CLUSTER PROPERTY ESTIMATION WITH GAIA DR3 AND ISOCHRONE FITTING

Abhishek Kattuparambil[1]

[1] *Department of Astronomy, University of California, Berkeley, CA, USA 94720*

Keywords: Color-Magnitude Diagrams — Synthetic Photometry — Theoretical Isochrones — GAIA DR3 — Hertzsprung-Russell Diagrams — Evolution

1. INTRODUCTION

Gaia is a space observatory launched by the European Space Agency in 2013, which has recorded astrometry for roughly two billion sources with remarkable precision. The Gaia mission set out to document 1% of the Milky Way, focusing on bright stars falling within an extended visual photometric band. Photometry is done with a blue photometer (BP) spectra for wavelengths of 330–680 nm and a red photometer (RP) spectra for wavelengths of 640–1050 nm. These measurements allow for the estimation of metallicity, age, effective temperature, apparent/absolute magnitudes, and therefore, the creation of Hertzsprung-Russell diagrams (HRD) for select clusters in the Milky Way. After the Gaia Data Release 3 (GAIA DR3) earlier this year, the measurements have been made public. In this paper, we will create HRDs for the Hyades, M67, and NGC 6397 clusters, and estimate their age and metallicity by comparing them to theoretical isochrones generated from both MIST and PARSEC models. Furthermore, we will analyze the evolutionary path of each cluster and compare the real photometric data to the synthetic data created by the model.

2. GAIA ARCHIVE

During its conception, GAIA was an acronym, standing for Global Astrometric Interferometer for Astrophysics. Gaia's interferometry technique has changed since then, and it has lost its lengthy moniker, yet it continues to record some of the best data we have on members of our own Milky Way Galaxy. The European Space Agency provides well-documented public access to all of Gaia's data in the Gaia Archive[1], which can be queried directly from the website, or through API requests abstracted by the `astroquery.gaia` package. All queries must be written in Astronomical Data Query Language (ADQL), which is syntactically very similar to SQL, but contains added functionality to simplify complex queries for stellar data. Each source has 152 astrometric and photometric variables, including `parallax`, `ra` (right ascension), `dec` (declination), `pmra` (proper motion along the axis of right ascension), `teff_val` (effective temperature), `bp_rp` (BP-RP color magnitude), among others. We will be collecting the initial datasets for each cluster with a simple positional filter on `ra` and `dec`, querying all stars in a circular region of the sky around the cluster. This is clearly not enough, as we must filter on a third spatial dimension to query the cluster in three-dimensional space, and further filter on photometric and proper motion variables to eliminate extraneous sources which share the same space as the cluster.

[1] https://gea.esac.esa.int/archive/

3. DATA FILTRATION

Due to the precision of Gaia, we will be doing the majority of filtering using astrometric variables, and will be making further quality cuts with photometric variables. We will employ a variety of quality cuts over the data set to ensure we are not selecting any faux cluster members (Babusiaux 2018). Firstly, leveraging the precision of our parallax measurements, we will only keep sources with `parallax_over_error > 10`. Furthermore, we will also filter on `phot_g_mean_flux_over_error > 50` and `phot_rp_mean_flux_over_error > 50` in hopes of removing variable stars discussed in (Eyer 2018). These constants are unique to Hyades, and will be relaxed for M67 and Hyades, since their photometric measurements are expected to have larger natural variation. Other suggested cuts, such as `visibility_periods_used > 8` and additional constraints on `phot_bp_rp_excess_factor` were omitted as they had no result on the final output; further astrometric filtering was able to remove these sources.

3.1. *Celestial Coordinates*

As mentioned before, we will employ our first (and most important) cut based on the celestial coordinates in our ADQL query. We center our search at the celestial coordinates of the cluster and search over a radius

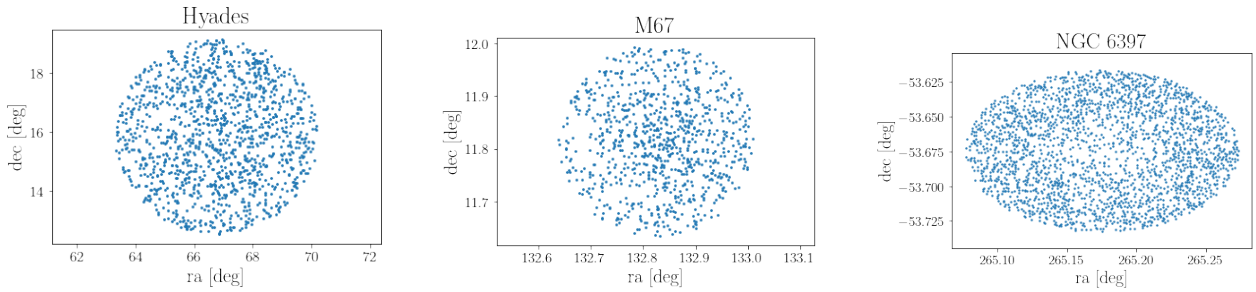

Figure 1. We can see that the Hyades portion of the sky is very dense, our query returned roughly 1400 sources and Hyades only contains almost 400. M67 has some definition towards the larger right ascensions and shows a visually dense cluster. NGC 6397 is a very apparent, tightly-bound cluster on the sky, and almost all of the sources returned by this query truly belong to the cluster.

slightly larger than the cluster's radius. To examine a circular portion of the sky, we construct a query of the following form:

```
# cluster_width is measured in arcminutes
# ra, dec are measured in degrees
"""
SELECT * FROM gaiadr3.gaia_source
WHERE 1 = CONTAINS(
   POINT(ra, dec),
   CIRCLE(<cluster_ra>, <cluster_dec>,
↪   <cluster_width>/60.))
"""
```

This ADQL query alone allows us to visualize these cluters. Figure 3.1 shows each cluster plotted in celestial coordinates. We have already cut our region out of celestial space, so any members falling outside this circle have been classified as outliers.

3.2. *Parallax*

Right ascension and declination are measured perpendicular to the line of sight, and this query is returning results from a two-dimensional patch of sky. The stellar clusters exist in three-dimensional space, so we must also account for the distance from the observer, which we can accomplish by filtering on parallax. Parallax is the apparent angular shift of an object compared to background sources due to a change in the observer's point of view. When measured in arcseconds ["], parallax has a direct inverse relationship with distance measured in parsecs [pc].

$$d = 1/p \tag{1}$$

The cosmic distance ladder suggests that a star's parallax is too small to accurately measure at a distance of roughly 10,000 light years. The furthest cluster we are querying is NGC 6397, which is around 7,800 light

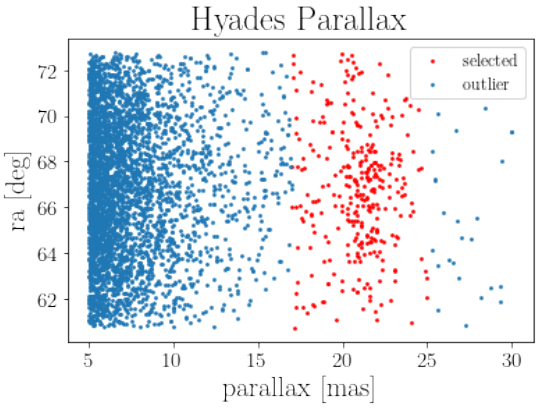

Figure 2. As suggested by the literature, the Hyades cluster is centered around a parallax of 21 mas. We permitted a difference of 4 mas on either side for the best results on the HRD. Interestingly enough, we found that sources with parallaxes larger than 6 mas were not outliers on the Hyades HRD, and actually improved the definition of the main sequence band.

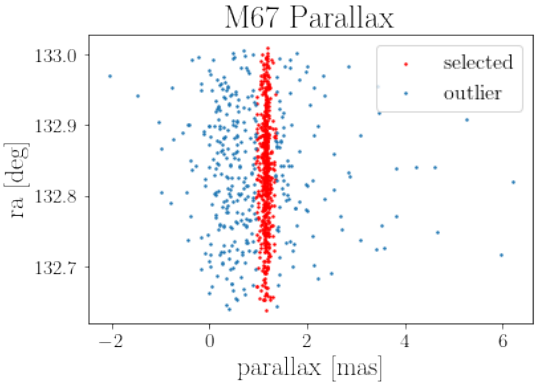

Figure 3. There is a strong individual band in the middle of this plot that is the M67 cluster. We also attempted creating HRDs after selecting over a larger range of parallaxes, but it seemed to include a large amount of late main sequence stars. This will become more apparent in the future section.

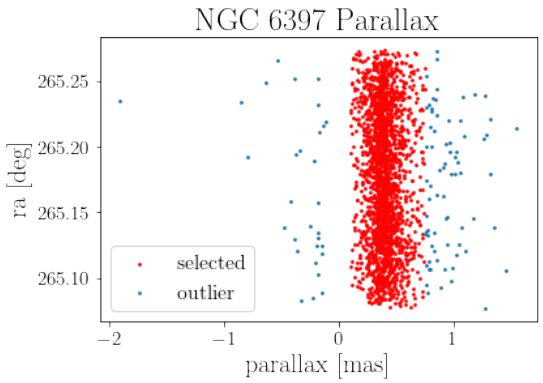

Figure 4. As with M67, there is a definite group in the parallaxes of potential NGC 6397 stars. We can make a tight selection here as this is a larger cluster, so we can afford to filter out a few genuine members.

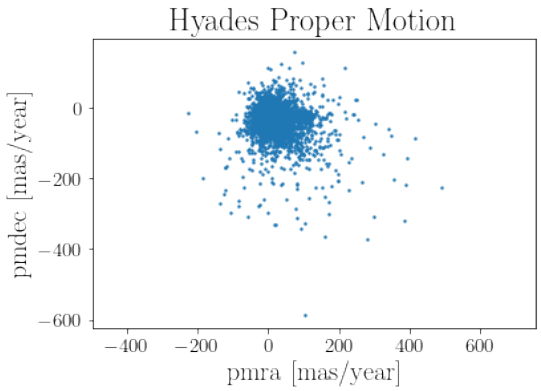

Figure 5. For reference, this is the unfiltered Hyades query plotted in proper motion space. There is a clear cluster definition, but it includes many outliers and stray sources which will widen the Main Sequence band. See Figure 6 for the selected cluster.

years away, so parallax measurements should be reliable throughout our study.

In each of these plots, we can identify a group of stars with similar parallaxes and make straightforward cuts to isolate them. Since the parallax measurements are extremely precise, this filter is one of the most accurate available in the Gaia data. We are able to see a third positional variable and often times, the cluster is revealed within its three-dimensional window. We must now turn to common cluster velocities to further validate cluster members. See Figures 2, 3, and 4.

3.3. *Proper Motion*

Proper motion (μ) is the actual angular velocity of a source across the sky. It is often split into two components along the celestial axes, `pmra` ($\mu_{\alpha*}$) and `pmdec` (μ_σ), such that

$$\mu^2 = \mu_{\alpha*}^2 + \mu_\sigma^2 \tag{2}$$

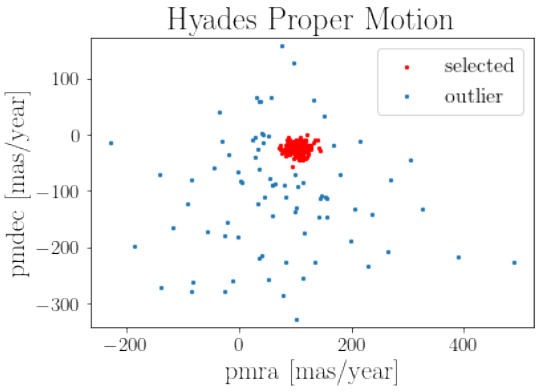

Figure 6. Here, we see there is a clear cluster defined around `pmra` = 100 and `pmdec` = -30. We can make simple rectangular cuts to select the cluster members. We experimented with a larger region including the sources close to our selected cluster, and they seemed to closely follow the evolutionary path of Hyades, but there were too many outliers.

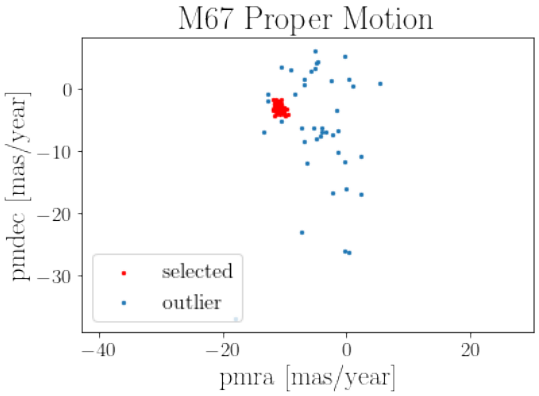

Figure 7. Although this cluster is less clearly visible, this proper motion filter underwent the most testing. Sources around the cluster are outliers to the M67 late main sequence evolutionary path. The sources in the cloud to the right of our selected cluster cause the 'flair' at the bottom of our unfiltered HR diagram.

Since each of the clusters is a gravitationally bound system, their stars will exhibit similar proper motions along each axis. This provides us for our next filtering candidate: `pmra` and `pmdec`. We will plot each cluster in proper motion space (i.e `pmdec` v `pmra`) and try to further isolate cluster members. See Figures 6, 7, and 9.

We experimented with different methods of clustering, including trials with the K-Nearest Neighbors (KNN) and K-Means Clustering algorithms, both implemented with `scikit-learn`[2]. KNN is a supervised learning algorithm, and works by classifying a data point as the

[2] https://scikit-learn.org/stable/

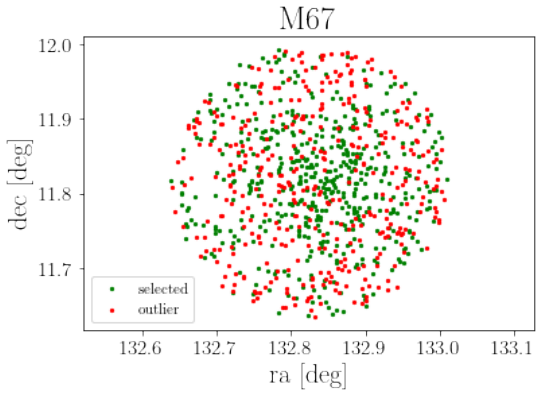

Figure 8. To reiterate, the physical position of the star does not always show whether it belongs to a cluster or not; each cluster is arbitrarily sparse, and we must allow for spatial outliers in initial queries.

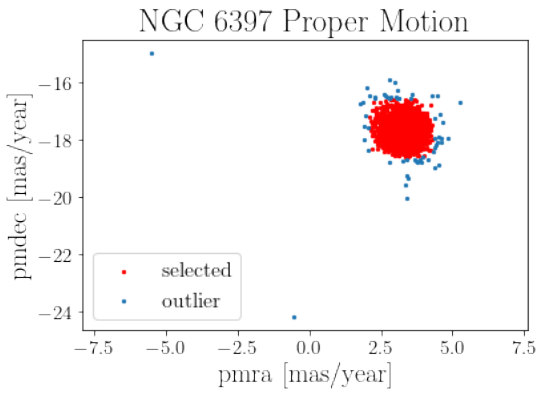

Figure 9. As mentioned before, almost every source from the positional query was truly a member of NGC 6397. It is further apparent here, as we have only two real outliers to this cluster in proper motion space. We are very stringent with the cuts since this is a much larger dataset, and keep only the centroid of this stretched cluster.

mode of its k nearest neighbors. Initializing this environment by handpicking points we knew were in the cluster was more tedious than simple rectangular cuts with `astropy.Table` slicing[3]. On the other hand, K-Means clustering is an unsupervised learning algorithm; it will classify the points into k clusters by iterating on its initial k guesses and updating optimal cluster assignments. Unfortunately, the use case for clustering in proper motion space is ill-defined, as we are simply attempting to find one cluster out of the entire dataset. K-Means is not designed for this environment and performs best with disjoint clusters, not intertwined data. Regardless, the clustering algorithm would seem to work best as some initial guess iterating on a KNN-esque loss function which evaluates the probability of a star belonging to a cluster based on its distance to an existing centroid. Once the values were plotted, all the algorithms seemed like an overkill as the rectangular cuts could be made just by looking at the plots. Therefore, we did our filtering by simply slicing indices to enclose an area limited to the cluster (Wang 2022).

[3] This is the 'supervised' portion of the algorithm. Since it requires the labels of a source's neighbors, we must have an existing set of labels. However, we have no points which we can confidently classify, so we need an unsupervised algorithm.

4. COLOR-MAGNITUDE DIAGRAMS

Now that we have filtered our datasets to remove outliers and identify actual cluster members, we can create our HRDs. Given Gaia's photometric data, we will create a Color-Magnitude Diagram (CMD) for each cluster by plotting the absolute magnitude in the Gaia G band (M_G) against the color index ($G_{BP} - G_{RP}$). Setting d to be the distance in parsecs to the star, or p to the star's parallax, we can solve for M_G accordingly.

$$M_G = \texttt{phot\_g\_mean\_mag} + 5 - 5 * log(d) \quad (3)$$

$$M_G = \texttt{phot\_g\_mean\_mag} + 5 * log(p) - 10 \quad (4)$$

This data is simpler to gather than the axes used on a traditional HRD, which are typically absolute magnitude (M) or luminosity (L) against temperature (T_{eff}). For each cluster, we will plot the CMD of the filtered dataset as well as a CMD of the filtered dataset overlaying the original. In each plot, we expect a clear indication of the Main Sequence, as well as select outliers, likely giants or white dwarfs belonging to each cluster.

5. STELLAR ISOCHRONES

A stellar isochrone is a path followed by a star on the HRD, usually spending most of its time on the Main Sequence. Stellar isochrones for our CMD can be created with three photometric variables: (G, G_{BP}, G_{RP}), the same three visual passbands used in Gaia DR3 (Babusiaux 2021). Therefore, we will create synthetic photometric data as a function of age and metallicity, and plot the generated isochrones against the selected sources from the previous section. This technique will allow us to estimate both the age and metallicity of the clusters, and further comment on their composition. Before making initial guesses, we must note that older clusters typically have lower metallicities. Since the formation of the older stars, generations of newer stars have fused and expelled metals, further saturating the metallicity of the universe. The older clusters were formed in a metal-poor universe and will have a noticeably lower metallicity.

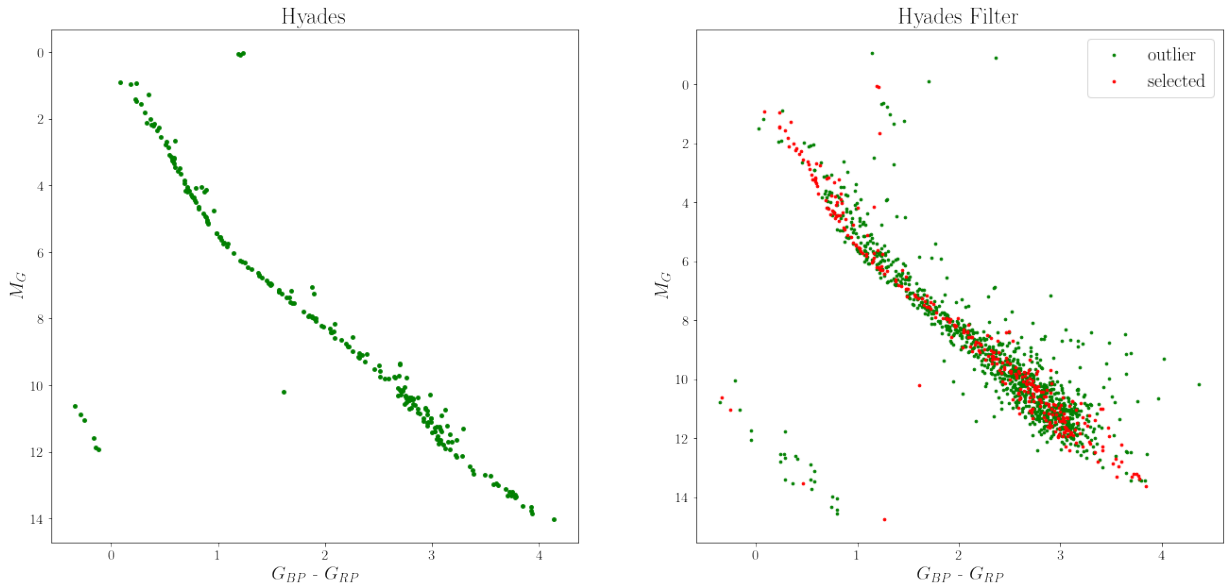

Figure 10. CMD for selected Hyades members. We did the most experimentation with the Hyades cluster, as many sources in the same celestial space follow the same evolutionary path down the CMD. As we can see from the plot on the right, we selectted roughly 20% of the total sources queried, and filtered many of the outliers that give the main sequence a spread look. The filtered CMD on the left is much more common in literature, with the characteristic yellow giants (to the upper right) and white dwarfs (to the lower left. Based on the magnitude and color of these stars, the majority of the Hyades consists of young adult stars on the Main Sequence (Röser 2018).

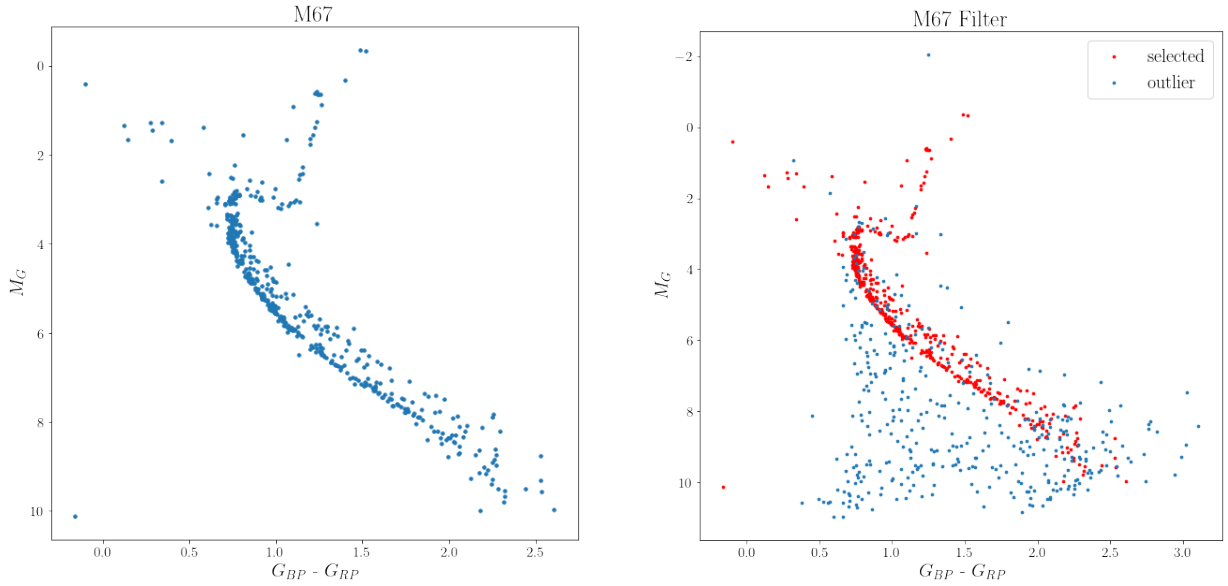

Figure 11. CMD for selected M67 members. This is the most exaggerated filter, with most of the lower flair on the CMD being filtered out in proper motion space. There are unexpected points to the left of the Main Sequence here: the upper outliers are blue stragglers (bluer than stars at the Main Sequence turnoff point), and there is one misclassified white dwarf in the bottom left (Nguyen 2022). Regardless, M67 hosts stars in all stages of their evolutionary cycle, with many along the Main Sequence and a significant amount of red giants before the turnoff point.

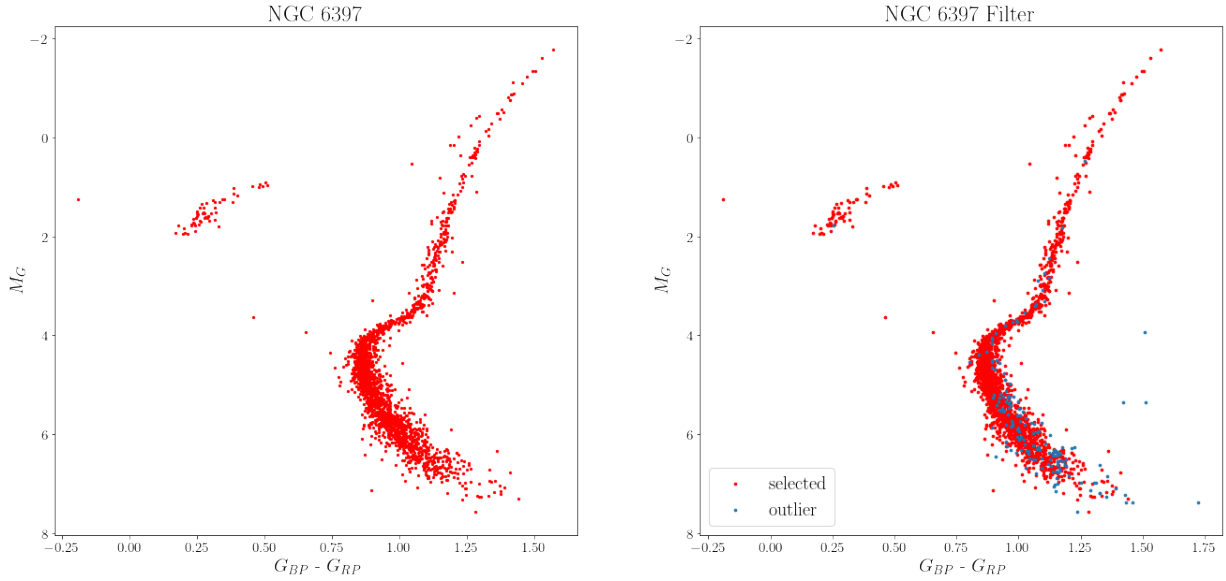

Figure 12. CMD for selected NGC 6397 members. As mentioned before, NGC 6397 is a dense, spatially concentrated cluster, and we have had to do minimal filtering to expose its true members. Therefore, the overlaid graph is quite unimpressive, we have simply filtered out a few of the redder stars straying from the Main Sequence (Torres 2019).

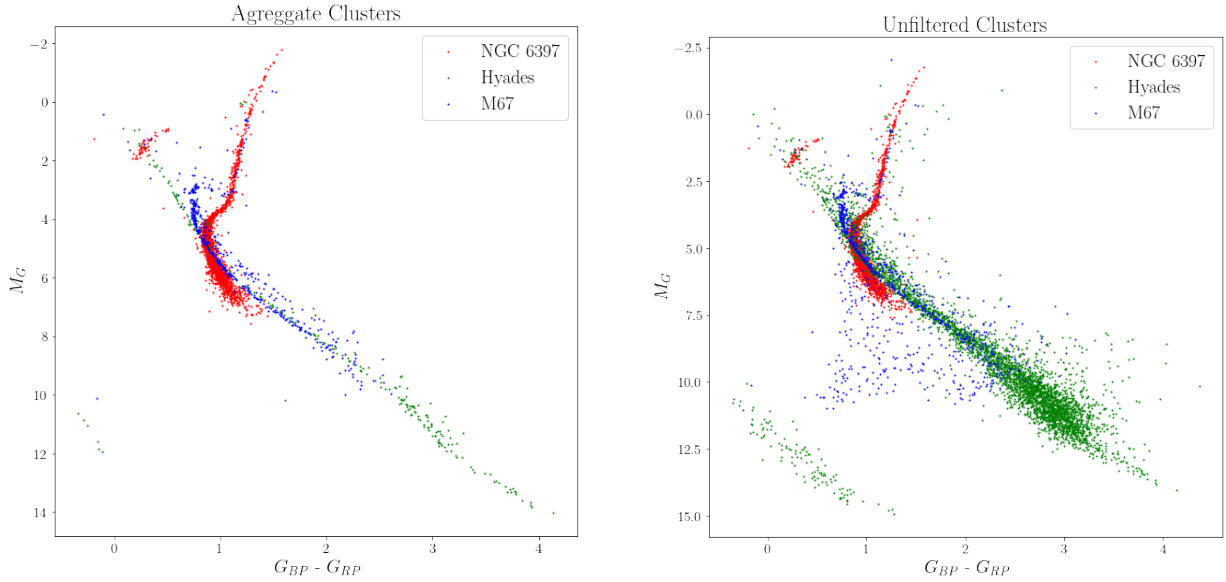

Figure 13. CMD for selected members from all three clusters. Individual CMDs are overlaid to be exhaustive. Throughout the selected data, the Main Sequence is apparent, and most of the sources fall along it. We have a significant amount of blue and red giants, found above to the left and right of the main sequence turnoff, respectively. The plot on the right with all the initial queried data is much more sparse and ill-defined; there are many outlier points and faux members of the selected clusters. The majority of stars still fall along the Main Sequence, but the stray sources with no correlation clearly represent outliers to these three clusters.

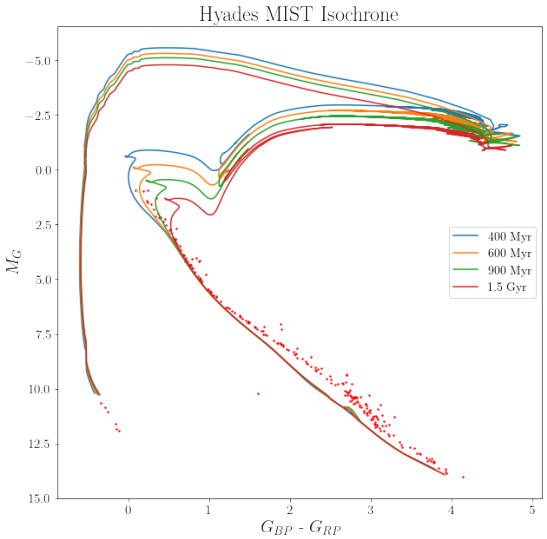

Figure 14. For clusters with ages in this range, the structure of the Main Sequence seems to be identical. The differences lie in the post-MS evolutionary tracks, of which the theoretical isochrone for a 600 Myr old cluster seems to match selected data the best. These isochrones even suggest that the white dwarfs are viable sources of this cluster, and are just very old and far along their stellar lifecycles. The Main Sequence does mismatch with data though, the MIST predictions are roughly one magnitude fainter on the CMD. As seen in later plots, isochrones for these ages were generated with a variety of metallicities (of each magnitude). For the Hyades cluster, we found that a metallicity $Z = 0.15$ led to the most accurate isochrones.

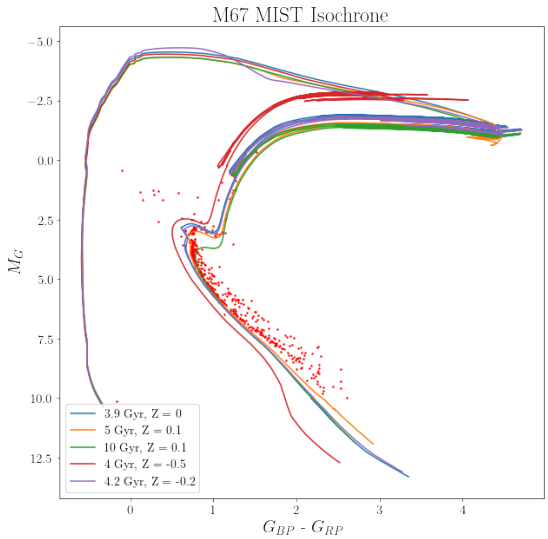

Figure 15. We plotted isochrones in an age range centered around 4 Gyr against a metallicity within [0, 0.1]. From the resulting isochrones, we can see that M67 has a low metallicity ($Z = 0.05$) and indeed has a range of roughly 4 Gyr. Most of the Main Sequence stars are concentrated in a line along the bottom of the band, which aligns with the theoretical isochrones. The only unexplained points are those to the left of the Main Sequence turnoff (blue stragglers)

.

5.1. *MIST Isochrone Fitting*

MESA Isochrones & Stellar Tracks (MIST[4]) was created in 2012 with the hopes of creating stellar evolutionary models with an unprecedented range in parameter space (Choi 2016). In this section, we will use theoretical isochrones generated by MIST to overlay synthetic photometry into our cluster selections. For each cluster, we will make initial guesses for the age and metallicity. While holding one constant, we will test a range on the other (first starting with orders of magnitude), such that we eventually converge at an optimal combination. This is almost like a manual gradient descent to the solution. If the isochrones were not so visually different, fitting one to the data points would be a good use case for any machine learning algorithm, even a simple linear least squares (LLS). Once we have identified an isochrone that fits the data well, we can directly estimate the age and metallicity of each cluster, listed in Table 5.3. Refer to Figure 14, Figure 15, Figure 16.

[4] https://waps.cfa.harvard.edu/MIST/

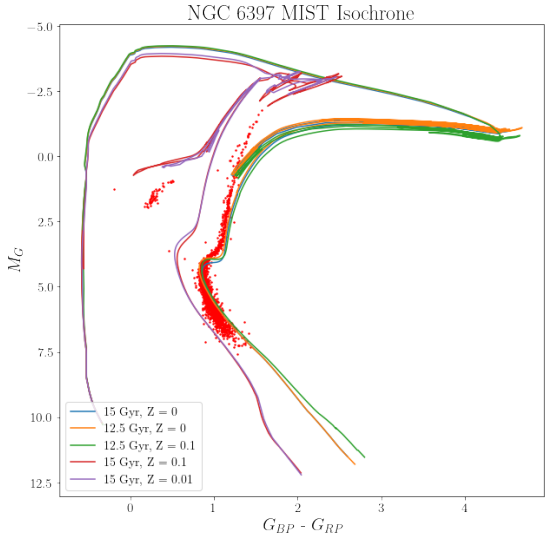

Figure 16. Since the discrepancies in ages are much larger at this magnitude, there is an apparent difference in the age bands, and we can tell that NGC 6397 tends towards being 12.5 Gyr old. Additionally, it has almost no metallicity, such that we can let $Z = 0.01$. The isochrones fit the Main Sequence fairly well, but do not fit the blue giants or the brightest red giants.

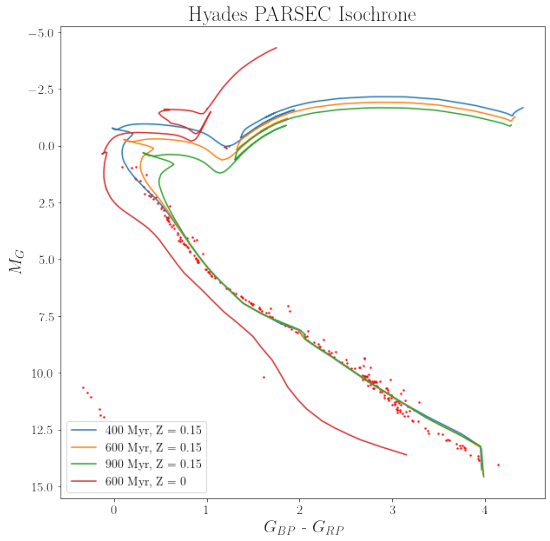

Figure 17. Once again the 600 Myr theoretical isochrone matches the data best for metallicities of $Z = 0.15$, so we have confirmed the conclusions drawn from the MIST model. However, the Main Sequence of Hyades is fit much better with this PARSEC model, which is apparent as the isochrone matches the cluster shape through the bottom right of the plot.

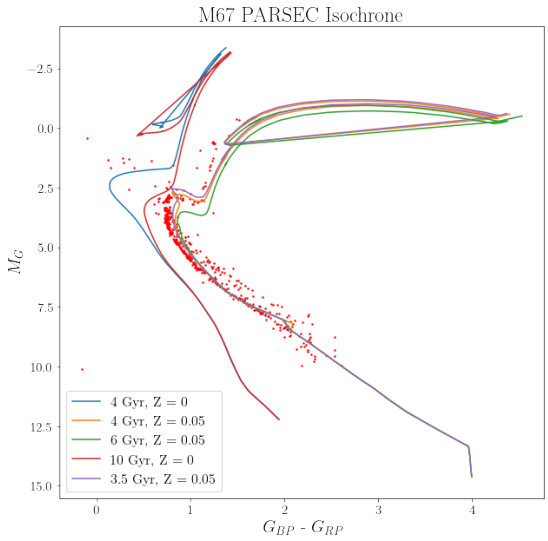

Figure 18. Building off of the last plot, we have ensured that the isochrone corresponding to an age of 4 Gyr and a metallicity of $Z = 0.05$ fits the data better than any other presented options. Once again, the Main Sequence of M67 is fit much better by the PARSEC isochrone, and the blue stragglers remain uncharted (their persistence implies they are legitimate cluster members). This isochrone also performs better at the turnoff point, where there is a plethora of stray sources which could not be filtered.

5.2. *PARSEC Isochrone Fitting*

PARSEC[5] is also an incredibly powerful tool to create synthetic photometry and theoretical isochrones given age and metallicity as inputs (Marigo 2012). While MIST computes its tracks with Modules for Experiments in Stellar Astrophysics (MESA) code, PARSEC has done the same with PAdova and TRieste Stellar Evolution Code. We will once again generate and overlay these isochrones to estimate the age and metallicity of our clusters, and will compare with the results found using the MIST models. Refer to Figure 17, Figure 18, Figure 19.

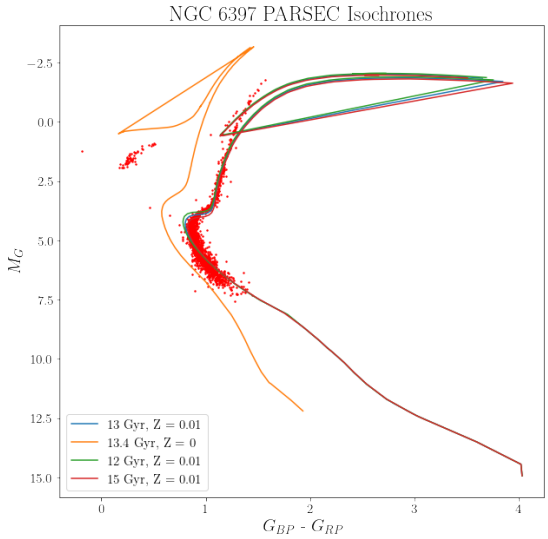

Figure 19. Once again, we have confirmed our findings from the MIST model, NGC 6397 can be estimated to have an age of 13.4 Gyr and a metallicity of $Z = 0.01$. And once again, the PARSEC data fits the Main Sequence more accurately. In this case, it is even able to fit the beginnings of the red giant branch, but still leaves the blue giants untouched.

5.3. *Comparison*

After plotting isochrones from both tools, it is apparent that the PARSEC isochrone definitely fit every cluster's Main Sequence better than its MIST counterpart. The PARSEC isochrones also managed to fit the red giant branch turnoffs for clusters which had them, but exhibited unexpected behavior for sources with very small or subzero absolute magnitudes. However, these magnitudes are not present in this data, so we can ignore their evolutionary implications. Unfortunately, neither model was able to explain the blue giants as a part of the cluster's evolutionary path. Nonetheless, PARSEC isochrones fit the data much better than MIST models, allowing us to estimate ages and metallicities more precisely; our results can be found in Table 5.3 below.

Table 5.2. Astrometric Data

CLUSTER	RA [deg]	DEC [deg]	PARALLAX [mas]	PMRA [mas/yr]	PMDEC [mas/yr]	METALLICITY [Fe/H]	AGE [Gyr]
Hyades	66.75	15.8533	21.1	101.	-28.5	0.15	0.6
M67	132.825	11.816	1.4	-10.9737	-2.9396	0.05	4
NGC 6397	265.175	-53.6743	0.35	3.30	-17.60	0.01	13.3

REFERENCES

Babusiaux, C. 2018, Gaia Data Release 2: Observational Hertzsprung-Russell diagrams. https://arxiv.org/abs/1804.09378

—. 2021, Gaia Early Data Release 3: Summary of the contents and survey properties. https://arxiv.org/abs/2012.01533

Choi, J. 2016, MESA ISOCHRONES AND STELLAR TRACKS (MIST). I: SOLAR-SCALED MODELS. https://arxiv.org/abs/1604.08592

Eyer, L. 2018, Gaia Data Release 2: Variable stars in the colour-absolute magnitude diagram. https://arxiv.org/abs/1804.09382

Marigo, P. 2012, PARSEC: stellar tracks and isochrones with the PAdova and TRieste Stellar Evolution Code. https://arxiv.org/abs/1208.4498

Nguyen, C. T. 2022, PARSEC V2.0: Stellar tracks and isochrones of low and intermediate mass stars with rotation. https://arxiv.org/abs/2207.08642

Röser, S. 2018, The Hyades tidal tails revealed by Gaia DR2. https://arxiv.org/abs/1811.03845

Torres, S. 2019, The white dwarf population of NGC 6397. https://arxiv.org/abs/1507.08806

Wang, Y. 2022, Detection of spatial clustering in the 1000 richest SDSS DR8 redMaPPer clusters with Nearest Neighbor distributions. https://arxiv.org/abs/2112.04502

[5] http://stev.oapd.inaf.it/cgi-bin/cmd