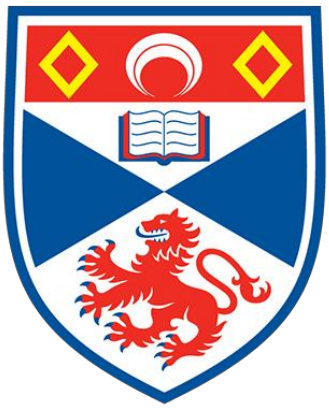
Mapping a star-forming region in 3D with Gaia astrometric data



University of
St Andrews

Callum Murison

cm469@st-andrews.ac.uk

Summer 2023

Supervised by Dr Paula Stella Teixeira

Abstract

This project uses Gaia DR3 data to map a star-forming region near the constellation Monoceros, in an attempt to identify new star clusters and associations. A different, more automated method was developed that looked at a previously-explored region in more detail. The work culminated in the discovery of two likely clusters - with approximate ages of 600 Myr and 15 Myr, and containing 51 and 27 stars respectively. Analysis of the internal motions of cluster members found neither cluster to be expanding.

Background

The Gaia telescope is a space observatory with a simple goal: to collect high-precision astrometric data for over a billion stars both inside and outside our galaxy [1]. The quality of this astrometric data (predominantly positions, distances and motions of stars) has allowed the construction of the most detailed 3D map of our Universe ever made.

Gaia data is made available in several **data releases** - the most recent being Data Release 3 (DR3) [2], released on 13th June 2022. 'Incomplete' data sets are also released in advance of these, known as **early data releases** (such as EDR3).

This project was carried out in the coding language Python [3], in conjunction with other tools which are explored in more detail throughout this essay.

Overview of terms

Several astronomical terms are crucial to understanding this project.

1. (RA, Dec), also written as (α, δ)
    - The astronomical coordinate system: Right Ascension and Declination, both in units of degrees.
2. Arcsecond
    - A small angle, 1/3600th of a degree.
3. Proper motion (μ^*_α, μ_δ)
    - The angular motion of stars on the sky, along both the Right Ascension (α) and Declination (δ) directions. Commonly in units of milli-arcseconds per year (mas/yr).
4. Transverse velocity (t^*_α, t_δ)
    - The physical motion of stars on the sky, along both the Right Ascension (α) and Declination (δ) directions. Commonly in units of kilometres per second (km/s).
5. (l, b)
    - The galactic coordinate system: Galactic Longitude l, Galactic Latitude b (both in units of degrees)
6. Parsec (pc)
    - Astronomical unit of distance. 1 pc = 3.086×10^{16} m (around 31 million billion metres).

Introduction

Stars form in 'stellar nebulae' (star nurseries), where massive clouds of cold gas collapse under their own gravity. In the process of collapsing, the gas heats up, nuclear fusion can begin, and stars are formed. The clouds are often massive enough that the stars are pulled together by the gravitational force between them; this collection of stars is then known as a **star cluster**. Over time, the group may become more spread out, to the extent where the stars are no longer categorised as 'gravitationally bound', but are still moving together; this is a **stellar association** [4].

Mapping these clusters and associations is worthwhile for several reasons. For one, having a well-constructed map of the Universe is very important for observational astrophysics. The impact of stellar clusters on the surrounding space can be significant, and only by knowing these effects are we then able to fully explain and understand other observations. Another is that clusters provide an ideal environment to test astrophysical theories: the cluster members all formed at roughly the same time, and so this removes a complicating factor when comparing the stars.

Specific to this project, Gaia was used to map a region of the sky near the Monoceros constellation. Associations have been identified previously, such as the Monoceros OB4 association found by Teixeira et al [5] in 2021. However, this previous work by Teixeira et al used data from Data Release 2 (DR2) [6], whereas my work uses a **newer data release in DR3**.

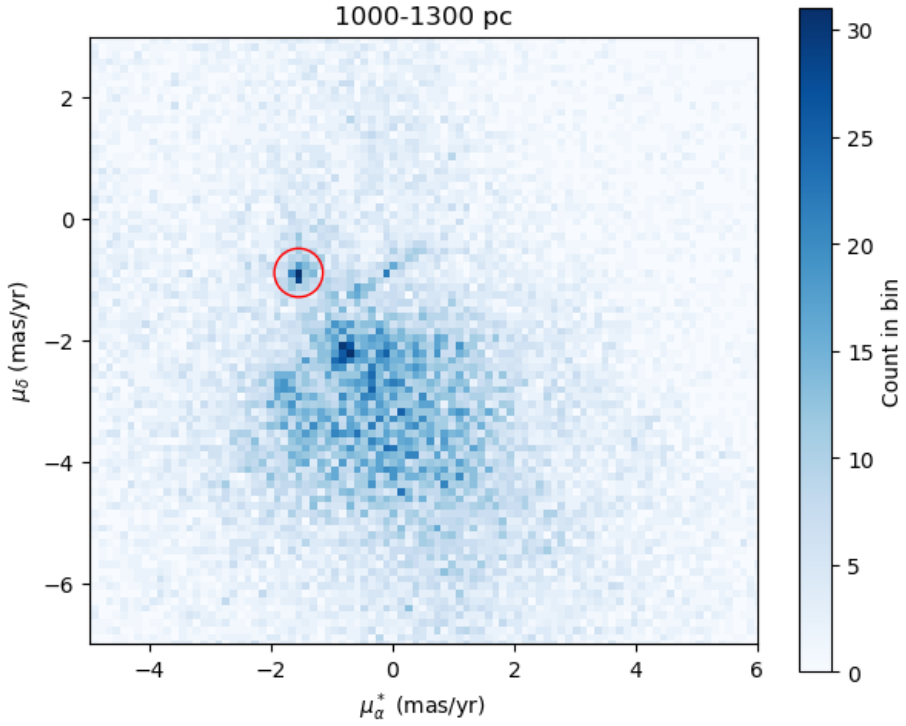


Figure 1: A histogram of proper motion, reproduced from Figure 2 of Teixeira et al (2021), showing stars within a distance slice of 1000-1300 pc. The OB4 association they identified is outlined approximately in the red circle.

The darker blue (high-density) regions in the above figure show large numbers of stars moving together - finding and exploring these regions formed the key interest for my project.

Outline of method

To identify stellar clusters, the goal was to find groups of stars that were **co-moving, co-distant and coeval** (moving together, at the same position in space, and of the same age). My method involved the following ordered steps, which are explored in more detail throughout this essay:

1. Selecting Gaia data for the desired region of the sky.
2. Finding stars that were grouped together in proper motion space (moving together).
3. Searching for trends in the positions of these stars, to see if they were close together in space.
4. Filtering the data to remove outliers.
5. Identifying the likely age of the stellar populations.

Lots of stars may just happen to be moving together - by coincidence, without actually being part of the same structure in space - and so the steps to refine by position and age are crucial.

Step 1: Data selection

Criteria

The 'region of interest' on the sky was located near the Monoceros constellation, bounded by the galactic longitude range $196.5° < l < 206.5°$, and galactic latitude range $-4.5° < b < 1.5°$. The same quality selection criteria were used as in Teixeira et al (2021) [5], filtering based on errors in parallax[1] and the proper motion in RA and Dec:

- Parallax/parallax error: $\bar{\omega}/\sigma_{\bar{\omega}} > 10.0$
- Proper motion (RA): $\sigma_{\mu^*_\alpha} < 0.2$ mas/yr
- Proper motion (Dec): $\sigma_{\mu_\delta} < 0.2$ mas/yr

(σ refers to the **standard deviation**, often taken as a good measure of error.)

Querying

Gaia data releases are accessible in several ways, all underpinned by ADQL (Astronomical Data Query Language) [8]. The interface between ADQL and Python was implemented using the *astroquery* package [9], ultimately creating a table containing the details of every star within the region of interest.

From this table, distance slices were made to analyse different subsections of the data, including the same slice explored by Teixeira et al (2021) [5], from 1000-1300 pc. Making these distance slices is a crucial step, because this selection of a smaller number of stars makes any patterns far more visible and simplifies the analysis.

All but one of the relevant columns were selected from the *gaiadr3.gaia_source* table, with the exception being the geometric distances calculated by Bailer-Jones et al (2021) [10] for the Early Data Release 3 (EDR3). These were stored in a different database table, called *external.gaiaedr3_distance*. An ADQL JOIN operation had to be performed between the two tables to select this distance column (see Appendix for further details).

Calculation of transverse velocities

Transverse velocity can be calculated from proper motion, eliminating any distance dependence and thus making it a useful tool for cleaning data sets.

Converting from one to the other is performed via the small angle approximation, combined with unit conversions. For a proper motion μ in mas/yr, and distance d in pc, the transverse velocity is found by $t \approx 4.74 \mu d$.

Dust

Dust has a significant impact on astronomical observations, through both **extinction** (blocking light), and **reddening** (a scattering effect which causes stars and galaxies to appear redder than they actually are).

This effect is hard to quantify, because dust is difficult to observe directly, and yet is critical for any analysis involving colour (something I require later when fitting isochrones). However, recent work by Green et al (2019) [11] managed to map a large region of the sky for dust out to several kiloparsecs, creating the **Bayestar dust map**.

The data underlying this 2019 paper was made available in the Python package *dustmaps* [12]. This contains a sort of 'raw reddening' - telling us how much reddening occurs in any direction, out to any distance. This 'raw redddening' was used to compute the visual extinction A_v for each star in the data set, as follows:

1. Finding the 'raw reddening' calculated in the Bayestar dust map for the coordinates and distance of every star.
2. Applying a multiplicative factor to convert from this 'Bayestar/raw reddening' to visual extinction A_v.

Finding this multiplicative factor to convert into A_v depends on something called the 'relative visibility' R_v, typically approximated in the Milky Way to $R_v \approx 3.1$. Table 6 from Schlafly and Finkbeiner (2011) [13] provides the factor to convert from Bayestar reddening into the V-band extinction A_v (listed as "Landolt V"): **2.742**.

Thus, multiplying the Bayestar reddening by 2.742 produces A_v and allows correction for the aforementioned 'reddening' phenomenon, necessary for further analysis (see Appendix for further details).

[1] Parallax is an angular measurement which defines the position of an astronomical object in space. This can be used directly to calculate distance, but only when there is no error in the measurement. In this case, that is not true, and so the problem becomes much more complicated, as discussed in Bailer-Jones et al (2015) [7].

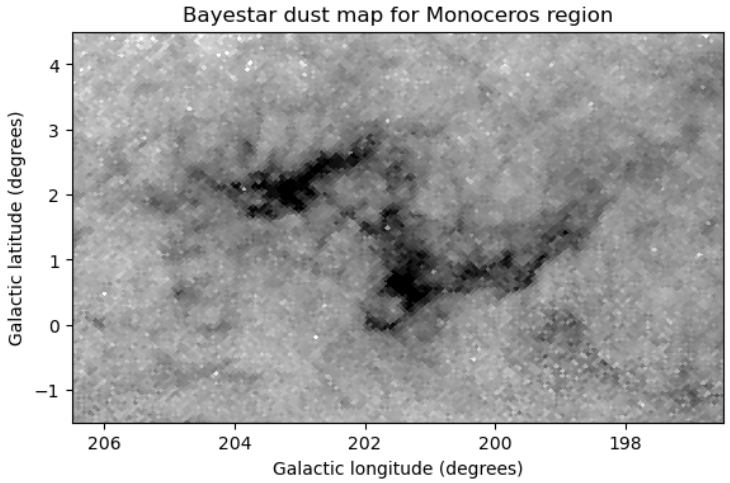


Figure 2: The Bayestar dustmap (2019) for the region of interest. Darker regions represent those with greater dust extinction.

Step 2: Gaussian Mixture Models

Whilst the Monoceros OB4 association in Figure 1 is obvious to the naked eye, this is not always the case. The main part of this project involved developing Gaussian Mixture Models (GMMs) to identify high-density regions in these plots which couldn't be identified by eye.

The basic premise behind a GMM is to fit N Gaussians to a dataset, in order to try and identify N clusters/groups. I implemented my model using the Scikit-learn package [14] - specifically the *GaussianMixture* class from *sklearn.mixture*. A GMM can be fitted to any number of 'dimensions', and I considered implementing either a 2D or 5D model. The 5D model appeared less reliable in testing (see later discussion), and so a 2D model was implemented.

This involved fitting a GMM to proper motion space, $[\mu^*_\alpha, \mu_\delta]$, to find the stars which are moving together. Note that this 2D GMM therefore gives no information on whether the stars are clustered in position space, [α, δ, distance], hence the need for Step 3 of the method.

The algorithm allowed the specification of the number of groups to try to fit to the data, with my choices outlined in the results. The most clustered of those fitted were found by considering the 'weight' - a measure of how tightly bound the data in a given group was. The top 3 most tightly-bound groups in a given GMM were selected for further analysis.

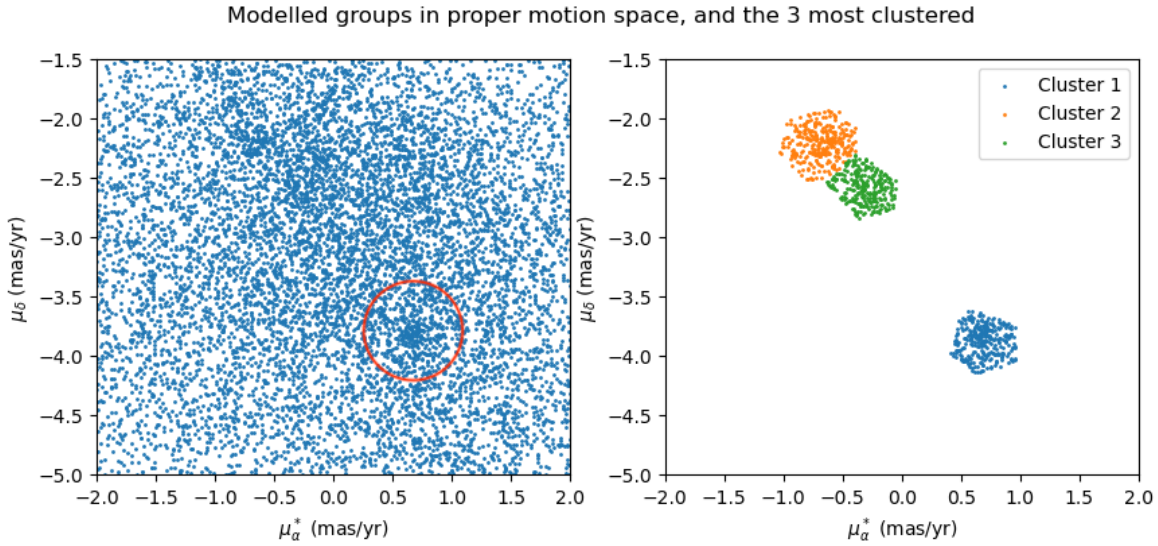


Figure 3: An example of the 2D Gaussian Mixture Model being applied to a data subset, between 1250-1500 pc. The raw data is displayed on the left, with the three most clustered regions highlighted in the right plot. An obvious high-density region in the left plot is circled, and this is correctly found by the algorithm (with the label 'Cluster 1', in the right plot).

Step 3: Position clustering

In contrast to the heavily-automated code of the previous section, searching for clustering in position had to be performed by eye. A key trend that could signify a cluster is where the members have a reasonable spread in distance, but are concentrated in galactic latitude and longitude, shown below:

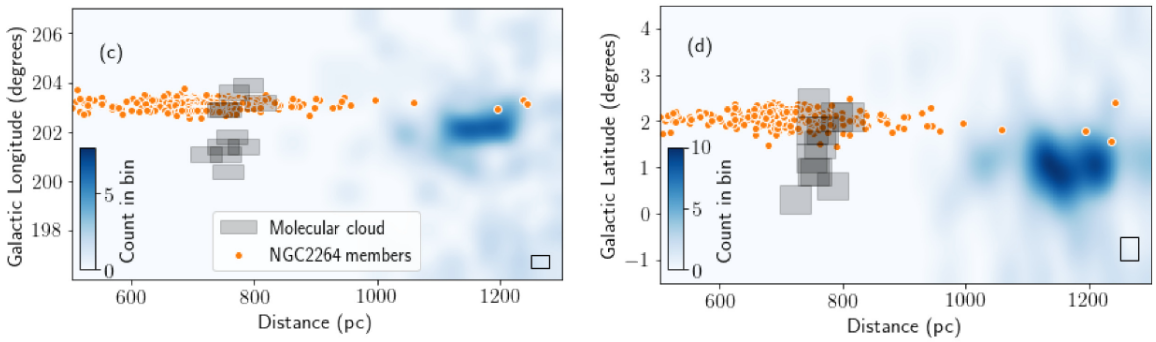


Figure 4: The NGC 2264 cluster, also lying in Moncoeros, shown by the orange circles - a pattern that I needed to look for in my results. Taken from Figure 3 in Teixeira et al (2021) [5].

Step 4: Data filtering

Gaussian models are extremely important tools for data refinement. In essence, they compute a mean (μ) and standard deviation (σ) for a dataset, allowing selection of data lying within desired confidence intervals (such as '3-sigma' tolerance: $\mu \pm 3\sigma$).

In 2D (known as a **2D bivariate Gaussian fit**), these 'confidence intervals' show up as ellipses. Visually, a 3-sigma tolerance is then interpreted as data lying inside the 3-sigma ellipse contour, with data points lying outside the ellipse considered to be outliers.

To perform the bivariate fit, I used an astrophysics-based machine learning package called AstroML [15], specifically the *fit_bivariate_normal* method from the *astroml.stats* package. I used an algorithm from the book "Statistics, Data Mining and Machine Learning in Astronomy" [16] to perform the fit, modifying a mathematical method to select the data points lying within the 3-sigma ellipse [17].

This process was performed in both proper motion and transverse velocity spaces simultaneously (i.e. only selecting data that lay within the 3-sigma confidence interval in both), to reduce the possibility of contamination by outliers.

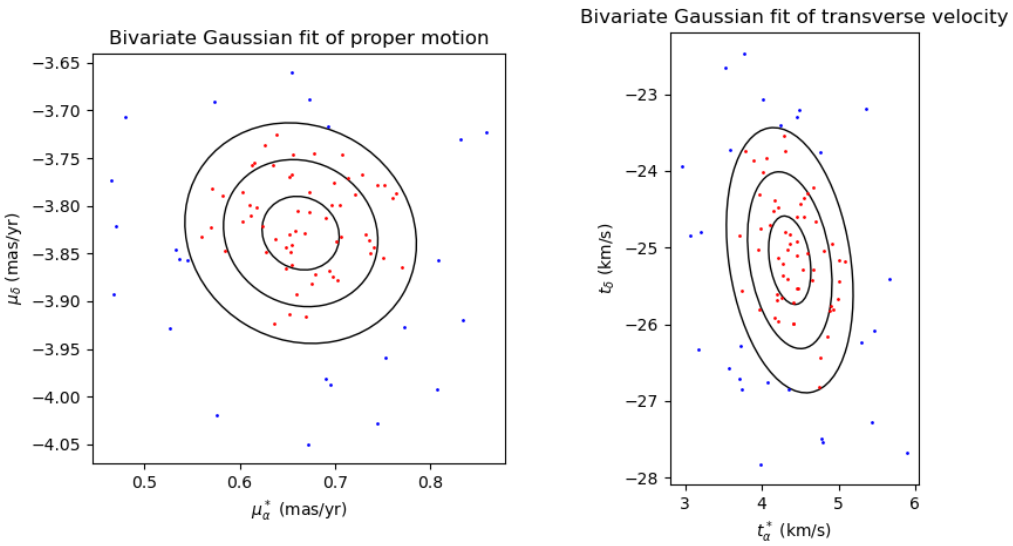


Figure 5: An example of performing a 2D Gaussian bivariate fit. The three black paths in each plot represent the 1σ, 2σ and 3σ confidence intervals. Points lying within the 3σ interval were selected and coloured red; the outlying points were discarded and are shown in blue.

Step 5: Determining stellar ages

The concepts of brightness and colour in astronomy are somewhat complex. Brightness is most often measured on the logarithmic magnitude scale, where a smaller magnitude corresponds to a brighter source. Magnitude is measured in various different 'bands' - for example, how bright the source looks through a blue filter, a green filter and a red filter. This 'observed' magnitude is known as the apparent magnitude.

Colour is then calculated as the **difference between apparent magnitudes** in different bands. For example, the relevant colour calculated from Gaia's BP (blue) and RP (red) filters is BP-RP (BP magnitude minus RP magnitude).

Combining colour and magnitude in a plot produces a **colour-magnitude (CM)** diagram. Here, we often use 'absolute' magnitude[2] in place of 'apparent' magnitude, which eliminates any dependence on distance. Plotting such a diagram allows estimation of the approximate age of a group of stars (or indeed, informs us as to whether they are the same age at all) by fitting **isochrones**. These isochrones are lines on a CM diagram that show what the shape of the diagram should look like if the stars were all the same age.

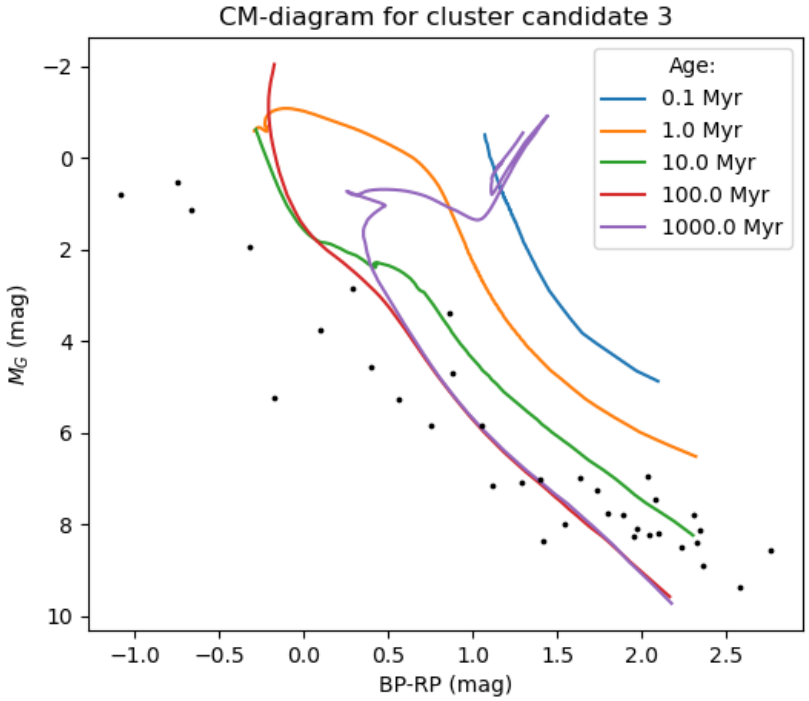


Figure 6: Colour-magnitude diagram for a group of stars found **not** to be a cluster. Absolute magnitude in the G-band is plotted on the y-axis, with BP-RP colour plotted on the x-axis. None of the isochrones (the coloured lines) fit the black data points. This tells us that the stars in this group are not of the same age, and so are unlikely to be part of a single cluster.

To implement isochrone fitting, I plotted MIST isochrones using the Python package *isochrones* [18]. Only stars for a cluster candidate whose data points provided a good fit to a given isochrone could be assumed to be of a common age, thus meeting the final criterion for finding a cluster.

[2]Absolute magnitude is defined as the apparent magnitude that an object would have if it were located at a distance of 10 parsecs away from us. This hence eliminates the distance dependence associated with apparent magnitude (where objects which are further away appear fainter, regardless of their intrinsic brightness).

Results

Several distance slices were analysed, and the notable results are described below, both of them located between **1250-1500 pc**.

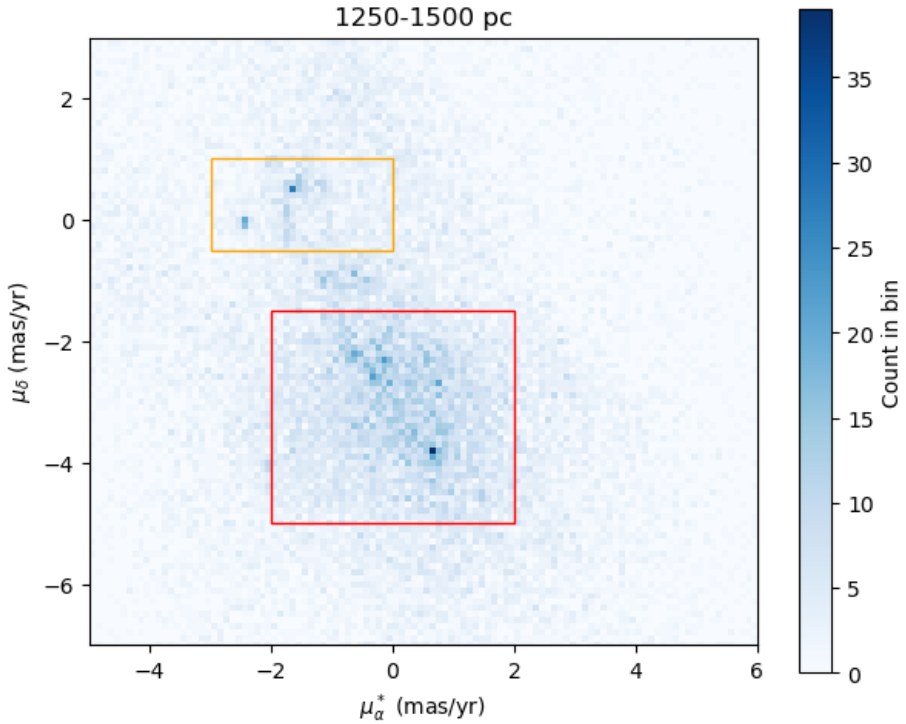


Figure 7: A histogram of proper motion for 1250-1500 pc. Fitting a Gaussian Mixture Model to the red region, with the parameters described later in this section, led to the discovery of cluster candidate 1. The same is true for the orange region and cluster candidate 2.

Cluster candidate 1

The first cluster candidate identified contains 51 stars, lying at a mean distance of (1390 ± 40) pc. The central coordinates of the cluster are estimated to be (l, b) = (203.6°, 0.11°). The cluster was found via a 50-group GMM fitted to the region contained within $-2 < \mu_\alpha^* < 2$; $-5 < \mu_\delta < -1.5$ (units of mas/yr). This is the red box in Figure 7.

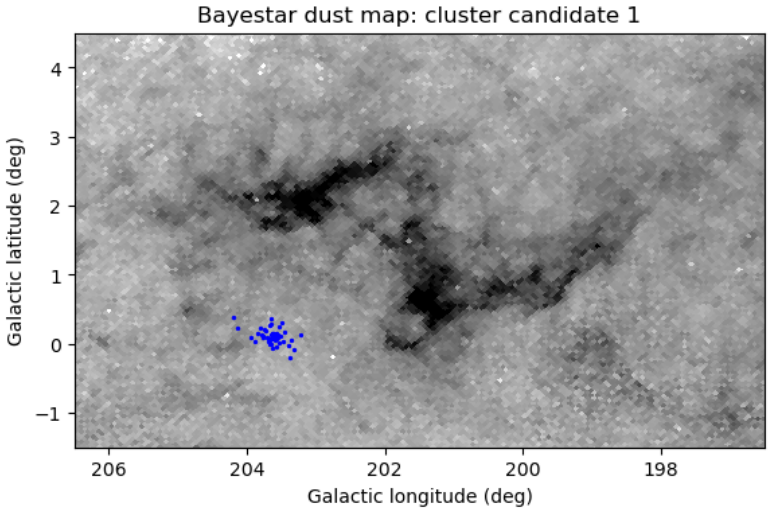


Figure 8: The location of cluster candidate 1, shown in blue, plotted over the Bayestar dust map of the Monoceros region.

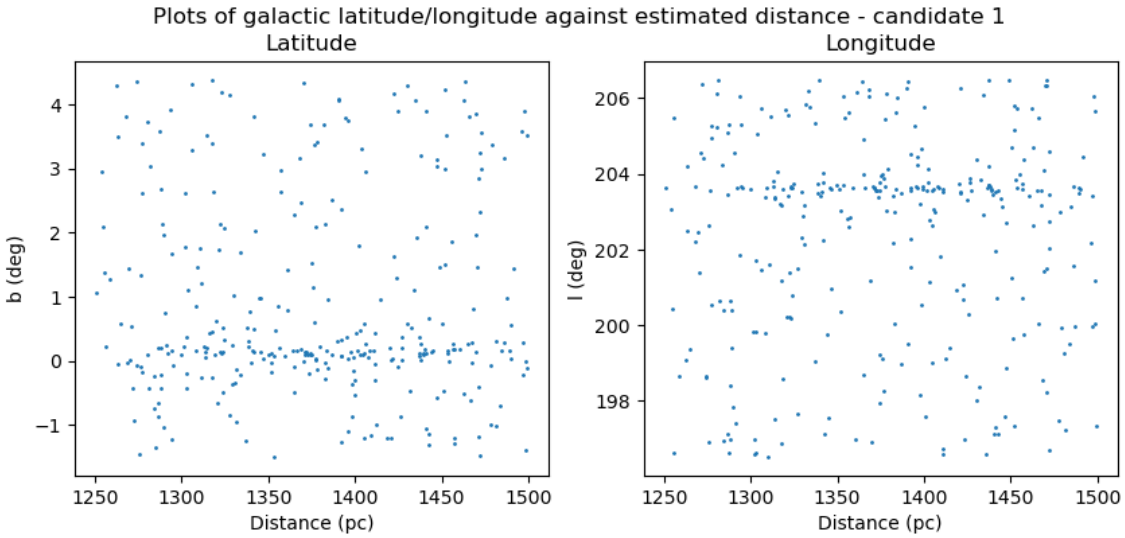


Figure 9: Plots of galactic latitude and longitude against distance for cluster candidate 1. The 'lines' of stars are of a similar pattern to the NGC 2264 cluster, shown in figure 4. Details of selecting the stars in these 'lines' can be found in the Appendix.

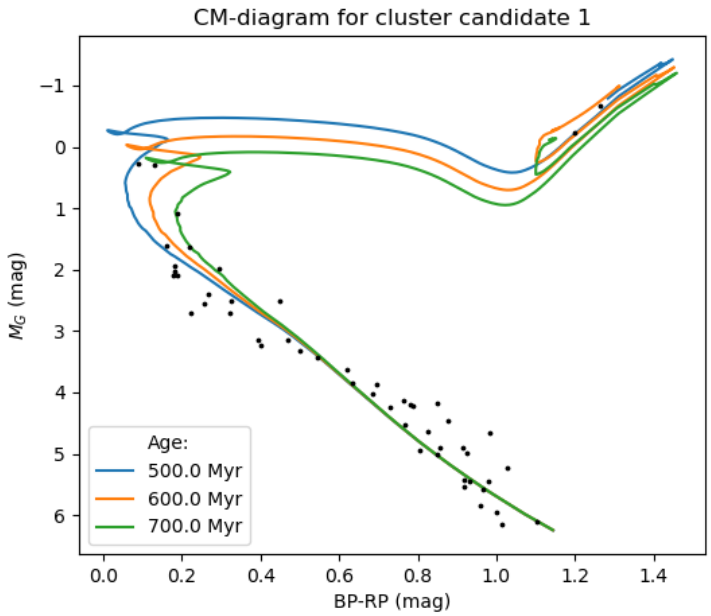


Figure 10: Colour-magnitude diagram for cluster candidate 1, with isochrones overlaid. The orange isochrone provides the best fit by eye to the black data points.

Cluster candidate 2

The second cluster candidate identified contains 27 stars, lying at a mean distance of (1420±40) pc. The central coordinates of the cluster are estimated to be (l, b) = (205.9°, −0.45°). The cluster was found via a 30-group GMM fitted to the region contained within $-3 < \mu_\alpha^* < 0$; $-0.5 < \mu_\delta < 1$ (units of mas/yr). This is the orange box in Figure 7.

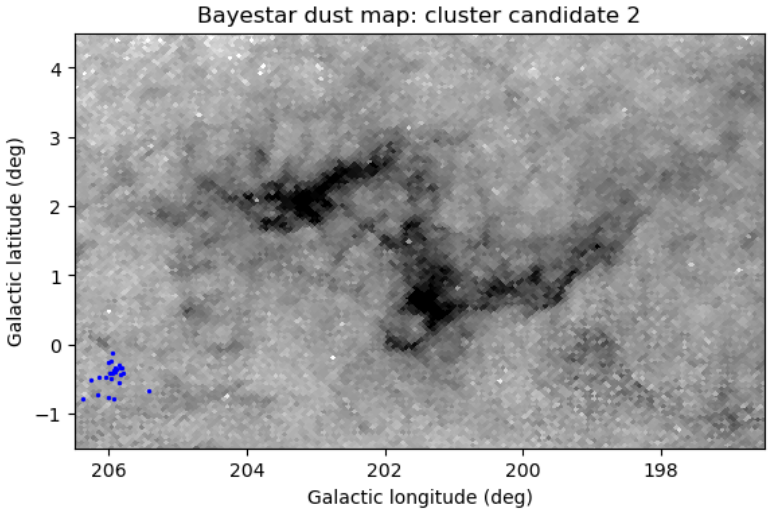


Figure 11: The location of cluster candidate 2, shown in blue, plotted over the Bayestar dust map of the Monoceros region.

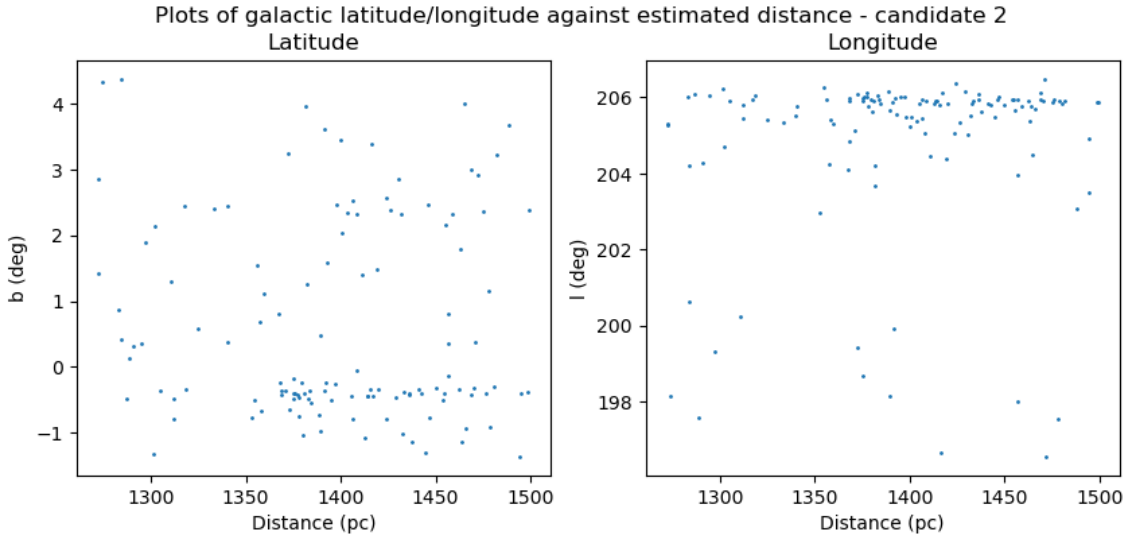


Figure 12: Plots of galactic latitude and longitude against distance for cluster candidate 2, again with visible 'lines' of stars. The selection criteria can again be found in the Appendix.

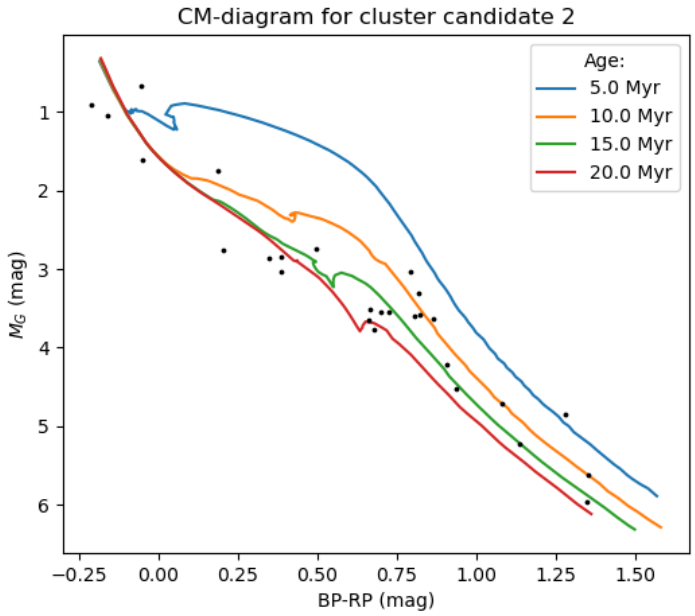


Figure 13: Colour-magnitude diagram for cluster candidate 2, with isochrones overlaid. The red, green and orange isochrones all provide a reasonable fit by eye to the data, suggesting the age of cluster members to lie between 10 and 20 Myr.

Note: the parameters to the Gaussian Mixture Model in both cases took a maximum number of iterations as max_iter = 500, and a tolerance of tol = 0.01. Both candidates were the most clustered groups found by their respective GMMs.

Discussion

Cluster 1 is an old cluster at 600 million years old, whereas Cluster 2 would be considered young. Other clusters in Monoceros have ages broadly agreeing with Cluster 2 - the age of Cluster 1 is therefore surprising, with more work required to explore why it is so different.

The motions of the stars in each cluster relative to the mean cluster velocities were found to be very small. This is indicative that neither cluster is expanding, which is another unusual feature of Cluster 1 – we would expect random motions of stars to increase over time, leading to expansion in older clusters.

Further potential cluster candidates were identified using the same method, but the MIST isochrone fitting showed the stars to not share a common age (such as for cluster candidate 3 in figure 6). Future steps would involve probing this region further, and perhaps considering different model isochrones (such as PARSEC [19]) to provide comparison with MIST.

Furthermore, I had attempted to develop a five-dimensional Gaussian Mixture Model (5D GMM), which searched for clusters in position and velocity spaces simultaneously. There was limited success using this method, and hence the final results coming from a combination of the 2D GMM and manual analysis of the data. Given more time, I believe that this 5D GMM method could be improved, providing a different means of finding clusters.

Finally, I would have liked to have explored the distances in the DR3 catalogue in more detail. Two calculated distances exist from different methods: the one I used from Bailer-Jones et al [10]; the other directly from the Gaia Consortium, GPAC [20]. My choice was based on the method used by Teixeira et al (2021) [5]. Differences were visible between the two sets of distances, and more work is required to test whether this would lead to meaningful differences in results.

Acknowledgements

I would like to thank my supervisor, Dr Paula Stella Teixeira, for embarking on this project with me, and providing invaluable help along the way. I would also like to thank the Laidlaw Foundation, and Lord Laidlaw for funding this scholarship, and the Laidlaw team at the University of St Andrews for their guidance and organisation throughout and beyond the six weeks.

I would also like to thank my fellow Laidlaw Scholars at the University of St Andrews, the other summer research students, and my friends for all of their help and support.

This project used data collected by the ESA Gaia mission (`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`).

References

[1] Gaia mission | Gaia in the UK. https://www.gaia.ac.uk/mission.

[2] Gaia Collaboration. Gaia Data Release 3 - Summary of the content and survey properties. *Astronomy & Astrophysics*, 674:A1, June 2023.

[3] Python.org, July 2023.

[4] Chelsea Gohd. Star Clusters: Inside the Universe's Stellar Collections. https://universe.nasa.gov/news/235/star-clusters-inside-the-universes-stellar-collections.

[5] P S Teixeira, J Alves, A Sicilia-Aguilar, A Hacar, and A Scholz. Monoceros OB4: a new association in Gaia DR2. *Monthly Notices of the Royal Astronomical Society: Letters*, 504(1):L17–L21, April 2021.

[6] Gaia Collaboration. *Gaia* Data Release 2: Summary of the contents and survey properties. *Astronomy & Astrophysics*, 616:A1, August 2018.

[7] Coryn A. L. Bailer-Jones. Estimating Distances from Parallaxes. *Publications of the Astronomical Society of the Pacific*, 127(956):994, October 2015. Publisher: IOP Publishing.

[8] Astronomical Data Query Language. https://www.ivoa.net/documents/ADQL/20180112/PR-ADQL-2.1-20180112.html.

[9] Astroquery — astroquery v0.4.7.dev439. https://astroquery.readthedocs.io/en/latest/.

[10] C. A. L. Bailer-Jones, J. Rybizki, M. Fouesneau, M. Demleitner, and R. Andrae. VizieR Online Data Catalog: Distances to 1.47 billion stars in Gaia EDR3 (Bailer-Jones+, 2021). *VizieR Online Data Catalog*, page I/352, February 2021. ADS Bibcode: 2021yCat.1352....0B.

[11] Gregory M. Green, Edward Schlafly, Catherine Zucker, Joshua S. Speagle, and Douglas Finkbeiner. A 3D Dust Map Based on Gaia, Pan-STARRS 1, and 2MASS. *The Astrophysical Journal*, 887(1):93, December 2019. Publisher: The American Astronomical Society.

[12] Gregory M. Green. dustmaps: A Python interface for maps of interstellar dust. *Journal of Open Source Software*, 3(26):695, June 2018.

[13] Edward F. Schlafly and Douglas P. Finkbeiner. Measuring Reddening with Sloan Digital Sky Survey Stellar Spectra and Recalibrating SFD. *The Astrophysical Journal*, 737(2):103, August 2011. Publisher: The American Astronomical Society.

[14] Pedregosa et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

[15] J.T. Vanderplas, A.J. Connolly, Ž. Ivezić, and A. Gray. Introduction to astroml: Machine learning for astrophysics. In *Conference on Intelligent Data Understanding (CIDU)*, pages 47 –54, oct. 2012.

[16] Ž. Ivezić, A.J. Connolly, J.T. Vanderplas, and A. Gray. *Statistics, Data Mining and Machine Learning in Astronomy*. Princeton University Press, 2014.

[17] tmdavison. Stack overflow answer, May 2016. https://stackoverflow.com/a/37032759.

[18] isochrones — isochrones 2.1 documentation. https://isochrones.readthedocs.io/en/latest/index.html.

[19] PARSEC online interface. http://stev.oapd.inaf.it/cgi-bin/cmd.

[20] DPAC Consortium - Gaia - Cosmos. https://www.cosmos.esa.int/web/gaia/dpac/consortium.

[21] Shu Wang and Xiaodian Chen. The Optical to Mid-infrared Extinction Law Based on the APOGEE, Gaia DR2, Pan-STARRS1, SDSS, APASS, 2MASS, and WISE Surveys. *The Astrophysical Journal*, 877(2):116, June 2019. Publisher: The American Astronomical Society.

Appendix

The method detailed in this essay should ensure reasonable reproducibility of results; however, certain additional information is provided below to make certain steps clearer.

Querying

The columns selected from the *gaiadr3.gaia_source* table were:

- source_id (a unique identifier for each star in the catalogue)
- l, b, ra, dec (galactic longitude and latitude; RA and Dec coordinates)
- pmra, pmdec (proper motion in RA and Dec)
- phot_g_mean_mag (G-band magnitude - the brightness of the star through Gaia's green filter)
- bp_rp (observed BP-RP colour)

The single column selected from the *external.gaiaedr3_distance* table was:

- r_med_geo (the median distance computed for the sources by Bailer-Jones et al [10])

Colour correction

As outlined at the end of Step 1 of the method, a multiplicative factor of 2.742 was applied to convert from 'raw reddening' into visual extinction A_v. A_v was then used to calculate the **colour excess**, a quantity allowing the colour to be corrected.

This colour excess was computed by considering the ratios in Table 3 of Wang et al (2019) [21] for the Gaia G-band (since my colour-magnitude plots use the G-band magnitude). The following equations from this table were solved simultaneously to get $E(BP-RP)$, the colour excess, in terms of A_v:

$$A_\lambda / A_v = 0.789$$

$$A_\lambda / E(BP-RP) = 1.890$$

$$\implies E(BP-RP) \approx 0.4174 A_v$$

With both the colour excess and the observed colour known, the intrinsic colour was then found through rearranging the following equation:

$$E(BP-RP) = (BP-RP)_o - (BP-RP)_i$$

where $(BP-RP)_o$ is the observed BP-RP colour and $(BP-RP)_i$ is the intrinsic/true BP-RP colour. This intrinsic/true colour is what was then plotted on the colour-magnitude diagrams.

Position clustering

As per Figures 8 and 12, the 'lines' of stars visible were selected by eye. For cluster candidate 1 (Figure 8), the cluster stars were taken to be within the ranges $203.2° < \mathrm{l} < 204.2°$ and $-0.3° < \mathrm{b} < 0.4°$.

For cluster candidate 2 (Figure 12), the cluster stars were taken to be within the ranges $205° < \mathrm{l} < 207°$ and $-1° < \mathrm{b} < 0°$.

Isochrone fitting

The MIST isochrones generated during the fitting process were done so with the following parameters:

- Masses (a range of roughly 0.4 to 5 solar masses)
- Metallicity (taken to be the solar metallicity: Fe/H = 0.05)
- Distance (10pc - just based on the definition of absolute magnitude)
- Extinction A_v (taken to be 0, since the raw data had already been corrected for dust in Step 1)

The ages of isochrones to plot were chosen by trial and error, to see which isochrones best matched up with the data points for a given cluster candidate.