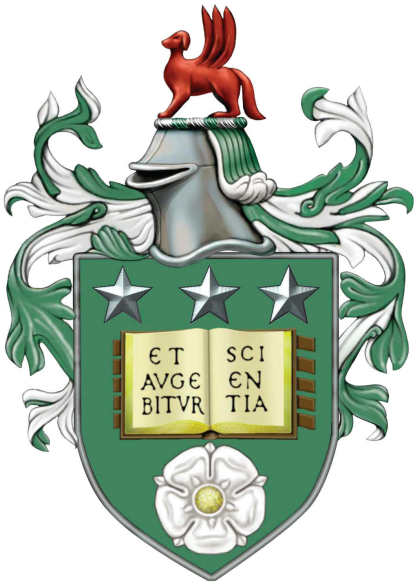
A Comedy of Uncertainties - Subclustering Stellar Clusters Using Spatial and Multi-Stage Methods

Joseph Eatson

Research Project - 2017



Contents

Abstract

Feasibility studies were performed on multiple subclustering techniques, focussing primarily on NGC 2264 and the Hyades cluster. Many techniques, such as age-based subclustering using markers such as lithium abundance and isochrone fitting were unsuccessful due to insufficient data and time being available. However, with newer data subclustering with these methods could be possible. Two of the studies were chosen to be pursued further, a novel 4-stage subclustering method involving DBSCAN and k-means on a precursor dataset containing protostar positions based on dust emission, as well as another round of DBSCAN and k-means on the main cluster dataset, derived from the MPCM database of the MYStIX survey. The second study involved a simplistic 2-stage DBSCAN/k-means method on 3-D position data produced from the Tycho-Gaia Astrometric Solution survey. Both were found to be in agreement with scientific literature, in particular the 4-stage method, which showed a marked improvement in accuracy compared to the 2-stage method used in a previous project.

1 Introduction

The aim of this project is to identify substructure and subclusters within the stellar cluster NGC 2264 using a multi-stage subclustering method, a young open cluster with 1173 members, as well perform 3-D spatial subclustering on the nearby Hyades cluster. While the use of multi-stage methods has been performed by students in previous years, this method is more complex, utilising 4-stages and a precursor dataset. In addition, 3-D spatial subclustering has not been performed before, due to the recent release of the TGAS survey dataset.

Feasibility studies were conducted with novel methods of subclustering, such as ageing through lithium and H^{α} emission lines as well as detection of embedded objects using colour excess and dust emission.

Stars within the same cluster are approximately the same age, having formed from the same giant molecular cloud. As the cloud collapses due to a perturbation resulting in Jeans instability, its density rapidly increases leading to a decreasing Jeans mass, resulting in multiple regions of the cloud independently collapsing into independent fragments. These clouds then further subdivide and collapse until star formation occurs. As the timescale of these collapses is $\approx 10^5 - 10^6$ years there is a disparity in formation time between stars in the clusters, necessitating the use of a hierarchy to describe relations between stars within a cluster, as such a comparison can be made to taxonomic classification.

It was initially assumed that fragments collapse more or less spherically. However, many star forming regions contain large filamentary regions where star formation is high, a good subclustering model should be able to work with both modes of collapse.

Subclustering is useful for refining models of cluster formation, as young clusters have a variety of morphologies, subclustering is necessary to compare clusters quantitatively to each other [Kuhn et al., 2014]. By determining common features and substructure using subclustering, further insight can be made into the formation of the cluster as a whole.

Subclustering is also essential for the continued understanding of stellar evolution, particularly in the field of young stellar object (YSO) evolution, as YSOs undergo drastic changes on timescales equivalent to the free-fall time of a molecular cloud. By categorising stars into subclusters, the age range can be reduced, allowing more accurate comparisons of stars with various properties, but similar ages, leading to the refinement of stellar evolution models.

1.1 Subclustering

Hierarchical subclustering has been used as a method of subdividing stellar since the 1970's [Aarseth and Hills, 1972], typically spatial subclustering was used early on. However, the advent of high-resolution spectrographs, as well as non-visible light telescopes have meant that other properties of cluster members can be observed, allowing for more refined subcluster models.

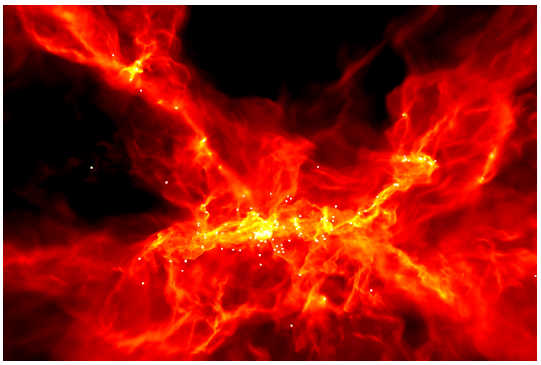


Figure 1: Snapshot of a hydrodynamic model showing fragmentation of a giant molecular cloud into smaller clouds, eventually the denser regions will self-attract and increase in density again, leading to further fragmentation, until a minimum mass is achieved, and star formation occurs, source [Bate, 2012]

Rory Boyer, who worked on a previous stellar clusters project [Boyer, 2016], using a single-stage subclustering method in order to determine cluster hierarchy, to display this he utilised dendrograms, which can display the fragmenting substructure of a cluster at a glance. However, dendrograms are not well suited for displaying positional relation, an essential aspect as spatial relation is by far the most important marker of relation in stellar subclustering.

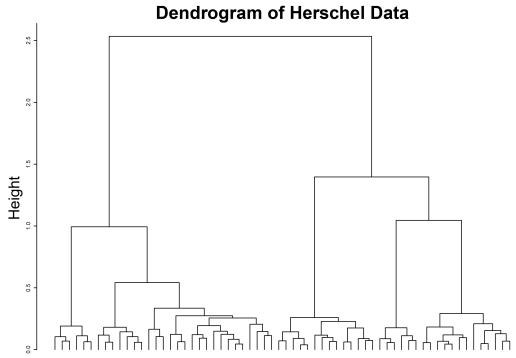


Figure 2: An example of a dendrogram from this project, utilising an early version of the 4-stage subclustering algorithm, dendrograms are well suited for displaying subcluster hierarchy, but do not show positional relation

Multiple methods for subclustering exist, typically relying on spatial relation algorithms, which are detailed in section 2. Spatial relation relies on stars being close to their initial formation site, and hence the subcluster they formed in, this is appropriate for young open clusters and gravitationally bound globular clusters, but for older open clusters spatial relation may not hold.

Other methods exist, typically based on determining the absolute or relative age of stars [Palla and Stahler, 1999], this can be accomplished by various methods, detailed in section 3.2. Additionally using more complex spatial data, such as 3-D parallax data could be used to subcluster stars more precisely, by removing the uncertainty of distance from the observer, this can also be used in conjunction with a method of ageing called kinematic ageing, in order to provide absolute ages for stars in a cluster.

The main drawback of this more complex methods is the data required, parallax data, in particular, is difficult to acquire, due to the extremely high angular resolution required, with parallax surveys until recently only reaching a maximum of a few hundred parsecs for point sources, precluding all but the closest clusters; ageing through colour or spectroscopic means, such as HR isochrone fitting requires either accurate spectroscopic data at high-resolution, or fluxes from sensitive telescopes that have been corrected for extinction with dust maps.

However, recently much of this required data has become available due to advances in telescope technology, which could revolutionise the field of subclustering. The Gaia telescope, detailed in section 3.1, has significantly greater angular resolution than Hipparcos, the previous parallax survey telescope, and is capable of deriving parallaxes of $\approx 1\%$ of stars in the galaxy. High-resolution dust maps due to the Herschel space telescope exist, as well as new theoretical models allowing for precise colour correction [Schlafly and Finkbeiner, 2011]. Finally, massive new telescopes, under construction, such as the E-ELT could perform spectroscopic surveys of every member in a cluster, allowing for precise abundances to be known [Ramsay et al., 2014].

This project aims to determine which methods are feasible with current or near-future data and to add additional stages and improvements to the algorithms used in previous years projects by Gorton and Boyer.

2 Subclustering Algorithms

After data is downloaded, the determination of subcluster within the structure can begin, this is achieved by the use of a clustering algorithm. Algorithms consists of two major categories, supervised and unsupervised methods.

The earliest, and simplest method of unsupervised subclustering is the "friends of friends" algorithm [Sneath, 1957], initially developed for taxonomy, was later found to be useful for determining cluster members. In simple unsupervised methods such as FoF, subclustering is entirely distance-based and defined by density. Supervised methods categorise objects based on classified examples, rather than simply clustering through position or emission.

2.1 Previous Project, DBSCAN and K-Means

A previous years project, performed by Jack Gorton [Gorton, 2015] used a combination of DBSCAN and the k-means algorithm in a 2-stage process to subcluster objects. DBSCAN was developed by as a density-based method of clustering objects [Ester et al., 1996], and was primarily used to reduce noise in the data, before k-means would be used for the actual subclustering.

DBSCAN has two parameters, `minpts` and ϵ (also referred to as `eps`), `minpts` refers to the minimum number of objects required to form a cluster, higher values are typically used for denser datasets, while ϵ refers to the maximum allowed distance from one

object to another for them to be considered related. An important aspect of implementing DBSCAN properly is selecting suitable values for both ϵ and `minpts`, otherwise too many objects could be considered cluster members, resulting in clusters as large as the dataset itself, or too many objects are considered noise, meaning that no clusters are generated. Gorton inspected his clusters visually to determine these values, and one of the goals of this project was to reduce the level of visual inspection required, to increase automation in the future.

DBSCANs main strength is that it can calculate the number of subclusters without being told in advance, unlike k-means which requires a parameter, n.

K-means clustering is an iterative subclustering algorithm that uses a different subclustering method to DBSCAN. By randomly selecting objects within the dataset to be "means", and then clustering the nearest objects with each mean, by iterating, subclusters based around centroids can be found [Lloyd, 1982]. However, k-means is quite sensitive to "noise" (elements in a dataset that do not fall into a particular cluster, in this instance) as it tries to include every object in a dataset. Additionally as previously stated, it requires an explicit number of centroids for subclustering, which can make subclustering quite involved for large datasets if visual inspection is used.

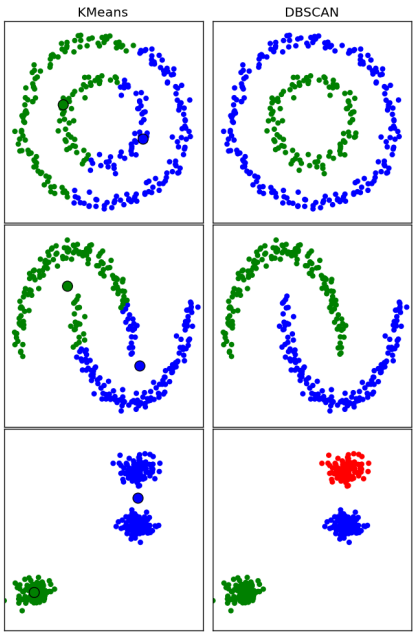


Figure 3: Comparison between k-means and DBSCAN, showing the benefits and drawbacks of each algorithm, source https://ogrisel.github.io/scikit-learn.org/sklearn-tutorial/modules/clustering.html

[1] $8\mu m$ flux data

By coupling the number of clusters found by DBSCAN to the k-means algorithm, a reduced amount of visual inspection needs to be performed, potentially increasing automation of the process.

In Gorton's project, DBSCAN was used to remove potential non-cluster objects; remaining stars were assumed to be cluster members. To further reduce the number of field stars, colour cuts were used to isolate all but YSOs, which have a fairly flat, featureless red spectrum. However, Gorton had not taken into account edge cases such as stars that were not deeply embedded in the star forming regions. These were entirely removed before the algorithm ran. In addition, DBSCAN noise reduction used an ϵ value based on visual inspection, which potentially biased the result. Finally, $F_8.0$[1] data could not be acquired for a majority of cluster members, meaning that rudimentary colour cuts not factoring $F_{8.0}$ data were used, despite the consensus that $F_{8.0}$ data is required for accurate colour cuts. Another method, such as utilising age data in conjunction with milder colour cuts, could potentially yield more accurate results.

2.2 Multi-Stage Methods

Gorton's subclstering method was a 2-stage model by using DBSCAN then k-means in succession, additional steps in the model may result in subclustering with a better fit. By adding another layer of data, such as a survey containing additional properties, embedded objects or spectroscopic data for instance and using those to produce centroids that would be imported into a secondary round of k-means using the original positional dataset, novel methods of subclustering could be found. This was the primary focus of this project, the main challenge was determining a suitable precursor dataset for initial subclustering.

2.3 Subclustering by MYStIX

The MYStIX survey uses a modified finite mixture model [Kuhn et al., 2014], which models spatial distributions of subclusters using isothermal ellipsoids. Ellipsoids are used as they are a reasonable shape for a subcluster formed from a collapsing molecular cloud, and can cover scenarios from spherical to filamentary collapse. In addition, multiple models are used and selected based on aptness for the particular cluster using the Akaike Information Criterion (AIC). This is a significantly more involved and accurate method than Gorton and Boyer utilised, however is beyond the scope of what is attainable in terms of complexity for this project.

The finite mixture model has multiple unique benefits, such as the nesting of subclusters, which is

exceptionally useful as subclusters may be in front of or behind others, an extremely useful factor if the dataset used does not contain parallax information, secondly, a critical density such as the DBSCAN derived method in [Gorton, 2015] is not assumed, a successful flexible model should have as few assumptions as possible to be truly effective. However, the method is highly involved and requires the foreground and background to be removed, this is done via measurement of X-ray emission and infrared excess; this is expanded upon in section 3.2.1. This complexity was considered beyond the scope of this project.

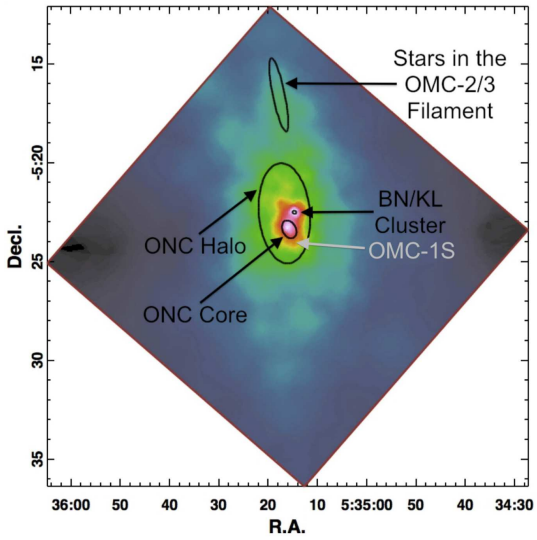


Figure 4: Finite mixture subclustering example from the Orion star-forming region, source [Kuhn et al., 2014]

3 Potential Subclustering Methods

The first half of the project consisted of feasibility studies on various novel methods of subclustering, using either recently available data or unconventional combinations of existing datasets.

Subclustering techniques can be roughly divided into spatial relation and age relation. Spatial relations subcluster based on the position and kinematics of stars within clusters; in young clusters ($<$ 10Myr), it is safe to assume that the position of stars within the cluster has not changed significantly, hence stars forming from the same cluster are still spatially near the stars that also formed from the same molecular cloud fragment.

Age relation allows for an independent subclustering method that could be used in conjunction with spatial relation in order to reduce uncertainty, especially for older clusters, where stars may have drifted beyond their formation sites.

Stars can be aged with a variety of methods, with semi-fundamental methods such as the lithium depletion boundary method and kinematic ages to model dependent methods such as abundances and HR/isochrone fitting. Methods of ageing will be expanded upon in section 3.2.

3.1 Spatial Mapping

The first study performed, and perhaps the most promising, was the use of Gaia astrometric data to group stars in 3-D.

3.1.1 Introduction

While spatial mapping using parallax data has been performed in the past, the lack of stars over 100 parsecs has limited its use to all but the nearest clusters, such as the Hyades [de Bruijne et al., 2001]. However, the Gaia telescope has a significantly higher angular resolution and sensitivity, able to detect fainter objects than Hipparcos, while still being able to determine the parallax of these objects.

	Hipparcos	Gaia
Limiting Magnitude	$\sim 12.4 mag$	~ 20
Angular Resolution	$\sim 1 mas$	$\sim 20 \mu as$
Catalogue Size	117,955	$\approx 1 \times 10^9$

Table 1: Comparison of Hipparcos and Gaia Performance, Hipparcos statistics reduced from [Van Leeuwen, 2007]

Unfortunately, as the telescope has only completed a single orbit around the sun, positional data has been measured, but not parallax data. In order to provide parallaxes, the Tycho Gaia Astrometric Solution (TGAS) [Michalik et al., 2015] is used. TGAS utilises a combination of the Tycho-2 parallax catalogue from Hipparcos as well as the Gaia DR1 positional data, released in September 2016. The Tycho-2 catalogue is a release that only contains positional data from multiple observations, however, calculating parallaxes was beyond the capability of Hipparcos. TGAS incorporates Gaia DR-1 data, and the enhanced sensitivity of Gaia reducing parallaxes from the data has been made possible.

To demonstrate the improvements TGAS offers over the Hipparcos catalogue, a region (table 2) was downloaded, comparing the parallax and associated error (figure 5). While there is no correlation between reducing parallax and error due to environmental reasons (density of region and extinction), the advantages are readily apparent. As can be seen, the uncertainty in parallax is considerably lower across all values, the region observed has a significantly larger number of detected objects and there is a greater number of $< 1 mas$ objects,

indicating that TGAS can discern parallaxes from a greater distance.

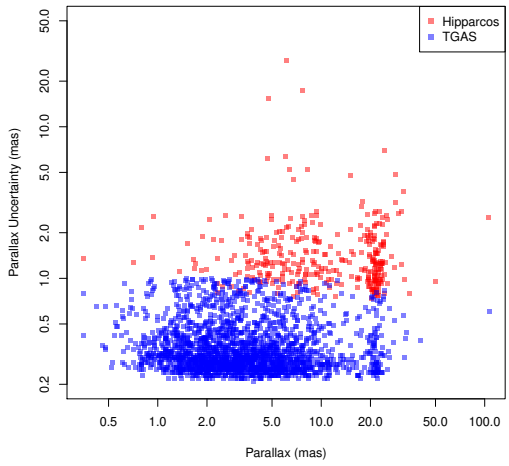


Figure 5: Comparison of parallax and its associated uncertainty between Hipparcos and TGAS datasets

Target	RA	DEC	Radius
Hyades	66.725°	15.867°	360’

Table 2: Download parameters for figure 5

3.1.2 Method

Data was acquired using the Gaia archive site[2] on 3 clusters in order to assess performance in different distance regimes.

Parallax was converted to distance, with uncertainty calculated with the formula:

$$\Delta D = \frac{\Delta\theta}{\theta} D \tag{1}$$

Erroneous parallaxes were removed, and clusters were assumed to be spherical, so objects not within the parameter $D_{\text{cluster}} \pm r_{\text{cluster}}$ were removed. Data not related to the spatial components of stars (magnitude, velocity) was removed, and rendered out into graphs using the `matplotlib` package.

3.1.3 Performance and Assessment

In the Hyades plot (figure 6) a suitable number of objects were detected within the distance range, with around 100 potential members, and 119 objects detected, this result seems plausible. The mean distance uncertainty was found to be 0.82 parsecs, up to a maximum uncertainty of 1.94 parsecs. This is more than adequate for subclustering, as it is within 10% of the total radius of the cluster.

[2] http://gea.esac.esa.int/archive/

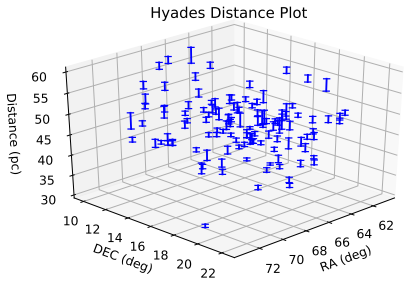


Figure 6: Hyades spatial plot using TGAS data

Blanco 1 (figure 7) seems to be fairly well populated. However, in reality, there are only parallaxes for around a third of the total population. Spatial subclustering proves to be even more unsuitable as the mean distance uncertainty is larger than the tidal radius of the cluster (28.99pc vs. 22.8pc). This is entirely inadequate for subclustering, and while Gaia will result in significant reductions to uncertainty, DR2 is not scheduled until April 2018, far beyond the scope of this project.

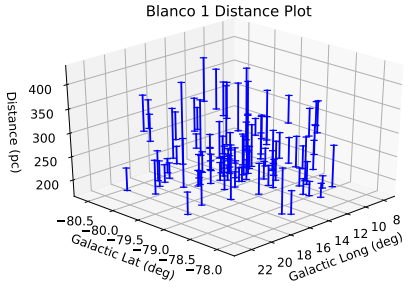


Figure 7: Blanco 1 spatial plot using TGAS data

Cygnus OB2 was still attempted (figure 8), and was found to be even more inaccurate. Despite being one of the most massive clusters in the galaxy, with approximately 2600 OB stars, only 30 parallaxes were found. Even worse than that is the distance error, with a mean error of 1180pc, and a max uncertainty of 2750 parsecs, the error at times is not only 2 orders of magnitude greater than the tidal radius of the cluster, but the distance uncertainty is on par with the distance itself.

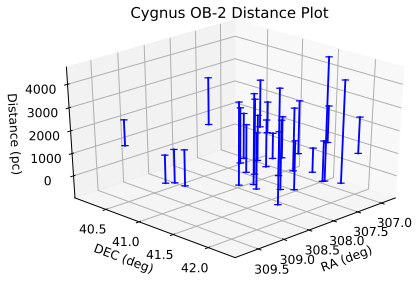


Figure 8: Cygnus-OB2 spatial plot using TGAS data

It is important to note that Cygnus contains over 100 O-type stars, which are most likely the stars that are being detected by TGAS, despite this, the uncertainty distance swiftly rises to unusable levels beyond 100pc.

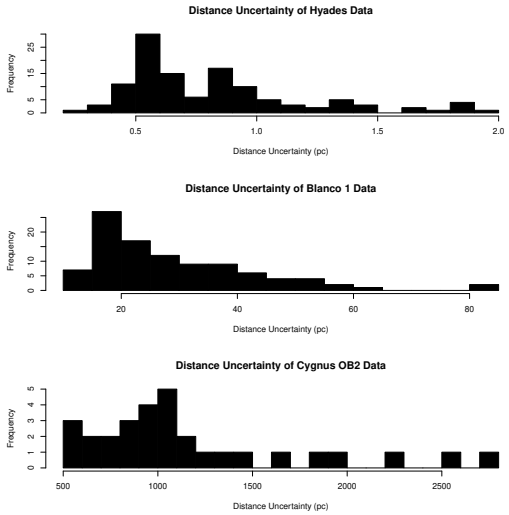


Figure 9: Histogram of distance uncertainties across all clusters

Despite the short range of TGAS, subclustering in 3-D could prove to be extremely useful when the data is available. And hence a simple clustering algorithm was created and applied to the Hyades data at a late stage in the project.

In the future, Gaia could prove revolutionary for subclustering. However, data outside of the design specifications does not exist, and its performance may suffer when dealing with dense fields, such as in distance clusters. Initial data releases will not release dense fields data above $M_v = 15$, which may hinder the adoption of spatial subclustering. Figure 5 clearly shows a significant decrease in uncertainty across all parallaxes, use of pure Gaia data in the future will reduce this further, due to drastically increased angular resolution.

3.2 Ageing

Ageing using various methods were attempted on NGC 2264, which was too distant to examine using TGAS, multiple catalogues were downloaded, containing various ageing metrics, and studies were performed on each metric in order to determine its viability.

3.2.1 Primary Cluster Catalogues - The MPCM & MIRES Surveys

The primary catalogue used to confirm members of NGC 2264 is the MYStIX MPCM[3] catalogue, containing 1173 members, this is merged with any catalogue in order to confirm that it is a part of the cluster.

The MYStIX MPCM catalogue consists of data from the MIRES[4] with additional verification using data from Chandra. MIRES utilises infrared excess to determine the presence of embedded sources using NIR data from UKIRT, as well less accurate 2MASS data for regions not covered by UKIRT, MIR data from the Spitzer GLIMPSE survey was then matched to the NIR data for a wider spectrum. Embedded objects were then found by determining the spectral energy distribution, any excess caused by absorption and re-emission of photons by the molecular cloud would cause an atypical emission profile and were flagged [Povich et al., 2013].

Embedded objects undergo verification with an independent process by using Chandra to find X-ray emission from shocks due to interaction with YSOs and the surrounding molecular cloud. X-Ray emission sources may have been wrongly associated with field stars instead of cluster members, meaning that "proximity-only" matching IR and X-Ray data could not be performed. This problem was compounded by the poor angular resolution of X-Ray telescopes, hence Bayesian statistics are used to combine MIRES and Chandra catalogue objects successfully [Naylor et al., 2013]. This does lead to potential problems within the sample, as the statistical matching may not be completely accurate especially in denser regions of space, which may have foreground and background field stars [Townsley et al., 2014].

Chandra	MIRES	MPCM
1328	805	1173

Table 3: Summary Source Populations for NGC 2264, source [Feigelson et al., 2013]

[3]MYStIX Probable Complex Member
[4]MYStIX Infra-Red Excess Source

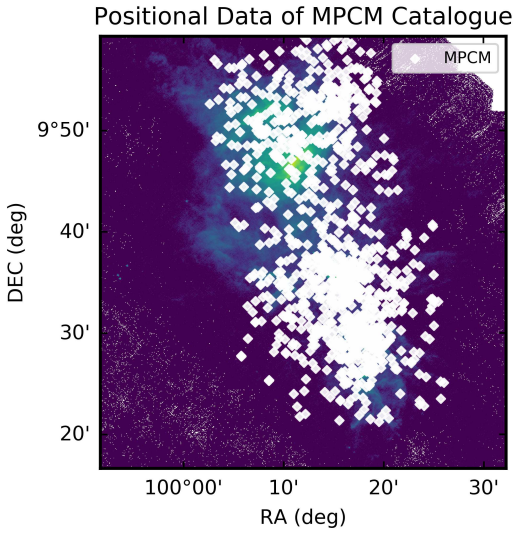


Figure 10: Herschel image of NGC 2264 overlaid with MPCM data

3.2.2 Surveys of the Cluster NGC 2264

In addition to MPCM, a survey containing age data was used as a secondary source [Mariñas et al., 2013], ages were determined through evolutionary model fitting [Baraffe et al., 1998]. Ageing through this manner is sensitive not only to uncertainty produced by interstellar extinction (which is only exacerbated by the cluster environment) but the uncertainty inherent in the model itself; because of this, multiple models tend to be used.

As the Mariñas survey only covers the southern region of the cluster, and lithium was not measured, additional surveys were acquired in order to determine a relation between the measured age, and stellar properties that have an age dependence.

The first of these tertiary sources was the Lim survey [Lim et al., 2016], which used the Hectochelle spectrograph to measure the lithium resonance doublet and H^{α} line emission for 86 objects distributed throughout the cluster. While the Lim survey contains calculated lithium abundances, other surveys did not as it was not the core focus of these surveys, hence only lithium equivalent line width was used, as it was considered sufficient.

The Bouvier survey [Bouvier et al., 2016] was the second used, as 86 objects was deemed insufficient for the cluster. This dataset uses the CSI 2264 and GES databases to determine line strengths of T-Tauri stars within NGC 2264. Having 201 members, some of which matching to the Lim survey, as well as black body temperature data, allowing for isochrone matching similar to the Mariñas survey.

A third survey, also measuring T-Tauri stars within the cluster, contains lithium, H^{α} and radial velocities, for the use of detecting accretion disks forming around T-Tauri stars [Sergison et al., 2013]. While this radial data could be used as a method of determining the ages of stars through kinematic ageing, this was quickly abandoned, due to insufficient data and uncertainty over whether it would be viable. However, radial data could be used in conjunction with Gaia parallax data, in order to trace star formation sites within clusters in the future.

The final tertiary survey used was the Dahm survey [Dahm and Simon, 2005]. This survey had significantly more members than other surveys, at 490 members, most of which containing H^{α} equivalent line widths, as well as Lithium emission line widths. Additionally the survey has a broad distribution of stars across both regions of the cluster, which made it an ideal candidate for the use of bootstrapping.

In addition to these surveys, Herschel 70 micron data was used to find strong dust emission regions in the cluster, to determine regions of strong extinction. Figure 11a contains the location of objects in all the tertiary catalogues used, overlaid onto the Herschel data.

3.2.3 Lithium

Lithium line width is an extremely good measure of relative age [Soderblom et al., 2014] as it is actively destroyed by fusion reactions in young stars, causing a dramatic reduction in line width over time. Figure 12 shows that there is a rapid dip in lithium content that directly correlates with age over a timescale of millions of years that is colour and spectral type independent. A large amount of this project was spent on the feasibility study of lithium ageing, as well as H^{α} ageing (section 3.2.4)

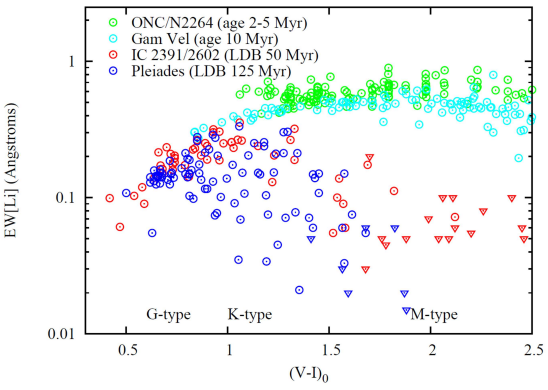
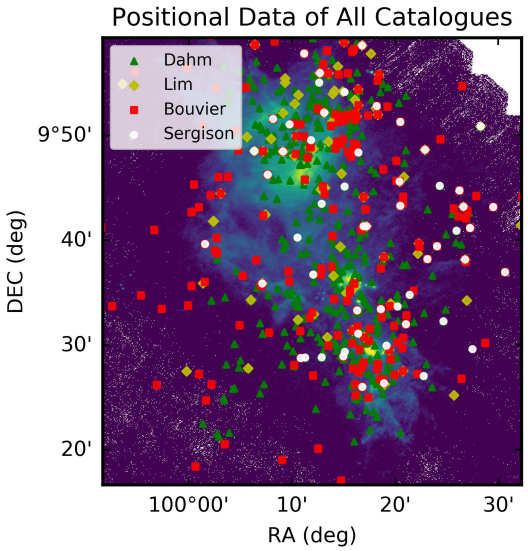
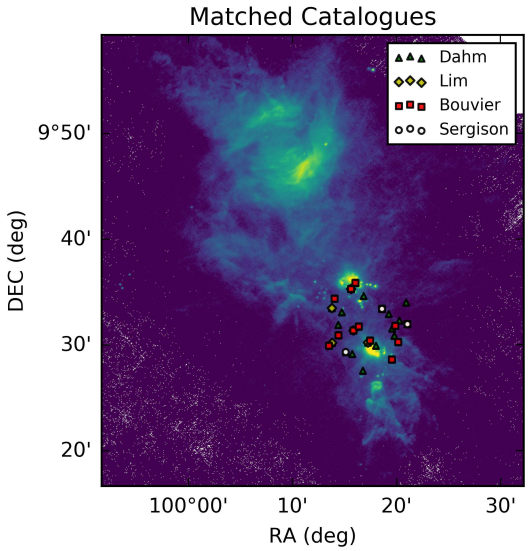


Figure 12: Li Equivalent Width vs. Colour, from samples of YSO's of different ages, source [Soderblom et al., 2014], fig. 8

However, high-resolution spectrographs are required in order to age objects to within tenths of millions of years, the timescale in which depletion can occur. While potentially lithium could be used



(a) Herschel image of NGC 2264 overlaid with tertiary sources



(b) Herschel image of NGC 2264 overlaid with matched tertiary sources

Figure 11: Comparison between raw and matched data

to determine whether some stars with large deviations in line width are in the foreground or background, rather than in the subcluster, this is of limited use for the amount of effort that it would have taken to implement and matching with MPCM data was used instead.

Additionally, for the cluster NGC 2264, there are only a limited number of stars that had lithium equivalent line widths, and matched with the Mariñas survey [Mariñas et al., 2013], which was the only survey found with stars that had been spectroscopically aged. In addition, the Mariñas survey only contained stars from the southern region of NGC 2264, which would have caused inaccuracies with bootstrapping ages to the more populous northern region.

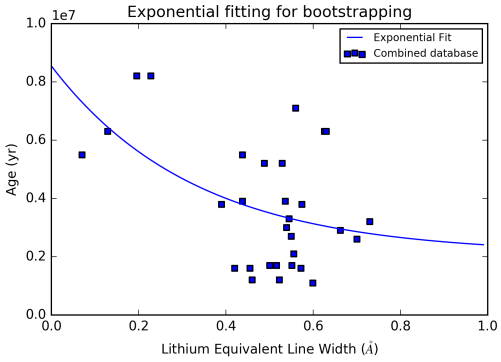


Figure 13: An attempt at exponential fitting to find a direct correlation between lithium and age

Figure 13 shows a sharp drop in equivalent line width with age, however a lack of stars with both age values and lithium equivalent line widths meant that there was insufficient data for adequate fitting. The cluster environment itself may have prevented addition sources from being detected. In the future however, more sensitive telescopes with higher resolution spectrographs may result in a greater number of sources being detected.

3.2.4 H^{α} Line Width

A number of surveys record H^{α} emission, which can be used as an indicator of relative age similar to lithium.

However, there is no adequate fitting between H^{α} equivalent line width and age in this cluster, as well as having the same problems as lithium ageing had. With even less data leading to no detectable correlation.

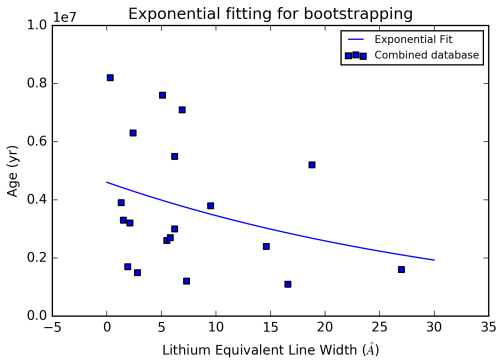


Figure 14: An attempt at exponential fitting to find a direct correlation between H^{α} and age

As estimating age directly was impractical considering the available data, an attempt was made

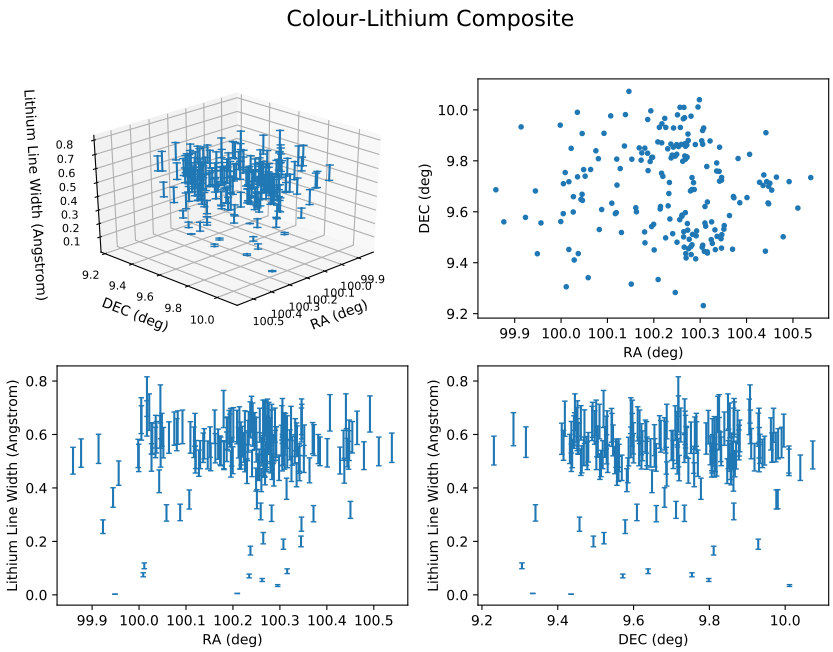


Figure 15: Composite plot of RA/DEC position and lithium equivalent line width, while there is a noticeable correlation between position and lithium abundance, there are not enough outliers to be considered useful

to determine a correlation between position and lithium abundance (figure 15), while some outliers were found, there were not enough to attempt subclustering, subsequently this method was not pursued further due to time limitations.

3.2.5 Alternative Ageing Methods

Due to the lack of ageing data in the northern region of the object, and an inability to bootstrap ages using markers, an attempt was made at ageing through evolutionary model fitting.

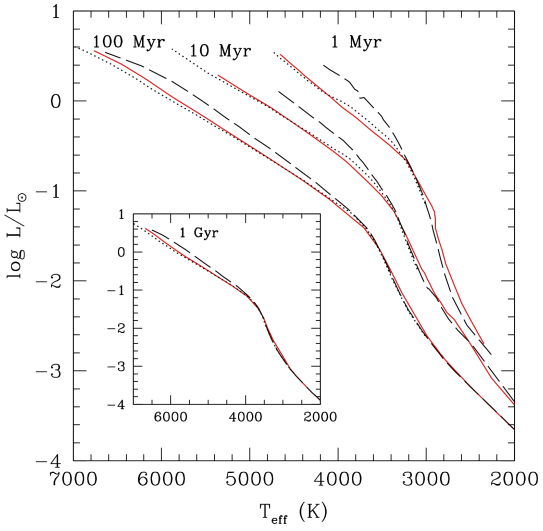


Figure 16: An example of change in luminosity as a function of age, source [Baraffe et al., 2015]

Evolutionary model fitting, also called isochrone fitting, is one of the most common methods of ageing stars [Soderblom et al., 2014]. Models are typically versatile, capable of being used with a wide variety of input data such as metallicity, colour, effective temperature and lithium abundance.

For young stars, isochrone fitting is especially useful, as properties change systematically during pre-main sequence contraction. However, while ageing young stars is a topic with a high degree of interest, due to the presence of dense molecular clouds there are still many unknown processes behind star formation, especially in the case of MYSOs. This means that the properties may not be monotonic, with similar properties across multiple ages, in order for an accurate age, multiple independent properties and models should be utilised.

The BHAC 2015 model [Baraffe et al., 2015] is an evolutionary model that supersedes the previous BCAH98 model [Baraffe et al., 1998] by using a more accurate, contemporary atmospheric model and 3D radiative hydrodynamic simulations. BHAC 2015 models stars in the range 0.01 to 1.4 $M_{\odot}$ on timescales from 0.5Myr to 1 Gyr. While using multiple models improves the accuracy of the estimate, for an initial assessment a single model was considered adequate.

In order to accomplish this, isochrones from the BHAC 2015 model were downloaded[5] and compared against the Bouvier survey dataset. J-K data

[5] http://perso.ens-lyon.fr/isabelle.baraffe/BHAC15dir/BHAC15_iso.ukidss

was added to the Bouvier dataset from the WFCAM science archive and bandpass corrections were applied using the models derived by Schlafly and Finkbeiner [Schlafly and Finkbeiner, 2011], dust maps were acquired from the NASA/IPAC Infrared Science Archive[6].

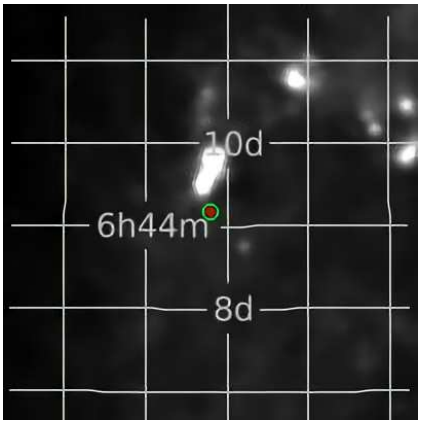


Figure 17: Dust emission at 100μm, from NASA/IPAC Infrared Science Archive

Mean dust emission (MJy/sr)	37.18 ± 2.26
Mean dust temperature (K)	16.66 ± 0.03
E(B-V) reddening (mag)	1.21 ± 0.07
UKIDSS J filter λ	1.248
J filter extinction (mag)	0.709
UKIDSS H filter λ (m)	1.659
H filter extinction (mag)	0.449
UKIDSS K filter λ	2.190
K filter extinction (mag)	0.302

Table 4: Dust emission and correction parameters for NGC 2264

Figure 18 is a composite of multiple regions in a colour-temperature HR diagram formulated from the BHAC 2015 model. It was found that the temperature uncertainty was too high to accurately perform isochrone fitting. In addition, the simple colour correction may not have been sufficient for a region of varying extinction, other properties garnered a similar result.

In addition, as BHAC 2015 only simulates low-mass stars, at odds with NGC 2264's abundance of high-mass stars, this mandated that additional models must be used, however due to time constraints this could not be adequately accomplished and this study was not pursued further.

Outliers

It was suggested that outliers in a colour-abundance plot may have a spatial relation, as a similar lithium content and reddening would suggest young stars that are contained within the same molecular cloud. These regions of matching lithium, colour and positions would then be used as centroids in a seeded subclustering algorithm.

In order to accomplish this, outliers were to be first sought out, and then compared positionally to determine if they were spatially near each other. Dust correction was applied to sources containing lithium and H^{α} equivalent line widths, which were then plotted against each other.

Figure 19a shows that there are few significant outliers that were later determined to not correlate spatially. The same was found for lithium abundance (figure 19b), while there was a clear (and expected) correlation between lithium abundance and J-K colour, but no groups of noticeable outliers in the data.

Kinematic ageing was not attempted, due to a lack of radial data and a lack of time remaining in the project, as the process is extremely involved.

3.3 Extending Previous Project Subclustering Algorithms

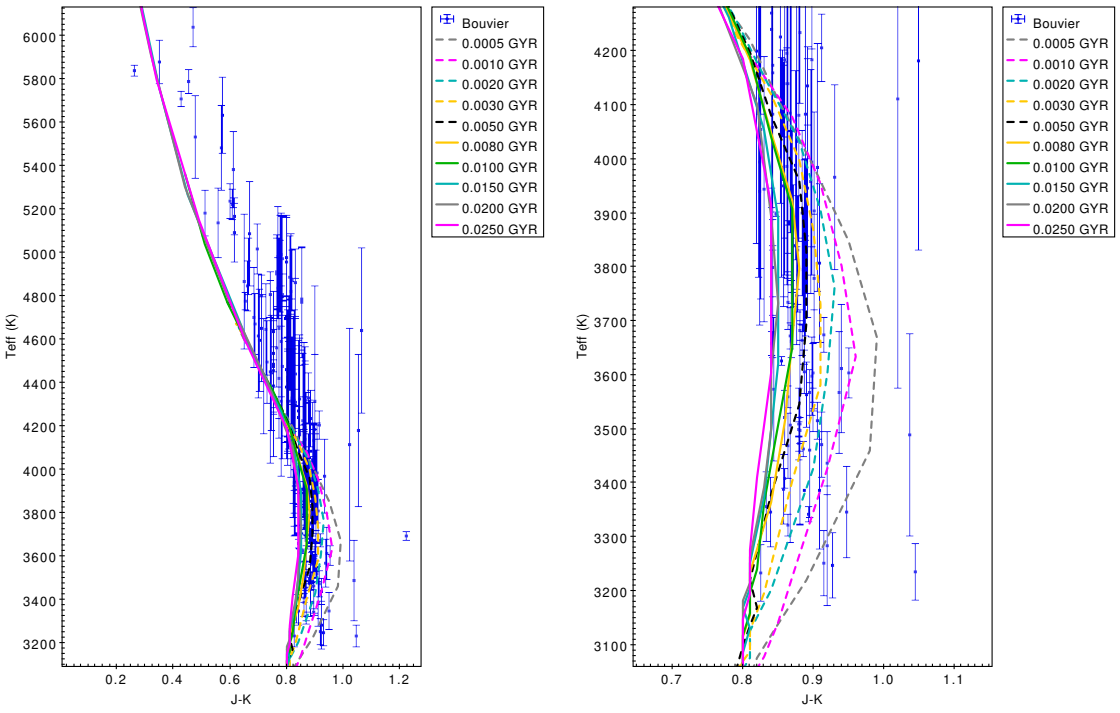
It was decided that a continuation of the subclustering done by Gorton and Boyer in previous years would be performed, by addressing a number of issues specified in their reports.

The primary issue with the previous projects is the degree of visual inspection required for subclustering, and a lack of automation. Gorton notes that the initial DBSCAN requires fine tuning on a per-cluster basis, a subroutine that optimises DBSCAN parameters (`eps` and `MinPts`) before the algorithm is utilised would allow for more rapid, initial subclustering of clusters. While more involved methods such as supervised finite mixture models could be used for greater accuracy, an automated method would be useful for an initial assessment; and could be used to provide initial parameters for a supervised method.

Another concern is the lack of available spectral data at the time; Boyer noted that simplistic colour cuts had to be used in order to determine members, this would inevitably lead to a number of false positive cluster members. Instead, using a well-established cluster catalogue such as MPCM removes this concern. In addition, dust emission can be used to detect embedded objects using Herschel 70μm data.

A third concern was the lack of advanced features, such as subcluster nesting, this is an inherent flaw of unsupervised models utilising DBSCAN and k-means, and cannot be corrected without utilising methods beyond the scope of this project.

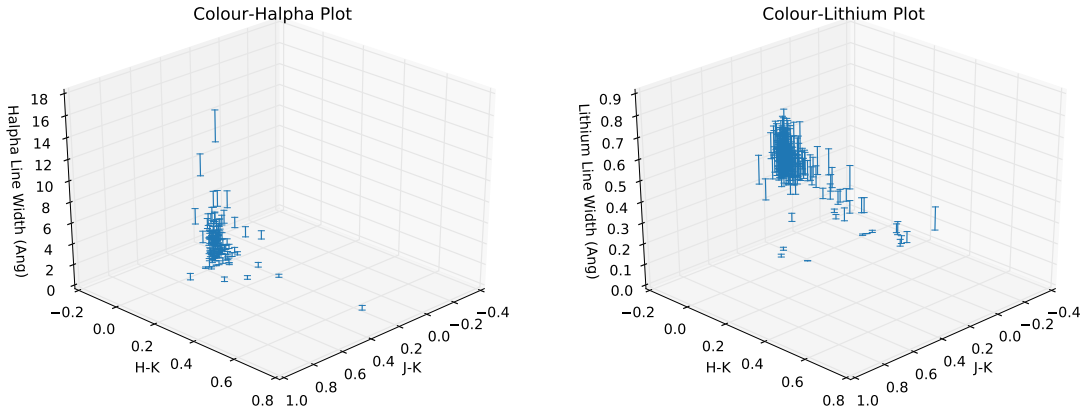
[6] https://irsa.ipac.caltech.edu/applications/DUST/



(a) Colour-temperature diagram with isochrones from the BHAC 15 model, high uncertainty prevents adequate fitting of isochrones

(b) Cropped section of HR diagram, the extremely high uncertainty prevents even low temperature objects from being fitted

Figure 18: Comparison between Bouvier survey and BHAC 15 protostar model



(a) Colour vs. H^{α} line width

(b) Colour vs. lithium line width

Figure 19: Attempts at finding significant abundance/colour outliers

4 Targets

4.1 NGC 2264

NGC 2264 was chosen to test the subclustering algorithm as it is fairly close ($2.6kpc$), well documented by many surveys (including the MYStIX survey), and is a young cluster, with a median age of 3Myr [Dahm, 2008], because if its age, NGC 2264 contains a large population of YSOs, in addition to more massive, embedded, main sequence stars.

As the cluster is quite young, it still has a high dust content, and hence has a high level of extinction within the cluster, this was especially noticeable in the northern region of the cluster, where the dust emits radiation strongly in the far-infrared bands. This can be largely corrected by using dust maps and extinction models, this may however still effect the detection of deeply embedded objects, as dust emission may result in difficulties finding the actual position of the embedded object.

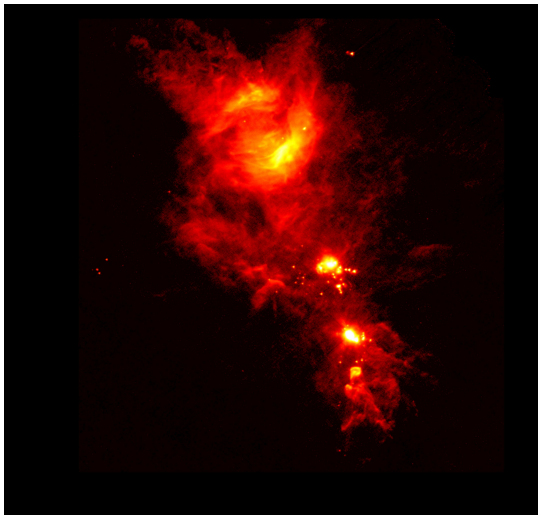


Figure 20: An image from the Herschel telescope in the 70μm band of NGC2264

4.2 Hyades

The Hyades cluster was chosen to test the 3-D 2-stage subclustering method as it is extremely close, and had an average distance uncertainty $\approx 10\%$ of the tidal radius of the cluster using TGAS data, which was considered acceptable for positional subclustering of the region.

Figure 21: HST image of Hyades cluster, credit NASA

Additionally, the Hyades cluster is fairly old for an open cluster, approximately 625Mr [Perryman et al., 1997], meaning that it is fairly dust-depleted, while colour data is not used at all, only RA/DEC and parallax positional data, this means that there are no embedded massive objects not being detected, or causing dust emission that obscures smaller, dimmer objects. However, because of the clusters age, stars above $\approx 3M_{\odot}$ are now off the main sequence [Tremblay et al., 2012] and may not be detected in the Tycho survey, meaning they would not be included in the TGAS dataset. This may cause the cluster to have a lower density of stars than anticipated.

5 Multi-Stage Seeding

In order to address the concerns of previous projects, a 4-stage, 2-mode subclustering method was devised, by subclustering a precursor dataset first, and using the centroids[7] to seed the final dataset.

As k-means clustering is computationally hard, and works faster and more accurately with less noisy data, DBSCAN is used to remove non cluster members, and estimate the number of subclusters, as k-means explicitly requires the number of centroids to be used, if seeds are not available[8]

1. Initial DBSCAN of precursor dataset
2. K-means subclustering of precursor dataset
3. DBSCAN of MPCM database to remove non-subcluster objects
4. Final kmeans with centroids seeded with precursor database subcluster centroids

[7] Subcluster centres

[8] In the case of seeds being provided, the number of centroids does not change, the seeds themselves are moved

The Herschel 70μm data was manipulated using Gaia and software in the Starlink suite, `figaro`, `medfilt`, and `isub` in particular, to generate a rudimentary point source dataset by estimating the centres of dust emission regions. This dataset acted as the precursor (figure 22), as it was deemed to be accurate enough, as dust emission is a direct result of the interaction between an embedded star and its molecular cloud.

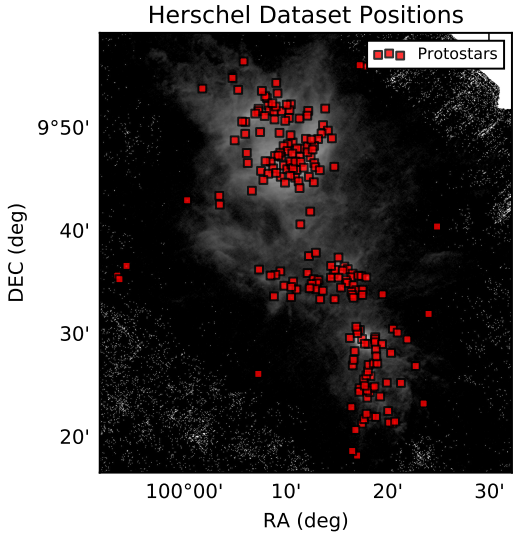


Figure 22: Overlay of Herschel point source estimate precursor onto Herschel 70μm image of NGC 2264

5.1 4-stage Subclustering in R

The statistical language R was used to perform subclustering, as it includes native packages for performing both the DBSCAN and k-means subclustering algorithms. While additional performance using a compiled language such as C could be used, the relatively small datasets meant that this was not necessary, however for subclustering of larger datasets, such as for globular clusters, this method may not have proved adequate.

Before subclustering could begin, ϵ values had to be generated for Herschel and MPCM data. This was performed by recursively comparing distances between objects within the cluster, out of which the nearest neighbours of each star are added to an array, then sorted and plotted, the point on the graph where the distance rapidly increases is taken as the value for ϵ. This technique was based on a novel method for determining ϵ, and was altered to be more accurate with smaller datasets by adding only the first nearest neighbour, rather than the 3 nearest neighbours per object [Rahmah and Sitanggang, 2016]. While it still requires some visual inspection, it is much faster than comparing points on a 2-D plot. The script used to find ϵ can be found in section 9.3.

Dataset	*Members*	*EPS*
MPCM	1173	0.0087
Herschel Dust Emission	217	0.0222

Table 5: Catalogue members and EPS comparisons for MPCM and Herschel point source estimate

DBSCAN is then performed on each database for the purpose of noise reduction. In each instance, the `eps` value is derived from the python code, and `MinPts` is derived from the density of the cluster, and found by visual inspection. For less dense data such as the precursors, 4 was chosen as a minimum number of links for a subcluster, for MPCM, which has significantly more members, 6 was chosen, in addition, the eps value was always significantly lower, due to the density.

Using this script DBSCAN appends an integer value to each cluster member, denoting which perceived subcluster it is in, all members with a 0 value are removed, as it denotes the lack of cluster membership.

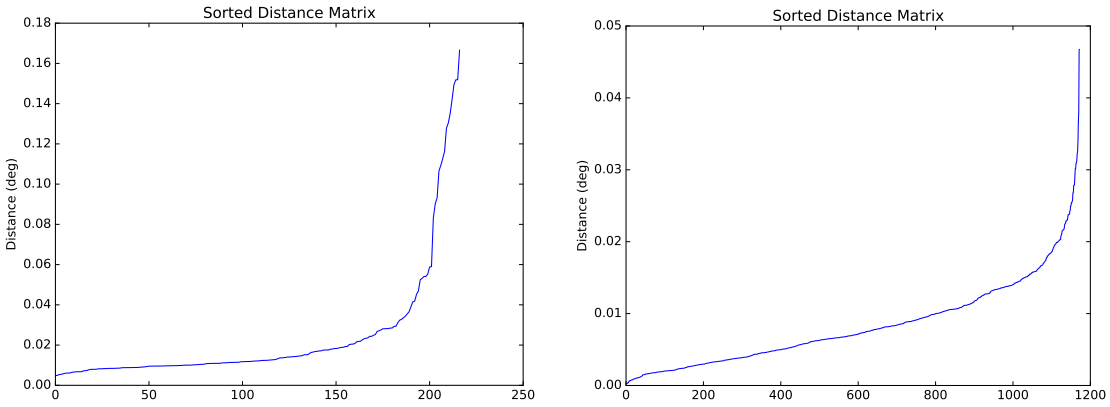
After noise reduction, k-means is applied to the Herschel 70μm data, 6 randomly placed centroids were initially chosen to adequately cover the northern and southern regions of the cluster.

After this, k-means subclustering was applied to the MPCM dataset, with the randomly placed centroids replaced with the list of final centroids from the Herschel precursor k-means subclustering.

After processing, the data was exported to CSV format, for contour plotting and addition of an image overlay. Output files included DBSCAN subcluster groups, individual k-means outputs for each subcluster, in order to generate contour plots for each one, and the k-means centroid locations for both precursor and MPCM datasets, in order to show the "drift" between centroids before and after the 4th stage. A full version of the 4-stage subclustering sript can be found in section 9.4.

Finally contour fitting and plotting is performed by a python script (section 9.5).

Figure 25 shows the result, using the Herschel point source estimate data as a precursor. Reasonable subclusters have been produced for the southern region, coinciding with the brighter regions of the cluster; this suggests a correlation between regions where star formation is still underway and current subclusters. This is sensible, as NGC 2264 is a fairly young cluster, using dust emission data may be sensible for use with a young cluster, however it may be less accurate for older clusters, where dust has been largely depleted by star formation, and a different precursor may be required.



(a) Sorted distance matrix of Herschel precursor dataset

(b) Sorted distance matrix of MPCM subclustering dataset

Figure 23: Sorted distance matrices, used to calculate ϵ values for DBSCAN noise reduction in 4-stage subclustering algorithm, ϵ cut-off shown in green

A major issue with this dataset is false positives due to bright, extended dust emission, this can be clearly seen in the northern region, where all centroids undergo significant drift, and in the central region, where a subcluster has moved significantly. This problem was mitigated by using a true protostar catalogue, and is covered in section 5.2.

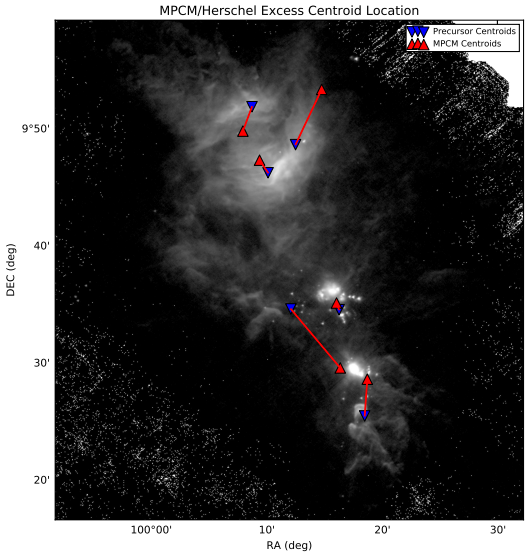


Figure 24: Centroid position drift between 2nd and 3rd stages, drift is extremely high for northern region, implying that the Herschel point sources estimate is not ideal as a subclustering precursor

5.2 Seeding Using Embedded Protostars

The Herschel-derived dataset was found to include dust emission, and lead to excessive, erroneous drift of the centroid locations, especially considering that an alternative precursor was available. A preliminary dataset containing protostar locations in NGC 2264 produced by Dr. Anne Buckner was used. This database was a significant improvement compared to the rudimentary precursor, and allowed for more accurate subclustering, despite the fact that the data is preliminary and incomplete.

Dataset	*Members*	*EPS*
Buckner Protostar Catalogue	124	0.0559
Buckner Southern Region	82	0.0117

Table 6: Catalogue members and EPS comparisons for Buckner catalogue

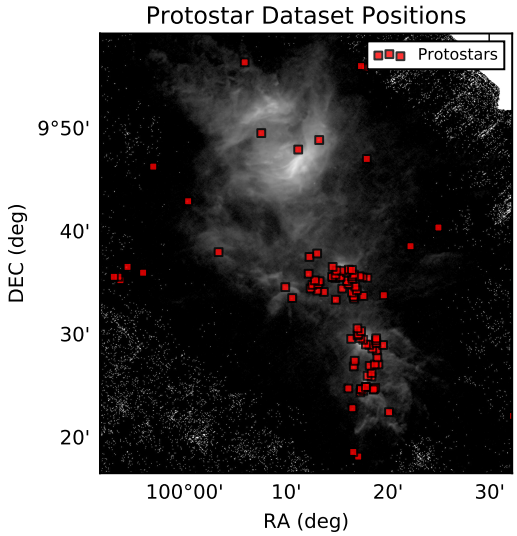


Figure 26: Overlay of protostar locations onto Herschel 70μm image of NGC 2264, the lack of available data in the northern region can be clearly seen

Additionally, due to extremely bright dust emission in the northern region of NGC 2264, it was decided that more focus should be made on the southern region, MinPts was dropped to 3, and a subset

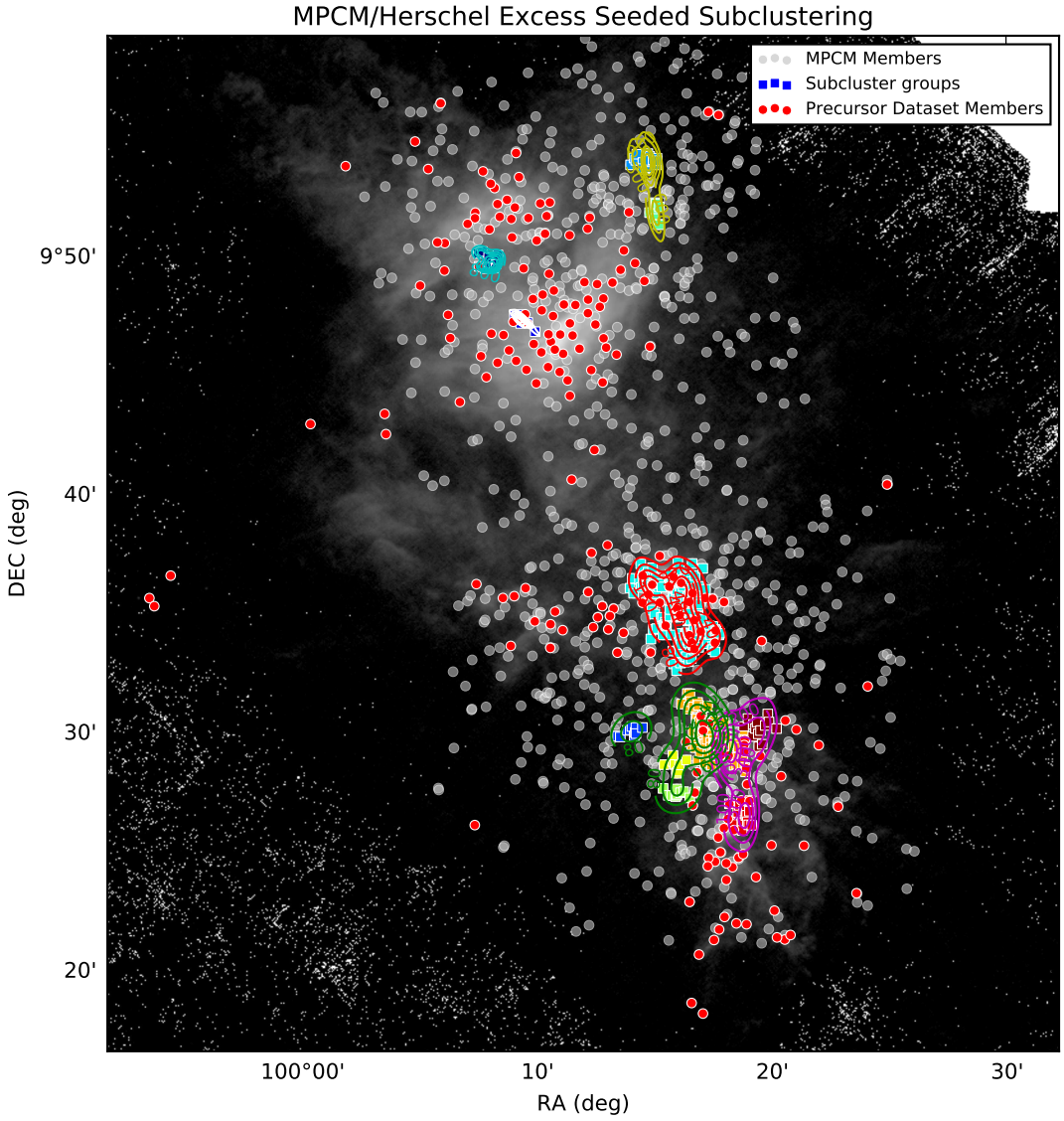


Figure 25: Subcluster contour plot of NGC 2264, using Herschel estimate catalogue as precursor dataset

of the MPCM data was made that only included data from below DEC 9°50′.

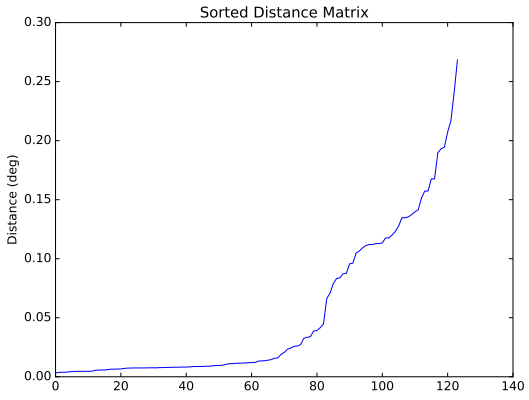


Figure 27: Sorted distance matrix of protostar catalogue precursor, fewer members meant that the slope increased rapidly, however the changes made to the ϵ calculation from [Rahmah and Sitanggang, 2016] meant that a suitable value was generated

As the code produced was fairly robust, all that had to be replaced were the ϵ values and the precursor dataset in order for subclustering to be commenced, fulfilling the goal of attempting to improve automation of the subclustering algorithm.

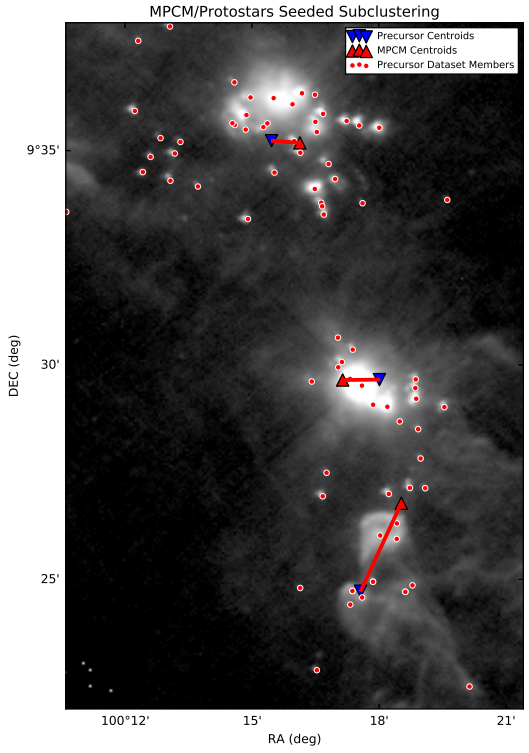


Figure 28: Centroid position drift for Buckner protostar dataset precursor, drift is significantly reduced

As can be seen in figure 28 the centroid drift is drastically reduced for each subcluster, as the precursor positions rooted in actual protostars, rather than just a rudimentary estimate. Figure 29 suggests that an immediate link between MPCM location and protostar location, which was expected due to the age of the cluster, as previously stated.

6 3-D Subclustering

As subclustering using Gaia parallax data for nearby clusters such as the Hyades was considered feasible, a reduced 2-stage version of the above programme was performed on the cluster data. No precursor data was used, as the main goal was to achieve subclustering through purely spatial means.

The standard set of libraries used in the 4-stage algorithm was declared and the data was imported.

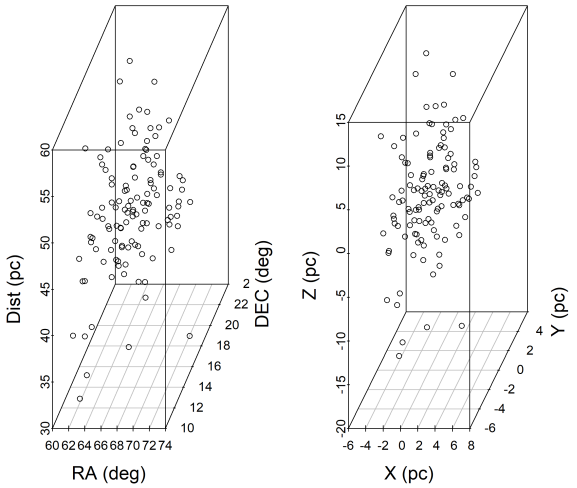


Figure 30: Before and after comparison of scaling, since a single ϵ value is used, all axes should use the same unit

In order to achieve subclustering each axis has to be roughly equivalent in size, for instance using an absolute distance metric would mean that the distance would have significantly more weight compared to right ascension and declination, hence all units are converted into parsecs using basic trigonometry, and the dataset was centred using the `scale` command (figure 30). After scaling, a modified version of the python ϵ calculator was used, that calculated the euclidean distance between each star in 3-D. ϵ was determined to be 1.681pc; comparisons should not be made to the 4-stage ϵ values due to the difference in units as well as the significantly lower density of the cluster due to a lack of complete data in the TGAS catalogue.

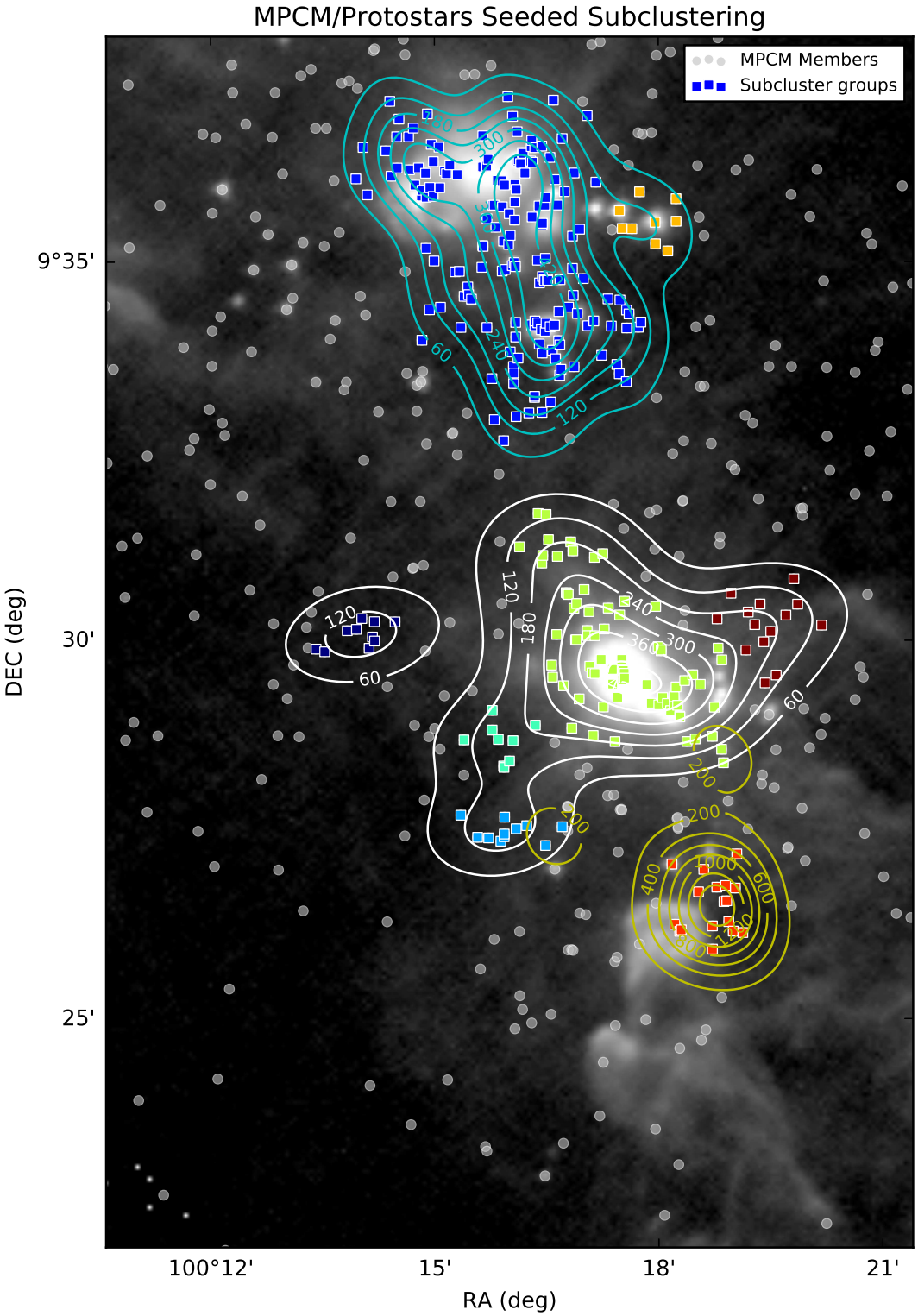


Figure 29: Subcluster locations of NGC2264, using protostar catalogue as precursor dataset

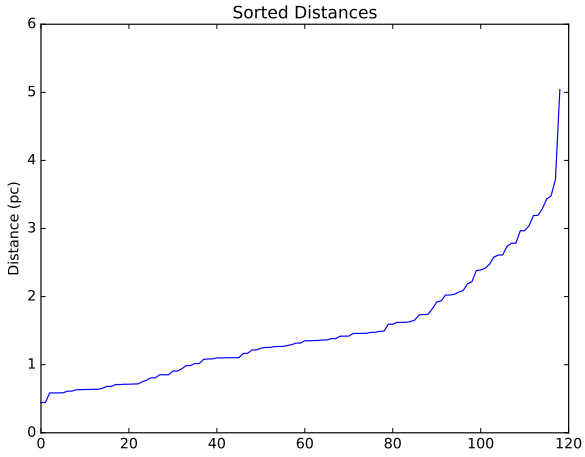


Figure 31: Sorted distances for Hyades 3D subclustering in order to determine ϵ

DBSCAN is performed on the cluster to reduce noise (non-subcluster objects), afterwards k-means is applied to the reduced noise dataset, the number of clusters in k-means is the same as the perceived number of clusters found in DBSCAN, this differs compared to the 4-stage algorithm, where n was found through visual inspection.

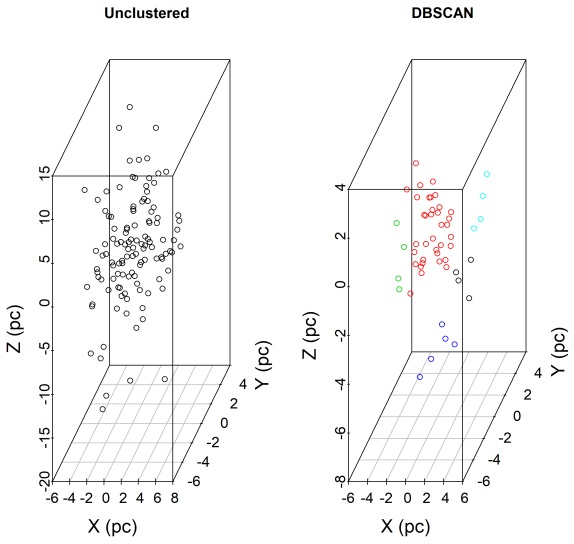


Figure 32: Comparison between scaled dataset and dataset after DBSCAN applied

Subset	Members
TGAS Data	119
DBSCAN Data	53

Table 7: Comparison of subset of Hyades TGAS data with data processed through DBSCAN

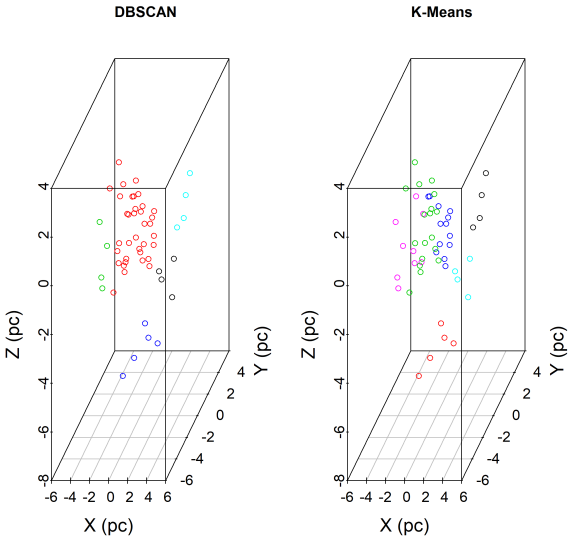


Figure 33: Comparison of DBSCAN and k-means subclustering, k-means is the main subclustering algorithm, while DBSCAN is used for noise reduction

Data was then plotted as a 3-D graph, as well as 2-D comparisons of each axis (figure 34), a modified version of the code used to generate the initial parallax plots (figure 6) was used, with subclusters identified using colours; ellipses or contours were not used in order to provide more visual clarity to the plot.

As can be seen in figure 34, clear subclusters arise from the data, and noise reduction through DBSCAN has been extremely effective in accurately defining the borders of the cluster; as instead of using crude distance cuts, the subcluster region has formed approximately the same size as the tidal radius. This suggests that noise reduction through the use of DBSCAN is crucial for the removal of useless data, especially for the datamining of cluster and subcluster members from the future Gaia data releases.

While the mean associated error for the X and Y axes (derived from right ascension and declination) is extremely small ($\pm$0.00048pc), the same cannot be said for the distance uncertainty on the Z axis ($\pm$0.816pc), which is 49% of the value for ϵ. While this was considered acceptable, this puts a harsh limit on positional subclustering using Hipparcos/TGAS data, this value may be reduced with Gaia DR2 and subsequent releases however.

6.1 Kinematic Subclustering and the Moving-Cluster Method

Gaia may also allow for determining cluster members accurately by measuring their kinematic properties [Galli et al., 2017], for nearby clusters in the

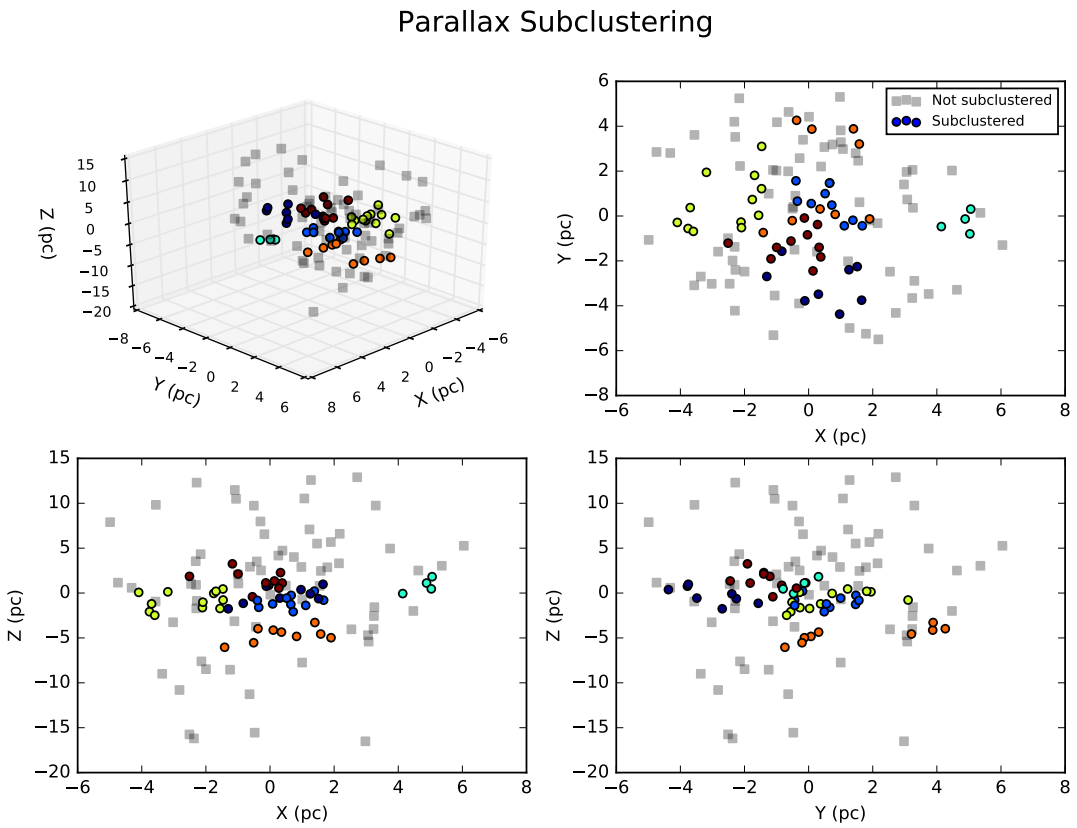


Figure 34: 3-D composite of subcluster positions within Hyades, subclusters are identified according to colour

future, this could potentially be used in old stellar clusters to find members of subclusters that are related through fragmentation, but not positionally due to drift of the stars from their molecular cloud. The moving cluster method was also used to determine subcluster members, in conjunction with direct motion through Doppler shift.

The moving-cluster method is a method to determine distance of star clusters using the convergence point of a group of stars, while was superseded by more accurate methods, however with the rise of high-resolution Doppler shift data from new spectrographs in Gaia and various ground based telescopes, combined with Gaia's high angular resolution the method may prove useful again as Galli's team displayed. Hypothetically clusters could be subclustered positionally even if direct observation of parallax proved infeasible, or it could be used to reinforce parallax observations with an independent method of calculating distance.

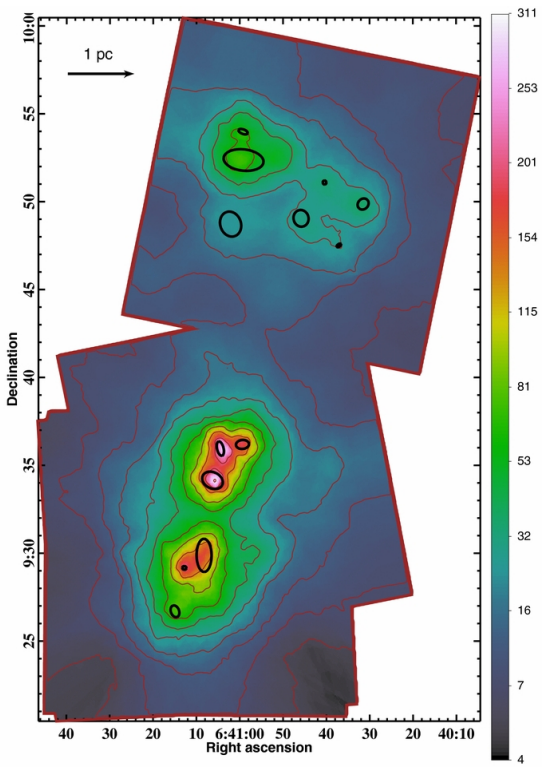
7 Discussion & Comparison with Literature

7.1 4-Stage Subclustering

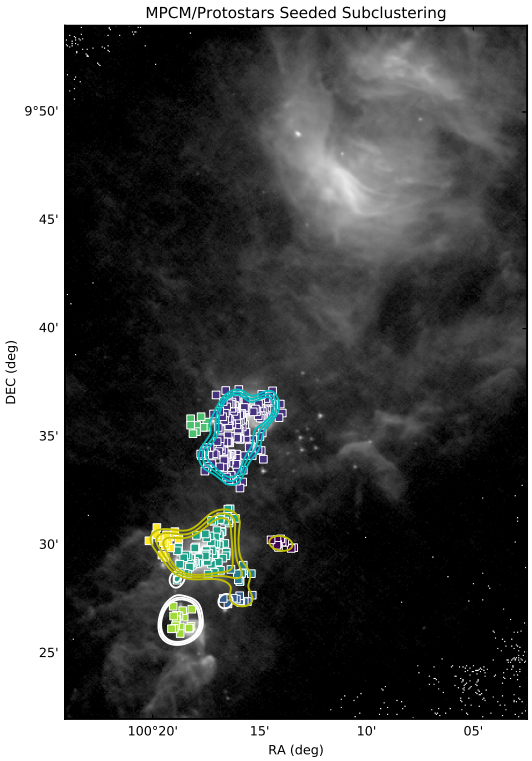
Both subclustering algorithms performed adequately given the high degree of reddening, particularly the incomplete protostar database. The main survey to compare the results of this project with is the MYStIX survey subclusters [Kuhn et al., 2014]. Figure 35 compares MYStIX isothermal ellipses with σ contours produced by the 4-stage subclustering method. The main distinction is that the northern region is not available, due to the incomplete precursor database, however the three k-means contours roughly coincide with the largest subclusters in the southern region.

Despite having subclusters in the northern region, 4-stage subclustering using the dust emission precursor dataset completely disagreed with the MYStIX survey subclusters. In addition, subcluster sizes were found to be extremely small (figure 25), with a low number of members and fairly nonsensical shapes. This suggested that the dust emission dataset was unsuitable for use as a precursor for the 4-stage method, additional evidence supporting this was the significant centroid drift between the 2nd and 4th stages, which can be seen in figure 24; this drift was drastically reduced with the protostar precursor.

A survey of primordial structure in NGC 2264 [Teixeira et al., 2005] confirmed that current star formation regions coincided with dusty regions within the cluster. This suggests that while the dust emission catalogue used in this project was not



(a) MYStIX subcluster isothermal ellipsoids, source [Kuhn et al., 2014]



(b) 4-stage protostar catalogue subcluster contours, at 3, 4 and 5σ levels

Figure 35: Comparison between MYStIX results and 4-stage subcluster methods

adequate for subclustering, a more refined version could potentially be used as a subclustering precursor, perhaps by measuring excess over various MIR frequencies, rather than just using 70μm data. An extension for this project could involve determining a robust common precursor for a variety of clusters.

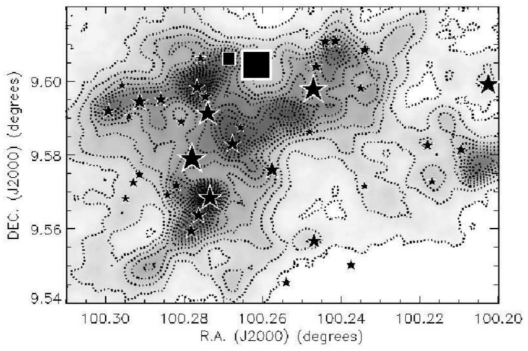


Figure 36: Correlation between dust emission and massive stars within NGC 2264, source [Teixeira et al., 2005]

Another explanation for the lack of feasible clusters within the northern region may be due to more significant differences in the environment of each region, as for subclustering, stellar density was assumed isotropic throughout the entire cluster. In reality stellar density is significantly lower in the region, meaning that subclusters only form in unusually dense parts of the northern region. This could be corrected by having slight variations on ϵ values between major regions of a cluster, this would also improve accuracy on larger clusters, with multiple large regions, and more varying cluster properties.

Furthermore, the multi-stage method could be improved by adding additional stages, and by subclustering using correlations such as line emission. A promising additional stage could potentially be utilising the Gaia-ESO public spectroscopic survey dataset [Gilmore et al., 2012]. While this was considered beyond the scope of the project due to the limited time remaining, using common line emissions or radial velocities as a second precursor prior to final subclustering could potentially yield vastly improved results. Finally, in order to determine the robustness of the method, additional testing in differing cluster environments, such as those envisioned with spatial subcluster, should be performed if this project is to be pursued further.

7.2 Spatial Subclustering

The spatial subclustering project shows promise, without TGAS, and the significantly reduced uncertainties as seen in figure 5, this method would not be possible, and with the additional reduction in uncertainty following DR-2 and subsequent data releases, additional clusters may be analysed in such a manner. Using Hipparcos data, attempts

were made to map the Hyades cluster in 3-D, however no subclustering was accomplished [Perryman et al., 1997]. Due to the recent release of the TGAS dataset, and the relative lack of interest in the field of spatial subclustering, no direct comparisons can be made with available literature the method devised.

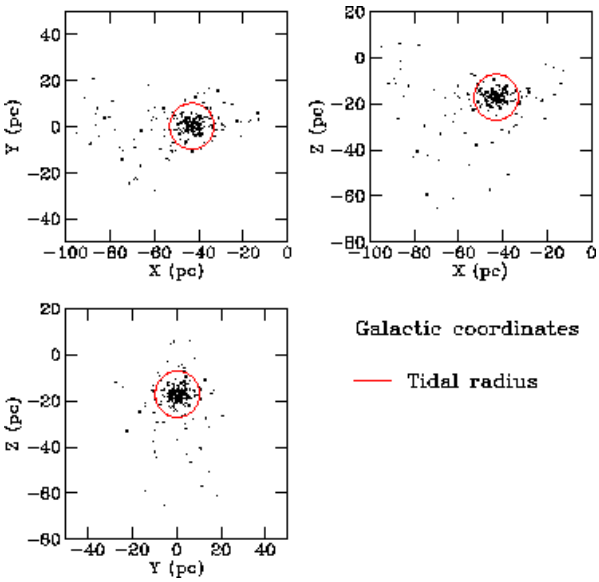


Figure 37: 3-D mapping of Hyades cluster using Hipparcos data, similarities between this figure and 34 can be seen, however there is no direct comparison as subclustering was not attempted, source [Perryman et al., 1997]

An additional survey of the Hyades was performed using secular parallax [de Bruijne et al., 2001], by using the motion of the solar system as it travels through the galaxy, long term surveys can be used to provide a significantly larger baseline for parallax measurements, however the relative velocity of stars must be known in order to derive accurate results. For stellar clusters, which have a relatively low velocity dispersion, this uncertainty due to stellar movement in this baseline can be reduced by measuring groups of stars in a cluster [Popowski and Gould, 1997].

Subclustering was not in the secular parallax survey, it is clear that the same regions of the cluster appear, despite utilising different coordinate systems the same filaments, and concentration of stars in the central region can be found (figure 38). Secular parallax was not considered for this project, hopefully the use of a combination of Tycho and Gaia positional data, in conjunction with accurate radial velocity metrics could result in accurate parallaxes for regions in a stellar cluster at a distance, rather than for individual stars. This could act as a stopgap until additional Gaia data is released, or in conjunction with traditional parallax measurements at a distance, where uncertainty is higher.

The subclusters within the Hyades display both ellipsoidal and elongated filamentary morphologies, consistent with accepted theories of cluster evolution. While dust emission was not compared with the subclusters produced, additional stages could be introduced to the spatial method utilising 2-D overlays as a continuation of this project, for instance, using spectroscopic data to determine abundances to match stars through relative age. In addition, radial velocities from the Gaia-ESO dataset could be utilised. However these were not attempted, due to a lack of available data, as well as being beyond the scope of the project. As with the 4-stage method, additional clusters should be analysed, when parallax data for more distant stellar clusters becomes available.

8 Conclusion

Subclustering at its heart is a largely subjective endeavour, relying on many disparate, independent methods to achieve the same goal. While combining multiple independent methods of subclustering was not attempted, doing so in the future, with more precise data may yield significantly improved results.

The main issue with this project was the severe lack of available data for more exotic methods, such as lithium abundance data. The main drawback of the spatial subclustering method is the lack of data due to the unavailability of pure Gaia parallax data, which will become available with DR-2 next year. However, despite the lack of general data, the subclustering methods produced are both novel, and in instances where they can be compared, compare favourably to more involved methods, such as the subclustering method used by MYStIX .

The 4-stage method closely matched the MYStIX subclustering model in the southern region of NGC 2264, and showed an improvement over Gorton's 2-stage subclustering model, for young clusters still undergoing star formation, with thick molecular clouds, using a precursor dataset containing star formation regions to bias these subclusters may yield more realistic subcluster borders. By combining this technique with spectroscopic ageing techniques using new equipment such as the E-ELT and Gaia-ESO to determine the age of stars using lithium spectra ageing techniques could also yield an additional improvements as soon as the data becomes available. Additionally using dust emission as a precursor for younger clusters may be viable if additional work is put into it, which was not available due to time constraints.

Subclustering in part utilising 3-D spatial data such as from the Gaia survey has a great deal of

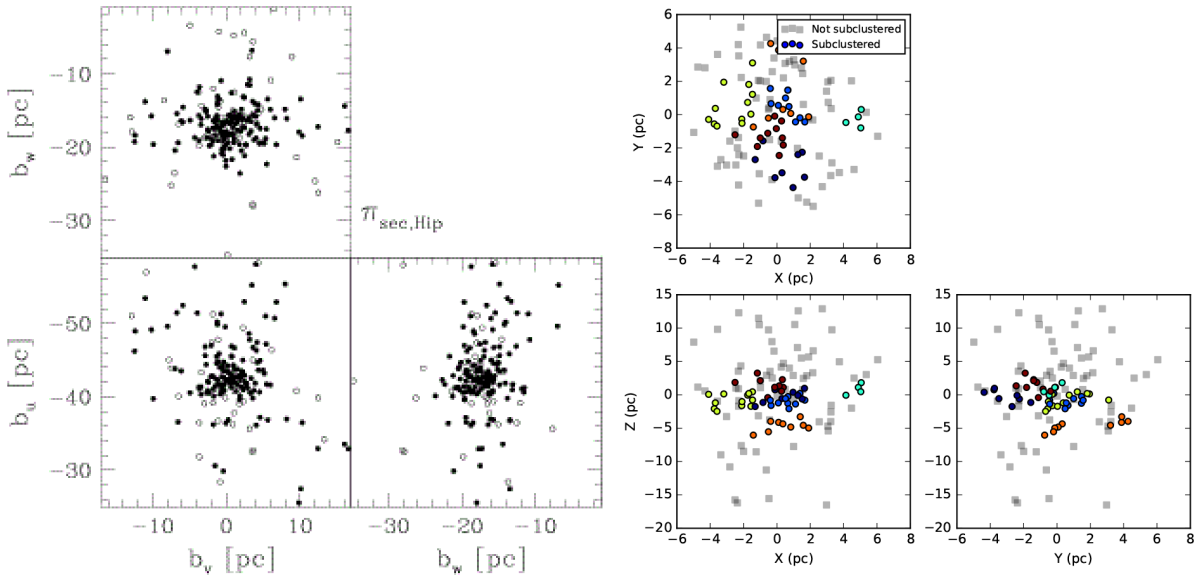


Figure 38: Comparison between Hyades 3-D survey using secular parallax [de Bruijne et al., 2001] and parallax subclustering, secular parallax survey uses Galactic Cartesian coordinates b_u, b_v, b_w, while survey conducted in this paper uses relative coordinates, however, the overall structure is in agreement with the literature

promise, as the limits of TGAS such as high uncertainty beyond 100pc, will be mitigated. For nearby stars, even simple subclustering algorithms such as the 2-stage method used provide convincing results when subclustering in 3-D, combined with additional methods, such as ageing data or dust emission maps, in the case of the 4-stage method used, these subclusters can be "biased" and better defined. "Hybrid" subclustering methods may prove useful, but with currently available data this method may prove to be fairly inconsistent, as can be seen with the various feasibility studies performed on NGC 2264.

Each of the two methods has their own benefits and drawbacks, suited to their respective distance and density regimes, with the direct observation of the 3-D spatial mapping technique being more suited to close range clusters, while the 4-stage, precursor dataset method being used for more distant clusters, assuming that they have a great deal of multi-spectral information available.

Ultimately, the imminent release of Gaia DR-2 suggests that the spatial subclustering method is the most promising, with a further reduction in parallax uncertainties from TGAS, distance uncertainty below 10% of the cluster radius may be possible for additional clusters beyond 100 parsecs, potentially leading to significant advances in the field of subclustering. Additionally, by grouping small regions of stars and using secular parallax to determine distance, accurate spatial subclustering could be determined for larger subclusters. Finally, hybridising the 2 methods shown may provide additional benefits for young, dust filled clusters.

9 Appendix

9.1 Acknowledgements

I would like to thank my supervisor Dr Stuart Lumsden for providing significant support and guidance during this project, and patience during the more fruitless studies. In addition I would like to thank Dr Anne Buckner for her protostar catalogue, which was instrumental in producing the 4-stage subclustering method used, and much more potent than the rudimentary excess point source catalogue used prior. Finally I would also like to thank Pruthvi Mehta, for keeping me sane throughout the more complex parts of this project.

9.2 Properties of TGAS downloads

Cluster	RA	DEC	Dimensions	Tidal Radius	Distance
Hyades	66.725	15.866	$330'$	$\approx 10pc$ [Röser et al., 2011]	$\approx 45pc$
Blanco 1	1.029	-29.833	$90'$	$22.8 \pm 3.8pc$ [Piskunov et al., 2008]	$\approx 260pc$
Cygnus OB2	308.3	41.317	$60'$	$46 \pm 10pc$ [Knodlseder, 2000]	$\approx 1400pc$

Table 8: Properties of TGAS downloads

9.3 Python Script to Calculate ϵ

This programme was written in python as appending lists in R doesn't scale with constant time ($\mathcal{O}(1)$ complexity), meaning that for larger datasets, such as MPCM, a prototype written in R took 8 minutes 15 seconds, whereas python took 14 seconds. The script was run on a notebook with an i5-6300U CPU (2.4GHz, 2 cores 4 threads) and 8GB of RAM. However, with larger datasets, the code may need rewriting in order to scale well.

```
import numpy as np
import matplotlib.pyplot as plt
printcounter,totalcounter=0,0
inputs=["Mystix/MPCM.csv",
"Herschel/Hcatalog-Finalmembers-Rreadable",
"Seed Transfer/Protostars/protostars_readable.csv"]
outputs=["mystix-dmat.pdf",
"herschel-dmat.pdf",
"protostars-dmat.pdf"]
for k in range(len(inputs)):
    data=np.genfromtxt(inputs[k],delimiter=",")
    distances=[]
    for i in range(0,len(data)):
        cache2=[]
        for j in range(0,len(data)):
            cache2.append(np.hypot(data[i][0]-data[j][0],data[i][1]-data[j][1]))
            cache2.sort()
            printcounter+=1
            totalcounter+=1
            if printcounter==1000:
                print"Finished "+str(totalcounter)+" iterations"
                printcounter=0
        distances.append(cache2[2])
    n=range(len(distances))
    distances.sort()
    fig = plt.figure()
    plt.plot(distances)
    plt.title("Sorted Distance Matrix")
    plt.ylabel("Distance (deg)")
    plt.savefig(outputs[k],
                dpi=300,
                bbox_inches="tight")
    plt.show()
    print "Finished " + outputs[k]
```

9.4 4-Stage Subclustering Script

```
#Library Dependencies
library(fpc)
library(MASS)
library(cluster)
```

```
library(plyr)
library(scales)
n_centroids<-6
output_name<-"protostars"
eps_precursor<-0.0222
eps_mpcm<-0.0087
coordinates_precursor<-read.("HERSCHEL_COORDINATES.csv",
header=TRUE)
coordinates_mpcm<-read.csv("MPCM_COORDINATES.csv",
header=TRUE)

#DBSCAN to find initial number of clusters
dbs_precursor<-dbscan(coordinates_precursor,
eps=eps_precursor,
MinPts=5,
method="raw")
precursor_cluster_combo<-cbind(coordinates_precursor,
dbs_precursor$cluster)
precursor_cluster_true<-precursor_cluster_combo[!grepl(0,precursor_cluster_combo[,3]),]
cluster_coordinates_precursor<-cbind(precursor_cluster_true[,1],
precursor_cluster_true[,2])
#DBSCAN to find cluster members of MPCM
dbs_mpcm<-dbscan(coordinates_mpcm,
eps=eps_mpcm,
MinPts=8,
method="raw")
mpcm_cluster_combo<-cbind(coordinates_mpcm,
dbs_mpcm$cluster)
mpcm_cluster_true<-mpcm_cluster_combo[!grepl(0,mpcm_cluster_combo[,3]),]
cluster_coordinates_mpcm<-cbind(mpcm_cluster_true[,1],
mpcm_cluster_true[,2])

#K-Means to find seed centroid locations using Herschel Data
kmeans_precursor<-kmeans(cluster_coordinates_precursor,
n_centroids,
iter.max=100,
algorithm="Hartigan-Wong")
kmeans_precursor_members<-cbind(cluster_coordinates_precursor,
kmeans_precursor$cluster)
kmeans_precursor_centres<-kmeans_precursor$centers[,c(1,2)]

#K-Means to find subcluster centroid locations using MPCM data, seeded with Herschel data
kmeans_mpcm<-kmeans(cluster_coordinates_mpcm,
kmeans_precursor_centres,
iter.max=100,
algorithm="Hartigan-Wong")
kmeans_mpcm_members<-cbind(cluster_coordinates_mpcm,kmeans_mpcm$cluster)
kmeans_mpcm_centres<-kmeans_mpcm$centers[,c(1,2)]

#DBSCAN Cluster Members
fname<-paste("Outputs/",output_name,"-DBSCAN.csv",sep="")
write.csv(mpcm_cluster_true,fname)
#4stage cluster members
for (n in unique(kmeans_mpcm_members[,3])){
fname<-paste("Outputs/",output_name,"-kmeans-",n,".csv",sep="")
out<-kmeans_mpcm_members[grepl(n,kmeans_mpcm_members[,3]),]
write.csv(out,fname)
}
#Cluster Centroids
fname<-paste("Outputs/",output_name,"-precursor-centres.csv",sep="")
write.csv(kmeans_precursor_centres,fname)
fname<-paste("Outputs/",output_name,"-mpcm-centres.csv",sep="")
write.csv(kmeans_mpcm_centres,fname)
```

9.5 Subclustering Plotting Script

```
from astropy.wcs import WCS
import numpy as np
from astropy.io import fits
import matplotlib.pyplot as plt
from matplotlib.colors import LogNorm
from scipy import stats as st
title="MPCM/Herschel Excess Seeded Subclustering"
mode="centres" #mode can be members or centres
mpcm_centres=np.genfromtxt("../../Outputs/herschel-mpcm-centres.csv",
delimiter=",")
```

```
precursor_centres=np.genfromtxt("../../Outputs/herschel-precursor-centres.csv",
delimiter=",")
precursors=np.genfromtxt("../../Herschel/Hcatalog-Finalmembers-Rreadable",
delimiter=",")
members=np.genfromtxt("../../Mystix/MPCM.csv",
delimiter=",")
members_sub=np.genfromtxt("../../Outputs/herschel-DBSCAN.csv",
delimiter=",")
contour_name="herschel-kmeans-"
n_contours=6
filename="../../../Herschel/pacs70-cal.fits"
herschel_image = fits.getdata(filename, ext=0)
herschel_coordinates = WCS(filename)
hdu = fits.open(filename)[0]
wcs = WCS(hdu.header)
fig = plt.figure(figsize=(8,10))
ax=fig.add_subplot(111, projection=wcs)
trans=ax.get_transform("world")
ax.imshow(herschel_image,
cmap="gray",
norm=LogNorm(vmin=0.01,
vmax=1))

plt.xlim(1250,
500)
plt.ylim(850,
1650)
if mode == "members":
ax.scatter(members[0:,0],members[0:,1],
transform=trans,
alpha=0.6,
color="0.75",
linewidth="0.5",
edgecolor="w",
label="MPCM Members")
ax.scatter(members_sub[0:,1],
members_sub[0:,2],
c=members_sub[0:,3],
linewidth="0.5",
marker="s",
edgecolor="w",
transform=trans,
label="Subcluster groups")
ax.scatter(precursors[0:,0],
precursors[0:,1],
transform=trans,
color="r",
linewidth="0.5",
edgecolor="w",
label="Precursor Dataset Members")
cols=["w","y","c","m","g","r"]
for n in range(1,n_contours+1):
fname="../../Outputs/"+contour_name+str(n)+".csv"
clustlab="Cluster "+str(n)+" Contour"
contdata=np.genfromtxt(fname,delimiter=",")
contx=contdata[1:,1]
conty=contdata[1:,2]
xmin=min(members[1:,0])
xmax=max(members[1:,0])
ymin=min(members[1:,1])
ymax=max(members[1:,1])
xx,yy=np.mgrid[xmin:xmax:1000j,
ymin:ymax:1000j]
positions=np.vstack([xx.ravel(),
yy.ravel()])
values=np.vstack([contx.ravel(),
conty.ravel()])
kernel = st.gaussian_kde(values)
f = np.reshape(kernel(positions).T,
xx.shape)
cset=ax.contour(xx,yy,f,colors=cols[n-1],
transform=trans,
label=clustlab)
ticks=range(0,200,10)
ax.clabel(cset, inline=1,
fontsize=8,ticks=ticks,
```

```
fmt = '%1.0f',)

if mode == "centres":
for n in range(1,n_contours+1):
ax.arrow(precursor_centres[n,1],
precursor_centres[n,2],
mpcm_centres[n,1]-precursor_centres[n,1],
mpcm_centres[n,2]-precursor_centres[n,2],
head_width=0, head_length=0,
color="r",
transform=trans,
length_includes_head=True)
ax.scatter(precursor_centres[1:,1],
precursor_centres[1:,2],
s=10**2,
marker="v",
c="b",
label="Precursor Centroids",
transform=trans)
ax.scatter(mpcm_centres[1:,1],
mpcm_centres[1:,2],
s=10**2,
marker="^",
c="r",
label="MPCM Centroids",
transform=trans)

plt.title(title)
plt.xlabel("RA (deg)")
plt.ylabel("DEC (deg)")
plt.legend(prop={'size':8})
plt.savefig("../../HERSCHEL-SUBCLUSTER-FINAL.pdf",
dpi=600,
bbox_inches="tight")
plt.show()
```

9.6 3-D Subclustering Script

```
#Library Dependencies
library(fpc)
library(MASS)
library(cluster)
library(plyr)
library(scales)
library(readr)
library(scatterplot3d)

data<-read_csv("~/Research/stellar-clusters/Data/Subclustering/3D/hyades-omniplot.csv")
radec<-cbind(data$ra,data$dec)
radec<-0.0174533*radec
radecscale<-scale(radec,scale=FALSE,center=TRUE)
xy<-data$dist*tan(radecscale)
x<-xy[,1]
y<-xy[,2]
z<-scale(data$dist,scale=FALSE,center=TRUE)
xyz<-cbind(x,y,z)
write.csv(xyz,"3dout.csv")
#EPS determined using 3dout data and eps python script
eps<-1.681
dbs<-dbscan(xyz,eps=eps,MinPts=4)
cluster_combo<-cbind(xyz,
dbs$cluster)
cluster_true<-cluster_combo[!grepl(0,cluster_combo[,4]),]
cluster_false<-cluster_combo[grepl(0,cluster_combo[,4]),]
cluster_coordinates<-cbind(cluster_true[,1],
cluster_true[,2],
cluster_true[,3])
kmeans<-kmeans(cluster_coordinates,
length(unique(cluster_combo[,4])),
iter.max=100,
algorithm="Hartigan-Wong")
kmeans_members<-cbind(cluster_coordinates,kmeans$cluster)
write.csv(kmeans_members,"subcluster_members.csv")
kmeans_centres<-kmeans$centers

write.csv(kmeans_centres,"subcluster_centres.csv")
```

```
write.csv(cluster_false,"not_sub_members.csv")
for (n in unique(kmeans_members[,4])){
fname<-paste("contour-data-",n,".csv",sep="")
out<-kmeans_members[grepl(n,kmeans_members[,4]),]
write.csv(out,fname)
}
```

9.7 σ Calculation For 4-Stage Method Contours

σ is the standard deviation of the kernel density of the kde2d contour plot of each k-means subcluster, and is as follows:

Subcluster	Colour (Figure 35b)	σ
Subcluster A	Blue	43.72
Subcluster B	Yellow	37.55
Subcluster C	White	70.72

Table 9: Cluster contours and associated σ values

List of Figures

List of Tables

References

[Aarseth and Hills, 1972] Aarseth, S. and Hills, J. (1972). The dynamical evolution of a stellar cluster with initial subclustering. *Astronomy and Astrophysics*, 21:255.

[Baraffe et al., 1998] Baraffe, I., Chabrier, G., Allard, F., and Hauschildt, P. (1998). Evolutionary models for solar metallicity low-mass stars: mass-magnitude relationships and color-magnitude diagrams. *arXiv preprint astro-ph/9805009*.

[Baraffe et al., 2015] Baraffe, I., Homeier, D., Allard, F., and Chabrier, G. (2015). New evolutionary models for pre-main sequence and main sequence low-mass stars down to the hydrogen-burning limit. *Astronomy & Astrophysics*, 577:A42.

[Bate, 2012] Bate, M. R. (2012). Stellar, brown dwarf and multiple star properties from a radiation hydrodynamical simulation of star cluster formation. *Monthly Notices of the Royal Astronomical Society*, 419(4):3115–3146.

[Bouvier et al., 2016] Bouvier, J., Lanzafame, A., Venuti, L., Klutsch, A., Jeffries, R., Frasca, A., Moraux, E., Biazzo, K., Messina, S., Micela, G., et al. (2016). The gaia-eso survey: A lithium-rotation connection at 5 myr? *Astronomy & Astrophysics*, 590:A78.

[Boyer, 2016] Boyer (2016). Using combined near-infrared and mid-infrared observations to reveal substructure within massive star forming regions.

[Dahm, 2008] Dahm, S. (2008). The young cluster and star forming region ngc 2264. *arXiv preprint arXiv:0808.3835*.

[Dahm and Simon, 2005] Dahm, S. and Simon, T. (2005). The t tauri star population of the young cluster ngc 2264. *The Astronomical Journal*, 129(2):829.

[de Bruijne et al., 2001] de Bruijne, J. H., Hoogerwerf, R., and de Zeeuw, P. T. (2001). A hipparcos study of the hyades open cluster-improved colour-absolute magnitude and hertzsprung-russell diagrams. *Astronomy & Astrophysics*, 367(1):111–147.

[Ester et al., 1996] Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231.

[Feigelson et al., 2013] Feigelson, E. D., Townsley, L. K., Broos, P. S., Busk, H. A., Getman, K. V., King, R. R., Kuhn, M. A., Naylor, T., Povich, M. S., Baddeley, A., et al. (2013). Overview of the massive young star-forming complex study in infrared and x-ray (mystix) project. *The Astrophysical Journal Supplement Series*, 209(2):26.

[Galli et al., 2017] Galli, P., Moraux, E., Bouy, H., Bouvier, J., Olivares, J., and Teixeira, R. (2017). A revised moving cluster distance to the pleiades open cluster. *Astronomy & Astrophysics*, 598:A48.

[Gilmore et al., 2012] Gilmore, G., Randich, S., Asplund, M., Binney, J., Bonifacio, P., Drew, J., Feltzing, S., Ferguson, A., Jeffries, R., Micela, G., et al. (2012). The gaia-eso public spectroscopic survey. *The Messenger*, 147:25–31.

[Gorton, 2015] Gorton (2015). Subclustering in massive star forming regions.

[Knodlseder, 2000] Knodlseder, J. (2000). Cygnus ob2-a young globular cluster in the milky way. *arXiv preprint astro-ph/0007442*.

[Kuhn et al., 2014] Kuhn, M. A., Feigelson, E. D., Getman, K. V., Baddeley, A. J., Broos, P. S., Sills, A., Bate, M. R., Povich, M. S., Luhman, K. L., Busk, H. A., et al. (2014). The spatial structure of young stellar clusters. i. subclusters. *The Astrophysical Journal*, 787(2):107.

[Lim et al., 2016] Lim, B., Sung, H., Kim, J. S., Bessell, M. S., Hwang, N., and Park, B.-G. (2016). A constraint on the formation timescale of the young open cluster ngc 2264: Lithium abundance of pre-main sequence stars. *arXiv preprint arXiv:1608.07798*.

[Lloyd, 1982] Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

[Mariñas et al., 2013] Mariñas, N., Lada, E. A., Teixeira, P. S., and Lada, C. J. (2013). Near-infrared imaging and spectroscopic survey of the southern region of the young open cluster ngc 2264based on observations obtained at the kpno. *The Astrophysical Journal*, 772(2):81.

[Michalik et al., 2015] Michalik, D., Lindegren, L., and Hobbs, D. (2015). The tycho-gaia astrometric solution-how to get 2.5 million parallaxes with less than one year of gaia data. *Astronomy & Astrophysics*, 574:A115.

[Naylor et al., 2013] Naylor, T., Broos, P. S., and Feigelson, E. D. (2013). Bayesian matching for x-ray and infrared sources in the mystix project. *The Astrophysical Journal Supplement Series*, 209(2):30.

[Palla and Stahler, 1999] Palla, F. and Stahler, S. W. (1999). Star formation in the orion nebula cluster. *The Astrophysical Journal*, 525(2):772.

[Perryman et al., 1997] Perryman, M. A., Brown, A., Lebreton, Y., Gomez, A., Turon, C., De Strobel, G. C., Mermilliod, J., Robichon, N., Kovalevsky, J., and Crifo, F. (1997). The hyades: distance, structure, dynamics, and age. *arXiv preprint astro-ph/9707253*.

[Piskunov et al., 2008] Piskunov, A., Schilbach, E., Kharchenko, N., Röser, S., and Scholz, R.-D. (2008). Tidal radii and masses of open clusters. *Astronomy & Astrophysics*, 477(1):165–172.

[Popowski and Gould, 1997] Popowski, P. and Gould, A. (1997). Mathematics of statistical parallax and the local distance scale. *arXiv preprint astro-ph/9703140*.

[Povich et al., 2013] Povich, M. S., Kuhn, M. A., Getman, K. V., Busk, H. A., Feigelson, E. D., Broos, P. S., Townsley, L. K., King, R. R., and Naylor, T. (2013). The mystix infrared-excess source catalog. *The Astrophysical Journal Supplement Series*, 209(2):31.

[Rahmah and Sitanggang, 2016] Rahmah, N. and Sitanggang, I. S. (2016). Determination of optimal epsilon (eps) value on dbscan algorithm to clustering data on peatland hotspots in sumatra. In *IOP Conference Series: Earth and Environmental Science*, volume 31, page 012012. IOP Publishing.

[Ramsay et al., 2014] Ramsay, S., Casali, M., González, J., and Hubin, N. (2014). The e-elt instrument roadmap: a status report. In *SPIE Astronomical Telescopes+ Instrumentation*, pages 91471Z–91471Z. International Society for Optics and Photonics.

[Röser et al., 2011] Röser, S., Schilbach, E., Piskunov, A., Kharchenko, N., and Scholz, R.-D. (2011). A deep all-sky census of the hyades. *Astronomy & Astrophysics*, 531:A92.

[Schlafly and Finkbeiner, 2011] Schlafly, E. F. and Finkbeiner, D. P. (2011). Measuring reddening with sloan digital sky survey stellar spectra and recalibrating sfd. *The Astrophysical Journal*, 737(2):103.

[Sergison et al., 2013] Sergison, D. J., Mayne, N., Naylor, T., Jeffries, R., and Bell, C. P. (2013). No evidence for intense, cold accretion on to ysos from measurements of li in t-tauri stars. *Monthly Notices of the Royal Astronomical Society*, page stt973.

[Sneath, 1957] Sneath, P. H. (1957). The application of computers to taxonomy. *Microbiology*, 17(1):201–226.

[Soderblom et al., 2014] Soderblom, D. R., Hillenbrand, L. A., Jeffries, R. D., Mamajek, E. E., and Naylor, T. (2014). Ages of young stars. *Protostars and Planets VI*, 1:219–241.

[Teixeira et al., 2005] Teixeira, P. S., Lada, C. J., Young, E. T., Marengo, M., Muench, A., Muzerolle, J., Siegler, N., Rieke, G., Hartmann, L., Megeath, S. T., et al. (2005). Identifying primordial substructure in ngc 2264. *The Astrophysical Journal Letters*, 636(1):L45.

[Townsley et al., 2014] Townsley, L. K., Broos, P. S., Garmire, G. P., Bouwman, J., Povich, M. S., Feigelson, E. D., Getman, K. V., and Kuhn, M. A. (2014). The massive star-forming regions omnibus x-ray catalog. *The Astrophysical Journal Supplement Series*, 213(1):1.

[Tremblay et al., 2012] Tremblay, P.-E., Schilbach, E., Röser, S., Jordan, S., Ludwig, H.-G., and Goldman, B. (2012). Spectroscopic and photometric studies of white dwarfs in the hyades. *Astronomy & Astrophysics*, 547:A99.

[Van Leeuwen, 2007] Van Leeuwen, F. (2007). *Hipparcos, the new reduction of the raw data*, volume 350. Springer Science & Business Media.