Extremely Large Telescope which should be able to produce proper motion data in the near infrared with precision near to that of Gaia, (see Sosey et al. 2021 and Matteo et al. 2019 for details).

Visual inspection of a large number of HDBSCAN plots has indicated that a match of 25% or greater of the HDBSCAN members to the UGPS Bonatto selected membership is most likely to correspond to a supporting match. The figure is relatively low because of the differences in depth and bandpass between Gaia and UGPS, however 66 candidates did pass this test. As expected, the brighter, less reddened Open candidates and clusters

dominate the matches. A full discussion and breakdown of the result is given in chapter 8.

I have included the results of the HDBSCAN analysis of each candidate in appendix A, columns *hdbscan_match* & *hdbscan_real*. In those cases where there are sufficient matches, and the HDBSCAN cluster has a central concentration, it does provide an additional strand of evidence to support the cluster hypothesis. In this particular application of the code it is being used to investigate how well it matches to the Bonatto codes member selection using an independent method. It can therefore be assessed on the basis of the percentage cross-match for the purposes of this work.

Chapter 8

Summary, Conclusions and Further Work

As described in the Introduction, the primary goal of this work is to utilise the high spatial resolution and sensitivity of the UKIRT-WFCAM to extend the sample of known young, compact, Galactic star clusters in a parameter space where previous surveys such as VISTA VVV/x, 2MASS and Spitzer GLIMPSE were unable to provide this data due to their limited resolution or depth. The same limitation applies to modern optical surveys such as the Gaia mission: because WFCAM operates in the near infrared, it retains significantly better sensitivity for young, embedded clusters and for distant clusters with high extinction. As a result this work should be able to identify more visually compact clusters, either at an early stage of evolution or more distant than before, as well as being able to identify compact regions of star formation within dark clouds and filaments, which are typically difficult to resolve. A lack of small, compact objects would itself be an interesting result: it could indicate, for instance, that small clusters have an unusually high rate of 'infant mortality' as described by Lada and Lada (2003), where small young clusters are rapidly disrupted by dynamic interaction and the rapid gas dispersal by feedback from the young stars.

8.1 Summary

To summarise the aims of the project:

1. Identify and provide basic characterisation for new star clusters and cluster candidates found during the quality control checking of the UKIDSS GPS DR10 data release.

2. Determine whether the UKIDSS GPS can be used to extend the sample of young, compact clusters into new parameter space with its greater resolution and depth.

3. Use data from ESA's Gaia DR3 release to investigate the proper motion and distances of both the existing and new / candidate clusters identified in 1 above.

4. Investigate the usefulness of a published, unsupervised machine learning technique to provide additional support to the cluster hypothesis for the new / candidate clusters where possible.

5. Provide confirmation of previously identified candidate clusters where possible.

The visual search of the UGPS data described in chapter 3 yielded 403 candidate objects. These were then analysed to determine a set of basic metrics, identify previously known objects and remove false clusters.

To achieve this the first stage of characterisation is to remove the field stars from the cluster area. This is a difficult task, as described in chapter 3, and leaves us with a selection of high-probability members which are then used to plot colour-magnitude and colour-colour diagrams (CMD and CCD respectively) as described in chapter 1. The CMD and CCDs give a view of the members as a population vs the field in which they sit and this provides information on the degree of reddening and the age of the cluster. A CMD with a clear main sequence displaying features such as main-sequence turn-off, Red Giant Branch, Red Giant Clump would suggest support for an evolved cluster hypothesis, whilst a CMD without a turn-off but a population of reddened proto-stars still approaching the main-sequence would provide strong support to the young cluster

hypothesis. A CCD showing a tight group of members at the lower left would support a more evolved open cluster, whilst members distributed along the reddening vector would indicate variable extinction and with proto-stars lying below the reddened main-sequence curve this would support a much younger, shrouded cluster.

Each candidate's CCD and CMDs were assessed with reference to the visual appearance of the candidate in the J, H and K band images, and any proximal star formation markers, and this information was used to assign the classification (as given in Table 1.1). In some cases the plots will contain groups of false members that are selected just by virtue of their colour being similar enough to the cluster population. These are usually easy to spot and remove or ignore from the plots. Detailed individual plotting and grouping helps to identify and remove these. There are also instances where a visually compelling image yields unconvincing CMD and CCDs. This can happen for valid clusters where the members lie below the detection threshold or otherwise fall outside the ability of the WSA process to render them. These objects are still listed but will have to wait for the next generation of instruments to investigate. In some cases, despite the initial visual suspicion of a cluster, none of the data from any sources supports the cluster hypothesis and these are marked accordingly. The overall results are summarised in Table 8.1.

Table 8.1 Breakdown of Results by Class and Status

Class	Total	Known	Known Candidate	New
Open Clusters	34	34	2	0
Open Cluster Candidates	136	57	33	79
Potential Open Clusters	6	4	1	2
Emerging Clusters	22	22	2	0
Emerging Cluster Candidates	79	48	32	31
Potential Emerging Candidates	14	3	3	11
Embedded Groups	7	1	1	6
No Cluster	105	12	8	93

For each of these candidates several physical properties were determined, as described in chapter 5,

- Algorithmically determine central coordinates
- Plot radial density profile and determine core and tidal radii
- Estimate angular diameter
- Determine number of decontaminated members
- Identify star formation markers within 10 arcmin
- Record associated J, H or K-band nebulosity

Each candidate was also cross-matched to the complementary data available from the WISE AllWISE and Gaia DR3 catalogues (see chapters 5 and 6). WISE provides a picture of the warm material, HII regions and other star formation markers in the vicinity of the candidate, and this was valuable when assessing whether a candidate fits with an Open, more evolved, or Emerging hypothesis. Suspected younger clusters show a clump or filament of warm material, and perhaps an HII region, whilst a more evolved open cluster would have cleared the majority of the progenitor material.

To bridge the gap between initial visual classifications and rigorous quantitative confirmation, Chapter 4 introduced an automated, multi-wavelength 2D MIST isochrone fitting pipeline. Because near-infrared colour-magnitude and colour-colour diagrams (K vs $J-K$ and $J-H$ vs $H-K$) often suffer from severe line-of-sight degeneracy—where extinction vector orientation can mimic age spreads—my pipeline introduced three structural innovations:

1. 3D Dust Map Anchoring: Distance moduli were constrained against the line-of-sight dust distribution from Bayestar19 (Green et al. 2019), preventing unphysical distance floating.

2. Photometric Q-Parameter Diagnostic: The reddening-free Q-index (Negueruela et al. 2007; Borissova, Bonatto, et al. 2011) was quantified across all members to identify intrinsic infrared excess ($Q < -0.05$ mag) indicative of circumstellar discs and Young Stellar Objects (YSOs).

3. Multi-Wavelength Ground-Truth Vetoes: Because extreme differential reddening

can artificially scatter background field stars across the Q-parameter threshold, photometric classifications were strictly cross-referenced against external multi-wavelength markers (WISE diffuse nebulosity, HII regions, and SIMBAD YSOs). Where high Q-parameters lacked any external mid-infrared emission, the external data acted as a conservative veto, reclassifying the system as an evolved open cluster ($\log t > 8.0$). Conversely, explicit WISE HII emission flagged a cluster as young ($\log t \leq 8.0$).

To guard against a known weakness of statistical field-star decontamination—that the subtraction can leave an apparent sequence behind in the regions where the field population is sparsest, even when the underlying distribution is uniform—my pipeline enforced a strict two-layered quality control framework. Candidates were evaluated against both quantitative criteria ($N_{\rm stars} \geq 10$ and reduced $\chi^2_\nu \leq 5000$) and a mandatory manual visual review of the physical sky imagery (J, H, K multiframes and spatial density overplots). The value of this safeguard is shown by my negative controls: 20 candidates (such as clus170 and clus378) satisfied quantitative richness thresholds or converged on low χ^2_ν values along differential reddening dust lanes, but lacked physical cluster cores upon visual review. These 20 cases were classified as Tier F artefacts and excluded from the final catalogue.

Applying this rigorous quantitative pipeline across the 198 evaluated new discovery candidates yielded a finalised catalogue of 178 High-Confidence Tier A/B Star Cluster Discoveries (89.9% of the evaluated sample). Classified on the combined evidence of the isochrone age, the reddening-free Q-parameter diagnostic and external mid-infrared star-formation markers, with $\log t = 8.0$ adopted as the nominal boundary, this high-confidence sample comprises:

- 154 Young / Emerging Cluster Discoveries: Systems retaining pre-main-sequence sequences, elevated YSO fractions, or mid-infrared emission markers, with fitted ages $\log t \leq 8.0$ (median $\sim$ 66 Myr).

- 24 Evolved Open Cluster Discoveries: Cleared stellar sequences exhibiting well-defined main-sequence turnoffs and an absence of star-formation markers,

with fitted ages $\log t > 8.0$.

Furthermore, the quantitative pipeline successfully evaluated and provided confirmation for 74 of the 82 previously identified candidate clusters from the literature (90.2% of the known candidate sample). My two-layered analysis demonstrated that 76 candidates originally classified as 'No Cluster' (NC) during early visual screening due to overwhelming background field noise were revealed upon Bonatto decontamination and 2D MIST fitting to possess compact, statistically significant stellar sequences ($N_{\rm stars} \geq 10, \chi^2_\nu \leq 5000$), elevating them directly into my Tier A/B High-Confidence catalogue.

It should be noted that the age is the least well constrained of these parameters. As set out in Section 4.2.8, the residual field population retained by statistical decontamination, combined with the near-degeneracy between age, distance and extinction in the near-infrared, limits the precision attainable on individual fits. The young/evolved classification is therefore supported by the Q-parameter diagnostic and independent mid-infrared markers, rather than resting on the isochrone age alone.

One of the hypotheses stated in the introduction was that UGPS should be able to resolve young, compact clusters that are beyond the reach of existing instruments. The synthesis of my final derived parameters confirms this. My 178 newly discovered candidates are, on average, 13% more compact (mean half-maximum radius 2.38 ± 0.9 arcmin) with a significantly narrower size distribution than previously known clusters (2.73 ± 2.0 arcmin). Moreover, my discoveries extend deep into obscured parameter space, exhibiting a median visual extinction of $A_V = 5.11$ mag, with extreme systems exceeding $A_V > 10$ mag. There are also 13 POC and PECs that are too compact and faint to be fully resolved but are clearly dense concentrations of stars, dust and gas.

The crossmatching to Gaia's DR3 data provided important proper motion and photogeometric distance data for both known and unknown candidates, where possible. However, linking directly back to Project Aim 3, optical astrometry demonstrated clear physical limitations when applied to embedded near-infrared clusters. Because optical G-band flux is heavily attenuated by dust ($A_G \approx 0.85 A_V$), an extinction of $A_V \gtrsim 10$

pushes cluster members below Gaia's detection threshold. Consequently, while Gaia cross-matching succeeded for evolved open clusters and bright foreground stars, it achieved low match rates ($\leq 25\%$) for deeply embedded objects. In all cases a median photogeometric distance was calculated and, where possible, compared to distance estimates obtained independently. Of the previously unknown, real candidates there are 9 with distance estimates available from WISE HII data. These are kinematic distances, derived from the radial velocities of the associated HII regions together with a model of Galactic rotation, and are therefore independent of Gaia parallax: the precision limit of $\sim$ 3–4 kpc discussed in Section 6.2 constrains the astrometric distances but does not apply to these. I can now compare these WISE kinematic distances directly against the photometrically derived distances from my 2D MIST isochrone fits ($d = 10^{(\mathrm{DM}/5+1)}$). In eight of the nine cases the isochrone distances fall systematically short of the WISE kinematic values. This offset is expected: distance modulus and extinction are near-degenerate in the near-infrared, and because statistical decontamination inevitably retains a residual population of foreground interlopers—which are both nearer and less reddened than genuine members—the fit is drawn toward the foreground, compensating with elevated A_V. The Bayestar19 prior cannot recover the difference, since it traces diffuse interstellar material rather than the natal cloud that dominates the extinction toward embedded clusters (Section 4.2.8). Notably, the same foreground bias affects the Gaia cross-match, where the WISE kinematic distances exceed the Gaia estimates in eight of the nine cases. That two independent techniques fail in the same direction, for the same physical reason, reinforces rather than undermines the conclusion that these are genuinely distant systems. There is one emerging candidate for which the Gaia and WISE distances agree at 1.5 kpc which illustrates the fact that Gaia struggles with fainter, extincted candidates. These findings are summarised in Table 8.2.

The proper motion distribution was calculated for each candidate and compared to the distribution expected at the given median distance. Of the known and new open candidates (OC and OCC) 22 out of 170 were able to meet this criterion, with 10 of the 136 OCCs meeting it. Of the EC and ECCs 18 met the criterion, with 14 being ECCs.

Table 8.2 Unknown Candidates with WISE Distances

Candidate Id	Class	Isochrone Dist. Est. (kpc)	WISE Dist. Est. (kpc)	WISE Dist. Method	Gaia Dist. Est. (kpc)
2	ECC	3.1	8.3	Kin	2.4
4	ECC	0.8	7.7	Kin	4
59	PEC	1.2	7.7	Kin	3.1
61	PEC	1.9	7.9	Kin	2.7
150	ECC	3.5	7.9	Kin	3.5
288	EGr	1.5	9.2	Kin	2.7
349	PEC	2.4	9.9	Kin	2.8
360	OCC	8.2	12.6	Kin	4
392	ECC	1.6	1.5	Parallax	1.5

An additional test was made using the unsupervised machine learning algorithm HDBSCAN as described in chapter 7. This provides an independent method for resolving clusters based on the vectorisation of the Gaia stellar sky coordinates, proper motion and photogeometric distance. Visual inspection of a large number of HDBSCAN plots has indicated that a match of 25% or greater of the HDBSCAN members to the UGPS Bonatto selected membership is most likely to correspond to a supporting match. The figure is relatively low because of the differences in depth and bandpass between Gaia and UGPS, however 66 candidates did pass this test. Of these, 37 were known clusters (9 of which are candidates) with 1 EC, 3 ECCs, 14 OCs, 14 OCCs, 2 PECs and 3 POCs. The 29 previously unknown matches included 3 ECCs, 24 OCCs, 1 PEC and one POC. As expected, the brighter, less reddened Open candidates and clusters dominate the matches. This can be a powerful technique and a valuable alternative method for assigning cluster membership once we have higher precision proper motion infrared data. Table 8.3 summarises the results for successful tests of both the Gaia proper motion and the HDBSCAN to Bonatto cross-match (it excludes emerging groups as these are not expected to - and did not - pass these tests).

Table 8.3 Summary of Successful Gaia and HDBSCAN Test Results

Class	Gaia Proper Motion Dispersion	HDBSCAN Cross-Match
Open Clusters	12	14
Open Cluster Candidates	10	38
Potential Open Clusters	1	4
Emerging Clusters	4	1
Emerging Cluster Candidates	14	6
Potential Emerging Candidates	1	3

The literature search was made against a number of curated databases as described in §5.4. Both object references and number of citations were taken into account when assessing known status, with one of the objectives of this work to help confirm or question current candidate clusters. Out of the 181 known clusters, 82 are candidates, and of those, 74 appear real and confirmed in my analysis.

8.1.1 Sensitivity of the 'No Cluster' Fraction to the Applied Tests

Of the 403 candidate overdensities put forward by the visual search, 105 (26%) were classified as 'No Cluster' (NC) during the initial visual screening — 93 of the new candidates and 12 previously catalogued objects. This figure, however, rests on a single method: visual inspection of crowded Galactic plane fields. Table 8.4 shows how the same 105 objects fare under each of the independent tests applied in this thesis, and demonstrates that the rejected fraction is highly sensitive to the test used. The kinematic entries are included for comparison only: as established in Chapters 6 and 7, HDBSCAN and the Gaia proper-motion test serve as secondary corroboration and are diagnostic only for the most open and evolved systems, not for the embedded and obscured candidates that dominate this subset.

Table 8.4 Number of candidates supported as clusters by each test, for the 105 objects initially screened as 'No Cluster' and for the full sample. The pipeline totals count only the new candidates it evaluated. The HDBSCAN and Gaia totals here include the NC-classified objects and therefore exceed the per-class totals of Table 8.3, which excluded them.

Test	Among the 105 NC	Among all 403
Initial visual screening	0	298
Decontamination + isochrone pipeline	76 of 93 evaluated	178 of 222 evaluated
HDBSCAN cross-match ($\geq$ 25%)	34	100
Gaia PM dispersion	1	43

Three points follow. First, the initial visual verdict has little predictive power for the quantitative outcome: the pipeline recovered 76 of the 93 new NC candidates (82%) as high-confidence clusters, a rate almost identical to its 80% recovery across all 222 new candidates. These 76 systems are precisely those whose background field contamination defeated visual inspection but yielded to statistical decontamination, and they are retained on that quantitative evidence together with the mandatory visual review of the fitted sequences. Second, the kinematic comparisons behave exactly as their limitations predict: HDBSCAN corroborates 34 of the 105 — of which 23 are among the 76 recovered discoveries, lending independent support to a subset of the recoveries — and the Gaia proper-motion test only one. These low rates reflect the optical detection limits established in Chapter 6, not evidence against the clusters, and no candidate in this thesis is accepted or rejected on either test alone. Third, combining all tests, the firmly rejected fraction falls from 26% at initial screening to 8% — the 20 Tier F artefacts together with the 12 previously catalogued objects for which no supporting evidence was found — with a further 24 candidates (6%) unresolvable because their photometry could not support a fit. The initial screen was therefore conservative, erring towards rejection, and the quantitative pipeline exists precisely to recover real clusters that visual inspection in these fields cannot distinguish from noise.

The combination of both the observed and derived parameters provides several

complementary strands of evidence which support, or refute the cluster hypothesis for a given candidate. The output of this entire process and my final classification breakdown is summarised in the comprehensive table provided in Appendix A.

8.2 Conclusion

Revisiting the primary goals established at the outset of this thesis, this project aimed to determine whether the superior angular resolution (0.8″) and near-infrared sensitivity of the UKIRT Galactic Plane Survey (UGPS) could systematically uncover young, compact, or deeply embedded star clusters in regions of the Milky Way where previous surveys (such as 2MASS and Spitzer) hit sensitivity bottlenecks. The final results of my multi-stage pipeline confirm this hypothesis across all five core aims:

1. Discovery and Characterisation of New Clusters (Aim 1): I have successfully identified, decontaminated, and derived a uniform set of physical parameters ($\log t$, DM, A_V, $N_{\rm stars}$, and $YSO_{\rm frac}$) for 178 High-Confidence Tier A/B Star Cluster Discoveries, comprising 154 young/emerging clusters and 24 evolved open clusters.

2. Extension into New Parameter Space (Aim 2): The new discoveries prove that UGPS unlocks previously inaccessible parameter space. My sample reveals exceptionally compact systems (13% smaller core radii than known clusters), deeply obscured populations at the far end of the Galactic extinction distribution ($A_V > 10$ mag, such as clus352 at $A_V = 14.99$ and clus360 at $A_V = 15.76$), and distant Galactic structure (with WISE HII kinematic distances placing 8 of my discoveries between 7.7 and 12.6 kpc, see Table 8.2).

3. Role of Optical Astrometry (Aim 3): My analysis demonstrated that while Gaia proper motions and photogeometric distances provide excellent secondary corroboration for well-resolved open clusters, optical astrometry cannot serve as the primary diagnostic for embedded infrared clusters due to severe dust attenuation and foreground interloper bias.

4. Unsupervised Machine Learning Validation (Aim 4): By applying HDBSCAN to 5D Gaia astrometric space, I established an independent clustering validation that successfully corroborated 66 candidates (including 29 previously unknown systems), demonstrating the power of non-parametric clustering when applied across multi-dimensional catalogues.

5. Confirmation of Existing Candidate Clusters (Aim 5): Out of 82 previously unconfirmed candidate clusters from the literature within my survey footprint, my quantitative decontamination and isochrone pipeline successfully confirmed 74 as real physical clusters (90.2%), while categorising the failure modes of the remainder.

In summary, by synthesising Bonatto colour-magnitude decontamination, 2D MIST isochrone optimisation, 3D dust mapping, and multi-wavelength ground-truth vetoes into a unified, critical decision-making process, this thesis has delivered a statistically robust and physically consistent catalogue of 178 new Galactic star clusters. This catalogue significantly enriches our census of Milky Way star formation and provides a prioritised target list for next-generation astronomical facilities.

8.3 Further work

While this thesis has successfully derived robust photometric parameters (DM, A_V, $\log t$, and membership counts) across all 178 high-confidence discoveries via 2D MIST isochrone fitting, the scale of the catalogue provides a rich target list that warrants high-resolution spectroscopic follow-up.

To confirm 3D kinematic dynamics, measure internal velocity dispersions, and derive absolute chemical abundances and metallicities ($[Fe/H]$), I highlight three primary tiers of follow-up candidates across distinct physical regimes:

- Rich Discoveries ($N_{\text{stars}} > 1000$): Clusters such as clus373 ($N_{\text{stars}} = 1851$), clus272 ($N_{\text{stars}} = 1264$ and clus354 ($N_{\text{stars}} = 1152$) represent the most statistically robust main-sequence and pre-main-sequence populations in the catalogue. They are

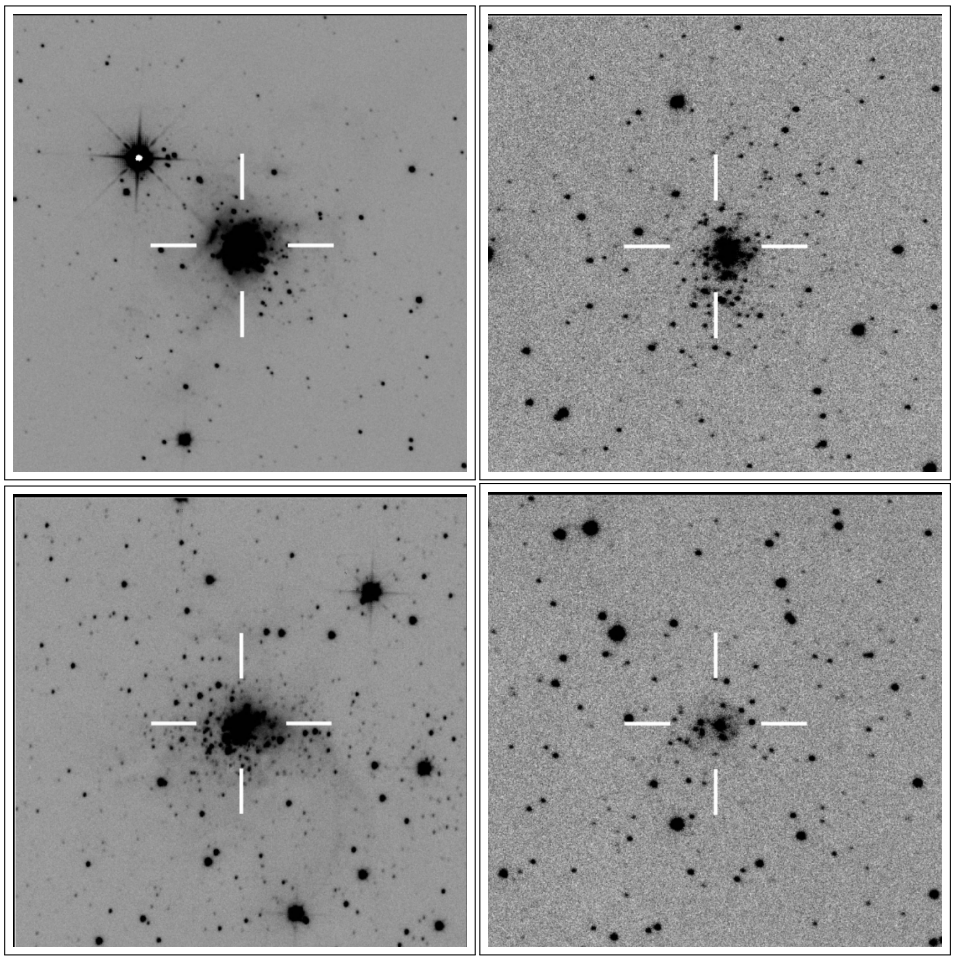


Figure 8.1 2x2 arcmin K-band images of new, distant candidates. The top left tile shows candidate 4 which has a WISE distance of 7.7 kpc. Top right tile shows candidate 61 which has a WISE distance of 7.9 kpc. The bottom left tile shows candidate 150 which has a WISE distance of 7.9 kpc. The bottom right tile shows candidate 349 which has a WISE distance of 9.9 kpc.

prime targets for deep multi-object spectroscopy (e.g., VLT/MOONS) to probe the detailed form of the Initial Mass Function (IMF) across the substellar boundary without low-number statistical noise.

- Extremely Obscured / Embedded Discoveries ($A_V > 10$): Systems such as clus352 ($A_V = 14.99$ mag) and clus360 ($A_V = 15.76$ mag) sit at the extreme edge of the Galactic Plane extinction distribution. High-resolution infrared spectroscopy (such as that enabled by JWST or upcoming ELT/MICADO instrumentation) is essential to penetrate their natal dust shrouds, confirm whether they are gravitationally bound cores, and study star formation under high-extinction conditions.

- Distant Spiral-Arm Discoveries ($d > 7$ kpc): The compact discoveries confirmed by independent WISE HII kinematic distances between 7.7 and 12.6 kpc (such as clus4, clus61, and clus360; see Table 8.2) serve as vital probes of outer Galactic spiral arm structure. Combining deep near-infrared proper motions from the Nancy Grace Roman Space Telescope or VVV/VIRAC2 with spectroscopic radial velocities will enable precise 3D orbital mapping of these distant star-forming regions across the Milky Way disc.

Finally, incorporating dynamic virial binding tests using velocity dispersions (Hunt and Reffert 2023) once high-precision infrared radial velocities become available will allow us to directly observe the transition from embedded infant clusters to unbound associations, directly answering lingering questions regarding stellar infant mortality in the Galactic Plane.

Appendix A

Candidate Data

The consolidated candidate data is too large to embed in this document and has been provided in the spreadsheet *appendix-a_candidate_data.xlsx* that accompanies this thesis. The tables below provide descriptions of the columns in each table.

Table A.1 Column Descriptions - *appendix-a_candidate_data.xlsx*

Column Name	Description
—— Summary Data	
clusNum	Allocated during the original visual search
vis_ra	Visually derived RA in degrees
vis_dec	Visually derived Dec in degrees
vis_l	Visually derived galactic longitude in degrees
vis_b	Visually derived galactic latitude in degrees
arith_ra	Arithmetic mean derived RA in degrees
arith_dec	Arithmetic mean derived Dec in degrees
arith_offset	Offset between arithmetic mean and visual coordinates in arcsec
vis_appr_rad	Visually assigned apparent radius in arcmin
kde_ang_size	KDE derived half-maximum radius in arcmin
eff_core_rad	EFF derived core radius in arcmin
jhk_neb	Is any nebulosity visible in JH or K bands
mem_3d	Number of Bonatto derived cluster members - 3D algorithm
mem_2d	Number of Bonatto derived cluster members - 2D algorithm
3d_clus	Do the 3D CCD and CMDs show a likely cluster
2d_clus	Does the 2D CMD show a likely cluster
wise	Does WISE AllWISE show any warm material at the cluster coordinates

Continued on next page

Table A.1 – continued from previous page

Column Name	Description
qual_grade	Isochrone pipeline validity grading (Grade A: Very Clean Sequence, Grade B: Sensible Sequence, Grade C: Marginal, either through sparse membership ($N < 10$) or marginal scatter about the fitted sequence on visual review, Grade F: Excluded Artefact)
fit_age	Derived cluster age from the 2D MIST isochrone pipeline, given in $\log t$
fit_extinction	Derived visual extinction (A_V) from the isochrone fit, in magnitudes
fit_dist_mod	Best-fit distance modulus (DM) anchored by the Bayestar19 3D line-of-sight dust maps
fit_mbrs	Number of decontaminated stellar members fitted to the sequence (N_{stars})
yso_frac	Fraction of fitted stellar members displaying infrared excess / YSO characteristics (Q-parameter diagnostic)
gaia_dist_med	Median photogeometric distance for all matched Gaia DR3 sources, in parsecs
exp_pm_disp	Expected proper motion dispersion at *gaia_dist_med* distance, in $mas\ yr^{-1}$
act_pm_disp	Actual proper motion dispersion at *gaia_dist_med* distance, in $mas\ yr^{-1}$
diff_pct	Difference in proper motion dispersion at *gaia_dist_med* distance, in $mas\ yr^{-1}$
gaia_real	Is *diff_pct* within 50% of *exp_pm_disp*
hdbscan_match	Percentage match between Bonatto members and HDBSCAN members
hdbscan_real	Is *hdbscan_match* $\geq$ 25%
known	Does the candidate have any references in the databases or literature
candidate	If known, is the cluster confirmed or still candidate
real	Final assessment of whether there is a cluster
class	Derived from taxonomy described in §1.2.1
—— SIMBAD	
sf_markers	Star formation markers within 5 arcmin of arithmetic centre
clus_markers	Cluster markers within 5 arcmin of arithmetic centre
lit_refs	Literature refs across all *clus_markers* object IDs within 5 arcmin of arithmetic centre
—— UCC	
UCC_ID	UCC Object Reference
public_id_ucc	Publicly known object name
half_max_rad	Half-max radius of object in arcmin
separation	Distance between arithmetic centre and UCC object coords in arcsec
—— DAML02	
source_id	Public name of object
clus_class	Cluster type, such as Visible, Infrared etc
appr_diam	The apparent diameter of the cluster in arcmin
dist	Derivied distance from Sun in parsecs
age	Given in log t
num_mem	Number of cluster members
separation	Distance between arithmetic centre and object coords in arcsec
—— HNR	
name	Primary public name
hnr_id	HNR Reference
all_names	All publicly known names

Continued on next page

Table A.1 – continued from previous page

Column Name	Description
num_mem	Number of cluster members
half_rad	Half-maximum radius in degrees
dist	Distance from Sun in parsec
separation	Distance between arithmetic centre and object coords in arcsec
—— MWSC	
mwsc_source_id	Catalogue source id
Name	Publicly known name
ang_dia	Angular diameter in degrees
dist	Distance from Sun in parsecs
log_age	Given as log t
source_type	ass = stellar association, ast = Dias: possible asterism/dust hole/star cloud, dub = Dias: dubious, objects considered doubtful by the DSS images inspection, emb = embedded open cluster/cluster associated with nebulosity, glo = globular cluster/possible globular cluster, irc = infrared cluster, irg = infrared stellar group, mog = Dias: possible moving group, non = Dias: "non-existent NGC"/ objects not found in DSS images inspection, rem = Possible cluster remnant, var = clusters with variable extinction
separation	Distance between arithmetic centre and object coords in arcmin
—— FSR	
source_id	FSR internal reference
known_name_poss	Publicly known name
num_mem_core	Number of members in the core
separation	Distance between arithmetic centre and object coords in arcmin
—— WISE	
hii_source_id	
hii_source_name	Publicly known name
hii_class	K = Known, G = Group, C = Candidate, Q = radio quiet, ? = no radio data
appr	Apparent radius in arcsec
hii_name	HII region identifier
hii_group_id	Group membership identifier
dist_est	Distance from Sun in kpc
dist_meth	Method used to derive distance: Parallax, OG/Kin = outer Galaxy kinematic,TP/Kin = tangent point kinematic, HIEA+HISA/Kin = HIE/A & HISA kinematic, HIEA/Kin = HIE/A kinematic,H2CO/Kin = H2CO absorption kinematic, IRDC+HISA (Mol)/Kin = kinematic distances from Dunham et al. 2011 (J/ApJ/741/110), Visible/Kin = inner Galaxy Sharpless HII region kinematic (near distances assumed).
separation	Distance between arithmetic centre and object coords in arcmin

Bibliography

Adams, Fred C. and Philip C. Myers (June 2001). "Modes of Multiple Star Formation". In: *The Astrophysical Journal* 553.2, pp. 744–753.

Allen, Lori E. et al. (Sept. 2004). "Infrared Array Camera (IRAC) Colors of Young Stellar Objects". In: *The Astrophysical Journal Supplement Series* 154.1, pp. 363–366.

Allison, Goodwin, Parker, de Grijs, et al. (June 2009). "Dynamical mass segregation on a very short timescale". In: *The Astrophysical Journal Letters, Volume 700, Issue 2, pp. L99-L103 (2009).* 700, pp. L99–L103.

Allison, Goodwin, Parker, Zwart, et al. (Apr. 2010). "The early dynamical evolution of cool, clumpy star clusters". In: *Monthly Notices of the Royal Astronomical Society, Volume 407, Issue 2, pp. 1098-1107.* 407, pp. 1098–1107.

Anderson, L. D. et al. (Dec. 2013). "The WISE Catalog of Galactic HII Regions". In: *Astrophysical Journal, Supplement Series* 212.1.

Andre, Philippe, Derek Ward-Thompson, and Mary Barsony (Mar. 1993). "Submillimeter Continuum Observations of rho Ophiuchi A: The Candidate Protostar VLA 1623 and Prestellar Clumps". In: *Astrophysical Journal v.406, p.122* 406, p. 122.

André, Ph et al. (July 2010). "From filamentary clouds to prestellar cores to the stellar IMF: Initial highlights from the Herschel Gould Belt Survey". In: *Astronomy & Astrophysics* 518.8, p. L102.

Aparicio, A. et al. (Oct. 1993). "The Intermediate Age Open Cluster NGC 7044". In: *AJ* 106.4, p. 1547.

Arellano Ferro, A. et al. (Oct. 2023). "Variable stars in the field of the Galactic bulge globular cluster NGC 6522". In: *Astrophysics and Space Science* 368.10, pp. 1–13.

Astraatmadja, Tri L. and Coryn A. L. Bailer-Jones (Nov. 2016a). “ESTIMATING DISTANCES FROM PARALLAXES. II. PERFORMANCE OF BAYESIAN DISTANCE ESTIMATORS ON A GAIA-LIKE CATALOGUE”. In: *The Astrophysical Journal* 832.2, p. 137.

— (Dec. 2016b). “ESTIMATING DISTANCES FROM PARALLAXES. III. DISTANCES OF TWO MILLION STARS IN THE Gaia DR1 CATALOGUE”. In: *The Astrophysical Journal* 833.1, p. 119.

Bailer-Jones, C. A. L., J. Rybizki, M. Fouesneau, M. Demleitner, et al. (2021). “Estimating Distances from Parallaxes. V. Geometric and Photogeometric Distances to 1.47 Billion Stars in Gaia Early Data Release 3”. In: *The Astronomical Journal* 161.3.

Bailer-Jones, C. A. L., J. Rybizki, M. Fouesneau, G. Mantelet, et al. (July 2018). “Estimating Distance from Parallaxes. IV. Distances to 1.33 Billion Stars in Gaia Data Release 2”. In: *The Astronomical Journal* 156.2, p. 58.

Bailer-Jones, Coryn A. L. (Oct. 2015). “Estimating Distances from Parallaxes”. In: *Publications of the Astronomical Society of the Pacific* 127.956, pp. 994–1009.

Baraffe, I. et al. (May 1998). “Evolutionary models for solar metallicity low-mass stars: mass-magnitude relationships and color-magnitude diagrams”. In: *A&A* 337, pp. 403–412. arXiv: `9805009 [astro-ph]`.

Bastian, Nate, Kevin R. Covey, and Michael R. Meyer (Aug. 2010). “A Universal Stellar Initial Mass Function? A Critical Look at Variations”. In: *Annual Review of Astronomy and Astrophysics* 48.1, pp. 339–389.

Bate, Matthew R. (Nov. 2011). “Collapse of a molecular cloud core to stellar densities: The formation and evolution of pre-stellar discs”. In: *Monthly Notices of the Royal Astronomical Society* 417.3, pp. 2036–2056.

Baumgardt, H. et al. (Feb. 2019). “Mean proper motions, space orbits, and velocity dispersion profiles of Galactic globular clusters derived from Gaia DR2 data”. In: *Monthly Notices of the Royal Astronomical Society* 482.4, pp. 5138–5155.

Bell, Cameron P. M. et al. (Aug. 2007). “Pre-main-sequence stars in the young open cluster NGC 1893. II. Evidence for triggered massive star formation”. In: *A&A* 471.2, pp. 485–497. arXiv: 0703706 [astro-ph].

Bentley, Jon Louis (Sept. 1975). “Multidimensional Binary Search Trees Used for Associative Searching”. In: *Communications of the ACM* 18.9, pp. 509–517.

Berba, Pepe (Jan. 2020). *Understanding HDBSCAN and Density-Based Clustering*.

Bergh, S. van den and G. L. Hagen (1975). “Astronomical Journal, Vol. 80, p. 11 - 16”. In: *Astronomical Journal* 80, pp. 11–16.

Bessell, M. S. et al. (Sept. 1988). “JHKLM Photometry: Standard Systems, Passbands, and Intrinsic Colors”. In: *PASP* 100, p. 1134.

Bica, E. et al. (June 2003). “New infrared star clusters in the Northern and Equatorial Milky Way with 2MASS”. en. In: *Astronomy and Astrophysics* 404.1, pp. 223–232.

Bland-Hawthorn, Joss and Ortwin Gerhard (Sept. 2016). “The galaxy in context: Structural, kinematic, and integrated properties”. In: *Annual Review of Astronomy and Astrophysics* 54, pp. 529–596.

Blandford, R D and D G Payne (June 1982). “Hydromagnetic flows from accretion disks and the production of radio jets.” In: *Monthly Notices of the Royal Astronomical Society, Vol. 199, 883-903 (1982)* 199.4, p. 883.

Blitz, L. (Apr. 1982). “Giant molecular-cloud complexes in the galaxy”. In: *Scientific American* 246, pp. 84–94.

Blitz, Leo and Erik Rosolowsky (Nov. 2004). “The Initial Mass Function of GMCs”. In.

Boesgaard, A. M. and G. Steigman (1985). “Big Bang nucleosynthesis: theories and observations.” In: *Annual Rev. Astron. Astrophys.* 23, pp. 319–378.

Bonatto, C. and E. Bica (May 2007). “Open clusters in dense fields: the importance of field-star decontamination for NGC 5715, Lynga 4, Lynga 9, Trumpler 23, Trumpler 26 and Czernik 37”. In: *Monthly Notices of the Royal Astronomical Society* 377.3, pp. 1301–1323.

— (Apr. 2009). “The nature of the young and low-mass open clusters Pismis5, vdB80, NGC1931 and BDSB96”. In: *MNRAS* 397.4, pp. 1915–1925.

— (July 2010). “The fate of the pre-main sequence-rich clusters Collinder197 and vdB92: dissolution?” In: *Astronomy and Astrophysics* 516, A81.

Bonnor, W. B. (June 1956). “Boyle’s Law and Gravitational Instability”. In: *Monthly Notices of the Royal Astronomical Society* 116.3, pp. 351–359.

Borissova, J., S. Ramírez Alegría, et al. (Oct. 2016). “Young Stellar Clusters Containing Massive Young Stellar Objects in the VVV Survey”. In: *AJ* 152.3, p. 74.

Borissova, J., C. Bonatto, et al. (2011). “New Galactic star clusters discovered in the VVV survey”. In: *Astronomy and Astrophysics* 532, A131.

Borissova, J., V. D. Ivanov, et al. (Dec. 2018). “New Galactic star clusters discovered in the disc area of the VVVX survey”. In: *Monthly Notices of the Royal Astronomical Society* 481.3, pp. 3902–3920.

Bressan, Alessandro et al. (2012). “parsec : stellar tracks and isochrones with the PAdova and TRieste Stellar Evolution Code”. In: *Monthly Notices of the Royal Astronomical Society* 427, pp. 127–145.

Camargo, Bica, and Bonatto (Jan. 2015). “New Galactic embedded clusters and candidates from a WISE Survey”. In: *New Astronomy* 34, pp. 84–97.

Campello, Ricardo J.G.B., Davoud Moulavi, and Joerg Sander (2013). “Density-Based Clustering Based on Hierarchical Density Estimates”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 7819 LNAI.PART 2, pp. 160–172.

Cánovas, H. et al. (June 2019). “Census of ρ Ophiuchi candidate members from Gaia Data Release 2”. In: *Astronomy & Astrophysics* 626, A80.

Cantat-Gaudin, T. and F. Anders (Jan. 2020). “Clusters and mirages: Cataloguing stellar aggregates in the Milky Way”. In: *Astronomy and Astrophysics* 633, A99.

Cantat-Gaudin, T., F. Anders, et al. (Aug. 2020). “Painting a portrait of the Galactic disc with its stellar clusters”. In: *Astronomy & Astrophysics* 640, A1.

Casali, G., L. Magrini, et al. (Nov. 2020). “Stellar Population Astrophysics (SPA) with TNG: The old open clusters Collinder 350, Gulliver 51, NGC 7044, and Ruprecht 171”. In: *Astronomy and Astrophysics* 643, A12.

Casali, M., A. Adamson, et al. (May 2007). “The UKIRT wide-field camera”. In: *Astronomy and Astrophysics* 467.2, pp. 777–784.

Caswell, JL and SL Breen (2010). “An unbiased pilot survey for Galactic water masers”. In: *Monthly Notices of the Royal . . .* 407.4, pp. 2599–2610.

Chabrier, Gilles (July 2003). “Galactic Stellar and Substellar Initial Mass Function”. In: *Publications of the Astronomical Society of the Pacific* 115.809, pp. 763–795.

Chi, Huanbin et al. (May 2023). “LISC Catalog of Open Clusters III. 83 Newly Found Galactic Disk Open Clusters Using Gaia EDR3”. In: *Research in Astronomy and Astrophysics* 23.6, p. 065008.

Choi, Jieun et al. (May 2016). “MESA ISOCHRONES AND STELLAR TRACKS (MIST). I. SOLAR-SCALED MODELS”. In: *The Astrophysical Journal* 823.2, p. 102.

Cohen, M. and L. V. Kuhi (Dec. 1979). “Observational studies of pre-main-sequence evolution”. In: *The Astrophysical Journal Supplement Series* 41, p. 743.

Cox, Donald P. (Aug. 2005). “The three-phase interstellar medium revisited”. In: *Annual Review of Astronomy and Astrophysics* 43.Volume 43, 2005, pp. 337–385.

Cummings, Jeffrey D. et al. (Oct. 2018). “ The White Dwarf Initial–Final Mass Relation for Progenitor Stars from 0.85 to 7.5 M ”. In: *The Astrophysical Journal* 866.1, p. 21.

Deason, Alis J., Vasily Belokurov, and Jason L. Sanders (Dec. 2019). “The total stellar halo mass of the Milky Way”. In: *Monthly Notices of the Royal Astronomical Society* 490.3, pp. 3426–3439.

Dejonghe, Herwig (Jan. 1987). “A completely analytical family of anisotropic Plummer models”. In: *MNRAS* 224.1, pp. 13–39.

Dias, W. S. et al. (Mar. 2002). “New catalogue of optically visible open clusters and candidates”. In: *A&A* 389.3, pp. 871–873.

Dirienzo, William J. et al. (Dec. 2012). “Testing triggered star formation in six h II regions”. In: *Astronomical Journal* 144.6, p. 173.

Dotter, Aaron (Jan. 2016). “MESA ISOCHRONES AND STELLAR TRACKS (MIST) 0: METHODS FOR THE CONSTRUCTION OF STELLAR ISOCHRONES”. In: *The Astrophysical Journal Supplement Series* 222.1, p. 8.

Dunham, M. M. et al. (2014). “The Evolution of Protostars: Insights from Ten Years of Infrared Surveys with Spitzer and Herschel”. In: *Protostars and Planets VI*, pp. 195–218.

Dutra, CM et al. (2003). “New infrared star clusters in the southern Milky Way with 2MASS”. In: *Astronomy and Astrophysics* 400, pp. 533–539.

Dye, S. et al. (Nov. 2006). “The UKIRT Infrared Deep Sky Survey Early Data Release”. In: *Monthly Notices of the Royal Astronomical Society* 372.3, pp. 1227–1252.

Ebert, R. (1955). “Über die Verdichtung von H I-Gebieten. Mit 5 Textabbildungen”. In: *ZA* 37, p. 217.

Eldridge, Justin, Mikhail Belkin, and Yusu Wang (June 2015). “Beyond Hartigan Consistency: Merge Distortion Metric for Hierarchical Clustering”. In: *Journal of Machine Learning Research* 40.2015.

Elia, D. et al. (June 2017). “Properties of Hi-GAL clumps in the inner Galaxy”. In: *MNRAS* 471.1, pp. 100–143.

Elson, Rebecca A. W. et al. (Dec. 1987). “The Structure of Young Star Clusters in the Large Magellanic Cloud”. In: *ApJ* 323, p. 54.

Evans, Neal J. (1999). “Physical conditions in regions of star formation”. In: *Annual Review of Astronomy and Astrophysics* 37.1, pp. 311–362.

Evans, Neal J. et al. (2009). “The spitzer c2d legacy results: Star-formation rates and efficiencies; evolution and lifetimes”. In: *Astrophysical Journal, Supplement Series* 181.2, pp. 321–350.

Faustini, F. et al. (Sept. 2009). “Properties of stellar clusters around high-mass young stars”. In: *Astronomy & Astrophysics* 503.3, pp. 801–816.

Fedele, D. et al. (Nov. 2009). “Timescale of Mass Accretion in Pre-Main-Sequence Stars”. In: *A&A* 510.1, A72.

Frayn, C. M. and G. F. Gilmore (Oct. 2002). “The analysis of red giant branch photometry in galaxies”. In: *Monthly Notice of the Royal Astronomical Society, Volume 337, Issue 2, pp. 445-458.* 337.2, p. 445.

Froebrich, D and S Schmeja (2010). “Old star clusters in the FSR catalogue”. In: *Monthly Notices of the Royal Astronomical Society* 409.July, pp. 1281–1288.

Froebrich, D, A Scholz, and CL Raftery (2007). “A systematic survey for infrared star clusters with| b|< 20° using 2MASS”. In: *Monthly Notices of the . . .* 374, pp. 399–408.

Gieles, M. (June 2011). “Dynamical Evolution of Stellar Clusters”. In: *Stellar Clusters & Associations: A RIA Workshop on Gaia*, pp. 118–122.

Green, Gregory M. et al. (Dec. 2019). “A 3D Dust Map Based on Gaia, Pan-STARRS 1, and 2MASS”. In: *The Astrophysical Journal* 887.1, p. 93.

Greene, Thomas P. and Michael R. Meyer (Sept. 1995). “An Infrared Spectroscopic Survey of the rho Ophiuchi Young Stellar Cluster: Masses and Ages from the H-R Diagram”. In: *ApJ* 450, p. 233.

Gutermuth, R. A., S. T. Megeath, et al. (2009). “A Spitzer survey of young stellar clusters within one kiloparsec of the sun: Cluster core extraction and basic structural analysis”. In: *Astrophysical Journal, Supplement Series* 184.1, pp. 18–83.

Gutermuth, R. A., P. C. Myers, et al. (Feb. 2008). “ Spitzer Observations of NGC 1333: A Study of Structure and Evolution in a Nearby Embedded Cluster ”. In: *The Astrophysical Journal* 674.1, pp. 336–356.

Haisch Jr., Karl E., Elizabeth A. Lada, and Charles J. Lada (June 2001). “Disk Frequencies and Lifetimes in Young Clusters”. In: *The Astrophysical Journal* 553.2, pp. L153–L156.

Hambly, NC and RS Collins (2008). “The WFCAM science archive”. In: *Monthly Notices of the Royal Astronomical Society* 662, pp. 637–662.

Hewett, P. C. et al. (Apr. 2006). “The UKIRT Infrared Deep Sky Survey ZY JHK photometric system: passbands and synthetic colours”. In: *Monthly Notices of the Royal Astronomical Society* 367.2, pp. 454–468.

Hodgkin, S. T. et al. (Apr. 2009). “The UKIRT wide field camera ZYJHK photometric system: Calibration from 2MASS”. In: *Monthly Notices of the Royal Astronomical Society* 394.2, pp. 675–692.

Hunt, Emily L. and Sabine Reffert (Feb. 2021). “Improving the open cluster census”. In: *Astronomy & Astrophysics* 646, A104.

— (May 2023). “Improving the open cluster census - II. An all-sky cluster catalogue with Gaia DR3”. In: *Astronomy & Astrophysics* 673, A114.

Hurley, Jarrod and Christopher A. Tout (July 1998). “The Binary Second Sequence in Cluster Colour–Magnitude Diagrams”. In: *MNRAS* 300.4, pp. 977–980.

Immer, K. et al. (Jan. 2012). “Recent star formation in the inner Galactic bulge seen by ISOGAL - II. The central molecular zone”. In: *Astronomy & Astrophysics* 537, A121.

Irwin, Mike J. et al. (Sept. 2004). “VISTA data flow system: pipeline processing for WFCAM and VISTA”. In: *SPIE Astronomical Telescopes + Instrumentation*, p. 411.

Jatenco-Pereira, V. (June 2013). “Damping of Alfvén waves as a heating source in protostellar accretion discs”. In: *Monthly Notices of the Royal Astronomical Society* 431.4, pp. 3150–3158.

Jeans, J. H. (Jan. 1902). “The stability of a spherical nebula”. In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 199.312-320, pp. 1–53.

Joos, M., P. Hennebelle, and A. Ciardi (July 2012). “Protostellar disk formation and transport of angular momentum during magnetized core collapse”. In: *Astronomy & Astrophysics* 543, A128.

Kennicutt, Robert C. and Neal J. Evans (Sept. 2012). “Star Formation in the Milky Way and Nearby Galaxies”. en. In: *Annual Review of Astronomy and Astrophysics* 50.1, pp. 531–608.

Kharchenko, N. V. et al. (Oct. 2013). “Global survey of star clusters in the Milky Way: II. the catalogue of basic parameters”. In: *Astronomy and Astrophysics* 558, A53.

King, Ivan (Oct. 1962). “The structure of star clusters. I. an empirical density law”. In: *Astronomical Journal* 67.7, p. 471.

Kleinmann, S. G. et al. (1994). “The two micron all sky survey”. In: *Experimental Astronomy* 3.1-4, pp. 65–72.

Könyves, V. et al. (Dec. 2015). “A census of dense cores in the Aquila cloud complex: SPIRE/PACS observations from the Herschel Gould Belt survey”. In: *Astronomy & Astrophysics* 584, A91.

Kounkel, Marina and Kevin Covey (Aug. 2019). “Untangling the Galaxy. I. Local Structure and Star Formation History of the Milky Way”. In: *The Astronomical Journal* 158.3, p. 122.

Kounkel, Marina, Kevin Covey, and Keivan G. Stassun (Nov. 2020). “Untangling the Galaxy. II. Structure within 3 kpc”. In: *The Astronomical Journal* 160.6, p. 279.

Kroupa, P, CA Tout, and G Gilmore (1993). “The distribution of low-mass stars in the Galactic disc”. In: *Monthly Notices of the Royal . . .* 262, pp. 545–587.

Krumholz, Mark R. (June 2014). “The big problems in star formation: The star formation rate, stellar clustering, and the initial mass function”. In: *Physics Reports* 539.2, pp. 49–134.

Lada, Charles J and Bruce A Wilking (Dec. 1984). “The nature of the embedded population in the rho Ophiuchi dark cloud : mid-infrared observations.” In: *Astrophysical Journal, Vol. 287, p. 610-621 (1984)* 287, p. 610.

Lada, Charles J. (1987). “Star formation: from OB associations to protostars.” In: *IAU Symposium* 115, p. 1.

Lada and Lada (Sept. 2003). “Embedded Clusters in Molecular Clouds”. In: *Annual Review of Astronomy and Astrophysics* 41.1, pp. 57–115.

Lawrence, A and SJ Warren (2007). “The UKIRT Infrared Deep Sky Survey (UKIDSS)”. In: *Monthly Notices of the Royal Astronomical Society* 1617.February, pp. 1599–1617. arXiv: `0604426v2 [astro-ph]`.

Levy, E. H. and J. I. Lunine (1993). “Protostars and planets III”. In: *prpl*.

Lin, C. C. and Frank H. Shu (Aug. 1964). “On the Spiral Structure of Disk Galaxies.” In: *The Astrophysical Journal* 140, p. 646.

Lindegren, L. and U. Bastian (Feb. 2010). “Basic principles of scanning space astrometry”. In: *EAS Publications Series* 45, pp. 109–114.

Lindegren, L., U. Lammers, U. Bastian, et al. (Nov. 2016). “Gaia Data Release 1 - Astrometry: one billion positions, two million proper motions and parallaxes”. In: *Astronomy & Astrophysics* 595, A4.

Lindegren, L., U. Lammers, D. Hobbs, et al. (Feb. 2012). “The astrometric core solution for the Gaia mission - Overview of models, algorithms, and software implementation”. In: *Astronomy & Astrophysics* 538, A78.

Liptai, David et al. (2017). “Does turbulence determine the initial mass function?” In: *Monthly Notices of the Royal Astronomical Society* 465.1.

Longmore, S. N. et al. (2014). “The Formation and Early Evolution of Young Massive Clusters”. In: *Protostars and Planets VI*, pp. 291–314.

Lucas, P. W. et al. (Nov. 2008). “The UKIDSS Galactic Plane Survey”. In: *Monthly Notices of the Royal Astronomical Society* 391.1, pp. 136–163.

Lucas P. W. et al. (2013). “A Billion Stars: The Near-IR View of the Galaxy with the UKIDSS Galactic Plane Survey”. In: *Thirty Years of Astronomical Discovery with UKIRT*. Ed. by Andy Adamson, John Davies, and Ian Robson. Dordrecht: Springer Netherlands, pp. 279–289. ISBN: 978-94-007-7432-2.

Luri, X. et al. (Aug. 2018). “Gaia Data Release 2 - Using Gaia parallaxes”. In: *Astronomy & Astrophysics* 616, A9.

Lyo, A-Ran et al. (Apr. 2006). “A deep photometric survey of the eta Chamaeleontis cluster down to the brown dwarf - planet boundary”. In: *Monthly Notices of the Royal Astronomical Society, Volume 368, Issue 3, pp. 1451-1455.* 368, pp. 1451–1455.

Magrini, L. et al. (Jan. 2023). “The Gaia -ESO survey: Mapping the shape and evolution of the radial abundance gradients with open clusters”. In: *Astronomy and Astrophysics* 669, A119.

Majaess, D. (Nov. 2012). “Discovering protostars and their host clusters via WISE”. In: *Astrophysics and Space Science* 344.1, pp. 175–186.

Matteo, P Di et al. (2019). “MICADO, the ELT first-light imager”. In: *sf2a*, p. Di.

McInnes, Leland (May 2024). *https://github.com/scikit-learn-contrib/hdbscan*.

McInnes, Leland, John Healy, and Steve Astels (Mar. 2017). "hdbscan: Hierarchical density based clustering". In: *The Journal of Open Source Software* 2.11, p. 205.

McKee, Christopher F. and Eve C. Ostriker (Sept. 2007). "Theory of star formation". In: *Annual Review of Astronomy and Astrophysics* 45.Volume 45, 2007, pp. 565–687.

McLaughlin, Dean E. and Roeland P. van der Marel (May 2006). "Resolved Massive Star Clusters in the Milky Way and its Satellites: Brightness Profiles and a Catalogue of Fundamental Parameters". In: *ApJS* 161.2, pp. 304–360.

Megeath, S. T. et al. (Jan. 2016). "THE SPITZER SPACE TELESCOPE SURVEY OF THE ORION A AND B MOLECULAR CLOUDS. II. THE SPATIAL DISTRIBUTION AND DEMOGRAPHICS OF DUSTY YOUNG STELLAR OBJECTS". In: *The Astronomical Journal* 151.1, p. 5.

Mellon, Richard R. and Zhi-Yun Li (Sept. 2007). "Magnetic Braking and Protostellar Disk Formation: The Ideal MHD Limit". In: *ApJ* 681.2, pp. 1356–1376.

Men'shchikov, A. et al. (July 2010). "Filamentary structures and compact objects in the Aquila and Polaris clouds observed by Herschel". In: *Astronomy & Astrophysics* 518.8, p. L103.

Mercer, EP and DP Clemens (2005). "New star clusters discovered in the GLIMPSE survey". In: *The Astrophysical Journal* 635, pp. 560–569.

Mermilliod, Jean-Claude (1995). "The Database for Galactic Open Clusters (BDA)". In: pp. 127–138.

Meyer, Michael R., Nuria Calvet, and Lynne A. Hillenbrand (July 1997). "Intrinsic Near-Infrared Excesses of T Tauri Stars: Understanding the Classical T Tauri Star Locus". In: *Astronomical Journal v.114, p. 288-300 (1997)* 114.2, p. 288.

Moeckel, Nickolas and Matthew R. Bate (May 2010). "On the evolution of a star cluster and its multiple stellar systems following gas dispersal". In: *Monthly Notices of the Royal Astronomical Society* 404.2, pp. 721–737.

Myers, P. C. et al. (Jan. 1998). "Evolution of the Bolometric Temperature and Luminosity of Young Stellar Objects". In: *The Astrophysical Journal* 492.2, pp. 703–726.

Negueruela, Ignacio et al. (Mar. 2007). “Pre-main-sequence stars in the young open cluster NGC 1893: II. Evidence for triggered massive star formation”. In: *Astronomy and Astrophysics, Volume 471, Issue 2, August IV 2007, pp.485-497* 471.2, p. 485.

Nelder, J. A. and R. Mead (Jan. 1965). “A Simplex Method for Function Minimization”. In: *The Computer Journal* 7.4, pp. 308–313.

Nguyen, C. T. et al. (Sept. 2022). “PARSEC V2.0: Stellar tracks and isochrones of low- and intermediate-mass stars with rotation”. In: *Astronomy & Astrophysics* 665, A126.

Nikolaev, Sergei et al. (Dec. 2000). “A Global Photometric Analysis of 2MASS Calibration Data”. In: *The Astronomical Journal* 120.6, pp. 3340–3350.

Ochsenbein, F., P. Bauer, and J. Marcout (Apr. 2000). “The VizieR database of astronomical catalogues”. In: *Astronomy and Astrophysics Supplement Series* 143.1, pp. 23–32.

Offner, Stella S. R. et al. (Dec. 2013). “The Origin and Universality of the Stellar Initial Mass Function”. In: *eprint arXiv:1312.5326*, p. 24.

Olmi, L. et al. (Oct. 2018). “Large-scale properties of the clump mass function”. In: *Monthly Notices of the Royal Astronomical Society* 480.2, pp. 1831–1849.

Palla, Francesco (Jan. 1996). “The quest for evolutionarydiagrams of young stellar objects”. In: *Disks and Outflows Around Young Stars*. Vol. 465. Springer Berlin Heidelberg, pp. 143–168.

Pandey, A. K. et al. (Feb. 2013). “Optical photometric and polarimetric investigation of NGC 1931”. In: *Astrophysical Journal* 764.2.

Peretto, N, G a Fuller, et al. (2013). “Global collapse of molecular clouds as a formation mechanism for the most massive stars”. In: *Astronomy and Astrophysics* 555.6.

Peretto, N. and G. a. Fuller (Oct. 2009). “The initial conditions of stellar protocluster formation”. In: *Astronomy & Astrophysics* 505.1, pp. 405–415.

Peretto, N. and G. a. GA Fuller (Nov. 2010). “A statistical study of the mass and density structure of infrared dark clouds”. In: *The Astrophysical Journal* 723.1, pp. 555–562.

Perren, Gabriel I. et al. (Oct. 2023). “The Unified Cluster Catalogue: towards a comprehensive and homogeneous data base of stellar clusters”. In: *Monthly Notices of the Royal Astronomical Society* 526.3, pp. 4107–4119.

Perryman, M. A.C. (1989). “Hipparcos: astrometry from space”. In: *Nature 1989 340:6229* 340.6229, pp. 111–116.

Pezzuto, S. et al. (Jan. 2021). “Physical properties of the ambient medium and of dense cores in the Perseus star-forming region derived from Herschel Gould Belt Survey observations”. In: *Astronomy & Astrophysics* 645, A55.

Platais, I. et al. (May 2012). “The Effects of Differential Reddening and Stellar Rotation on the Appearance of Multiple Populations in Star Clusters: The Case of Trumpler 20”. In: *ApJL* 751.1, p. L8.

Plummer, H. C., Plummer, and H. C. (Mar. 1911). “On the problem of distribution in globular star clusters”. In: *MNRAS* 71.5, pp. 460–470.

Portegies Zwart, Simon F., Stephen L.W. McMillan, and Mark Gieles (Sept. 2010). “Young massive star clusters”. In: *Annual Review of Astronomy and Astrophysics* 48, pp. 431–493.

Prim, R. C. (1957). “Shortest Connection Networks And Some Generalizations”. In: *Bell System Technical Journal* 36.6, pp. 1389–1401.

Prusti, T. et al. (Nov. 2016). “The Gaia mission”. In: *Astronomy & Astrophysics* 595, A1.

Ray, T.P. and J. Ferreira (Dec. 2021). “Jets from young stars”. In: *New Astronomy Reviews* 93, p. 101615.

Reid, Michael A. and Christine D. Wilson (Oct. 2006). “High-Mass Star Formation. III. The Functional Form of the Submillimeter Clump Mass Function”. In: *The Astrophysical Journal* 650.2, pp. 970–984.

Rudolph, Andreas et al. (May 2014). “Gaia Mission Operations Concept and Launch and Early Orbit Phase - In-Orbit Experience”. In: *SpaceOps 2014 Conference*. Reston, Virginia: American Institute of Aeronautics and Astronautics. ISBN: 978-1-62410-221-9.

Rybizki, Jan et al. (May 2020). “A Gaia Early DR3 Mock Stellar Catalog: Galactic Prior and Selection Function”. In: *Publications of the Astronomical Society of the Pacific* 132.1013, p. 074501.

Salpeter, EE (1955). “The luminosity function and stellar evolution.” In: *The Astrophysical Journal* 121, p. 161.

Shu, FH, FC Adams, and S Lizano (1987). “Star formation in molecular clouds-Observation and theory”. In: *Annual Review of Astronomy . . .* 25.1, pp. 23–81.

Silk, J., Silk, and J. (May 1977). “On the fragmentation of cosmic gas clouds. II. Opacity-limited star formation.” In: *ApJ* 214, pp. 152–160.

Skrutskie, M. F. et al. (Feb. 2006). “The Two Micron All Sky Survey (2MASS)”. In: *The Astronomical Journal* 131.2, pp. 1163–1183.

Smith, Rowan J., Paul C. Clark, and Ian A. Bonnell (Dec. 2008). “The structure of molecular clouds and the universality of the clump mass function”. In: *Monthly Notices of the Royal Astronomical Society* 391.3, pp. 1091–1099.

Smith, Rowan J., Simon C. O. Glover, Ralf S. Klessen, and Gary A. Fuller (Feb. 2016). “On the nature of star-forming filaments - II. Subfilaments and velocities”. In: *Monthly Notices of the Royal Astronomical Society* 455.4, pp. 3640–3655.

Smith, Rowan J., Simon C. O. Glover, and Ralf. S. Klessen (Dec. 2014). “On the nature of star-forming filaments – I. Filament morphologies”. In: *Monthly Notices of the Royal Astronomical Society* 445.3, pp. 2900–2917.

Solin, O., E. Ukkonen, and L. Haikala (June 2012). “Mining the UKIDSS Galactic Plane Survey: star formation and embedded clusters”. In: *Astronomy & Astrophysics* 542, A3.

Solomon, P. M. and A. R. Rivolo (Apr. 1989). “A Face-on View of the First Galactic Quadrant in Molecular Clouds”. In: *ApJ* 339, p. 919.

Sosey, Megan et al. (2021). “Astrometry with Roman and Its Relationship to Science Data Products”. In: *roman*, Technical Report Roman-STScI–000243.

Spina, Lorenzo, Laura Magrini, and Katia Cunha (Feb. 2022). “Mapping the Galactic Metallicity Gradient with Open Clusters: The State-of-the-Art and Future Challenges”. In: *Universe* 8.2, p. 87.

Springel, Volker (Dec. 2005). “The cosmological simulation code GADGET-2”. In: *Monthly Notices of the Royal Astronomical Society* 364.4, pp. 1105–1134.

Stahler, Steven W. and Francesco Palla (Jan. 2008). “The Formation of Stars”. In: *The Formation of Stars*, pp. 1–852.

Straižys, V. and Romualda Lazauskaitė (2009). “Intrinsic Color Indices and Luminosity Sequences of Stars in the 2MASS Two-Color Diagram”. In: *Baltic Astronomy, Vol. 18, p. 19-31* 18, p. 19.

Tarricq, Y. et al. (Mar. 2021). “3D kinematics and age distribution of the open cluster population”. In: *Astronomy & Astrophysics* 647, A19.

Taylor, M. B. et al. (2005). “TOPCAT & STIL: Starlink Table/VOTable Processing Software”. In: *ASPC* 347, p. 29.

Tokunaga, A. T., D. A. Simons, and W. D. Vacca (Feb. 2002). “The Mauna Kea Observatories Near-Infrared Filter Set. II. Specifications for a New JHKL M Filter Set for Infrared Astronomy”. In: *Publications of the Astronomical Society of the Pacific* 114.792, pp. 180–186.

Ungerechts, H. and P. Thaddeus (Mar. 1987). “A Co survey of the dark nebulae in Perseus, Taurus, and Auriga”. In: *The Astrophysical Journal Supplement Series* 63, p. 645.

Vallenari, A. et al. (June 2023). “Gaia Data Release 3”. In: *Astronomy & Astrophysics* 674, A1.

Wang, Shu and Xiaodian Chen (June 2019). “The Optical to Mid-infrared Extinction Law Based on the APOGEE, Gaia DR2, Pan-STARRS1, SDSS, APASS, 2MASS, and WISE Surveys”. In: *The Astrophysical Journal* 877.2, p. 116.

Wenger, M. et al. (Apr. 2000). “The SIMBAD astronomical database - The CDS reference database for astronomical objects”. In: *Astronomy and Astrophysics Supplement Series* 143.1, pp. 9–22.

Wright, Edward L., Peter R. M. Eisenhardt, et al. (Dec. 2010). “THE WIDE-FIELD INFRARED SURVEY EXPLORER (WISE): MISSION DESCRIPTION AND INITIAL ON-ORBIT PERFORMANCE”. In: *The Astronomical Journal* 140.6, pp. 1868–1881.

Wright, Nicholas J., Herve Bouy, et al. (Aug. 2016). “Cygnus OB2 DANCe: A high-precision proper motion study of the Cygnus OB2 association”. In: *Monthly Notices of the Royal Astronomical Society* 460.3, pp. 2593–2610.

Young, Alison K. et al. (Feb. 2018). “What can the SEDs of first hydrostatic core candidates reveal about their nature?” In: *Monthly Notices of the Royal Astronomical Society* 474.1, pp. 800–823.