

ASTR4004 Assignment 3

Wei Yun Tham U73929239

October 8, 2024

Task 2: Using ADQL to Search for Bright Stars Around the Open Cluster M67

This task involves querying the Gaia DR3 dataset and crossmatching it with the 2MASS catalog to identify bright stars in the vicinity of the open cluster M67. The focus is on applying quality cuts based on 2MASS photometry and Gaia parallaxes to narrow down the selection to stars with reliable data.

ADQL Query

The following ADQL query was used to search for bright stars around M67:

```
""
SELECT gaia.source_id, gaia.ra AS g_ra, gaia.dec AS g_dec,
       gaia.phot_g_mean_mag, gaia.bp_rp, gaia.parallax, gaia.pmra, gaia.pmdec,
       tmass.j_m, tmass.h_m, tmass.ks_m, tmass.ph_qual
FROM gaiadr3.gaia_source AS gaia
JOIN gaiadr3.tmass_psc_xsc_best_neighbour AS xmatch
  ON gaia.source_id = xmatch.source_id
JOIN gaiadr3.tmass_psc_xsc_join AS tmass_join
  ON xmatch.clean_tmass_psc_xsc_oid = tmass_join.clean_tmass_psc_xsc_oid
JOIN gaiadr1.tmass_original_valid AS tmass
  ON tmass_join.original_psc_source_id = tmass.designation
WHERE 1=CONTAINS(
  POINT('ICRS', gaia.ra, gaia.dec),
  CIRCLE('ICRS', 132.825, 11.8, 1))
AND gaia.phot_g_mean_mag < 14
ORDER BY gaia.source_id ASC;
""
```

Results

- Number of stars returned from the initial query: 1018
- Number of stars with bad 2MASS photometry (ph_qual not 'AAA'): 21
- Number of stars with non-positive parallaxes: 2
- Number of stars remaining after applying quality cuts: 988

Conclusion

After applying quality cuts, 988 stars remain in the dataset. These quality cuts ensure that only stars with reliable 2MASS photometry and positive parallaxes are considered. This dataset can be used for further analysis, and the number of stars suggests a manageable number of fibres would be needed for spectroscopic follow-up.

Plot

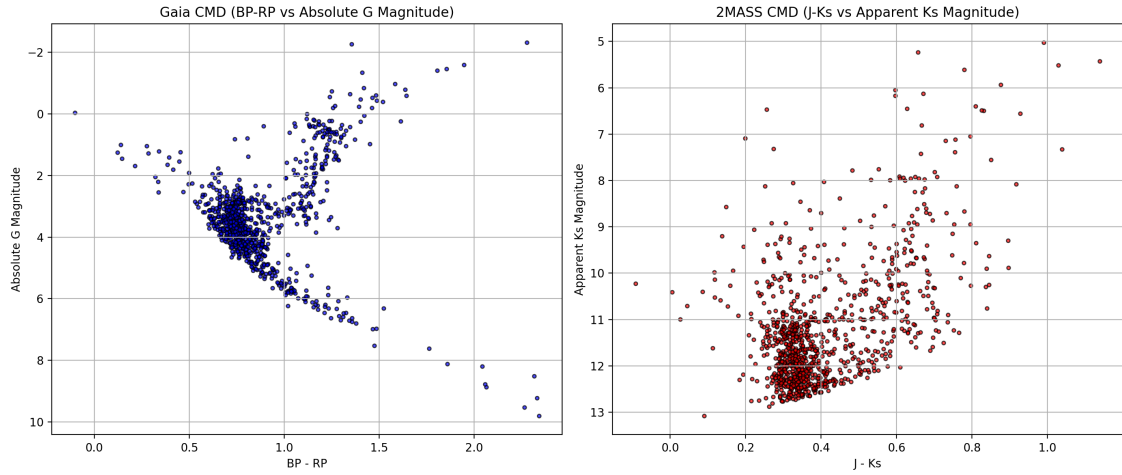


Figure 1: Colour-magnitude diagram (CMD) of Gaia BP-RP vs. absolute G magnitude (left) and 2MASS J-Ks vs. apparent K magnitude diagram (right)

Recommendation

Based on the number of stars remaining after applying quality cuts (988), I recommend moving forward with a proposal for spectroscopic fibre allocation. The reduced dataset ensures efficient use of fibres, focusing only on stars with good photometric and parallax data. By removing stars with bad 2MASS photometry and non-positive parallaxes, the dataset is refined, allowing the fibres to be allocated to scientifically valuable targets. The cuts reduce the number of stars from 1018 to 988, which makes fibre usage more efficient by ensuring that only stars with reliable data are observed. This approach balances the need to observe a substantial number of stars while maintaining the quality of the data, making the fibre allocation process more resource-efficient.

TASK 3: The radial metallicity relation in simulated data

Plot

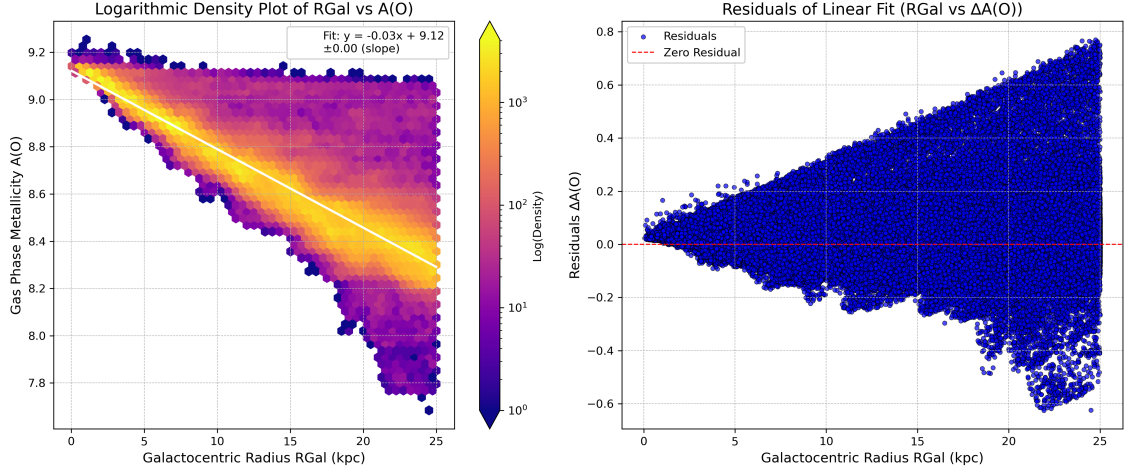


Figure 2: Logarithmic density plot of R_{Gal} vs. $A(O)$ with linear fit and legend (left) and residuals of the fit: R_{Gal} vs. $\Delta A(O)$ (right)

Model Performance

The model was evaluated using several performance metrics to assess its accuracy and fit to the data. Below are the key metrics:

- **Root Mean Square Error (RMSE):** 0.088
- **R-squared:** 0.873
- **Mean of residuals:** 0.000
- **Maximum absolute residual:** 0.770

Overall Fit of the Linear Model

- **Root Mean Square Error (RMSE):** The RMSE of 0.088 suggests that the typical deviation between the observed data and the predictions made by the linear model is about 0.088 in the units of $A(O)$ (gas-phase metallicity). This is a reasonably low value, indicating that the model does a fairly good job of predicting $A(O)$ based on R_{Gal} (galactocentric radius).
- **R-squared (R^2):** The R^2 value of 0.873 indicates that 87.3% of the variance in the gas-phase metallicity $A(O)$ is explained by the linear relationship with R_{Gal} . This is a strong indication of a good linear fit, but it also means that about 12.7% of the variation in $A(O)$ remains unexplained by the model.
- **Interpretation:** While the linear model captures most of the trend, there are likely other factors (or non-linearities) that contribute to the remaining variance.
- **Mean of Residuals:** The mean of the residuals is 0, which is a good sign as it indicates that the model does not systematically overestimate or underestimate the values of $A(O)$.
- **Maximum Absolute Residual:** The maximum absolute residual of 0.770 indicates that there are individual data points where the model performs poorly. These outliers likely represent regions where the relationship between $A(O)$ and R_{Gal} is not well captured by a linear model.

Where the Model Fits Well

- **Small Residuals in the Core Regions:** The linear model performs well in the *central regions of the galaxy* (where R_{Gal} is moderate), as evidenced by the residuals plot. In these regions, the trend of metallicity decreasing with increasing radius is well captured by the model. The RMSE of 0.088 further supports that, on average, the model is quite accurate in most regions.
- **Overall Fit ($R^2 = 0.873$):** The high R^2 value indicates that the linear model does a good job of explaining most of the variability in the data, especially in regions where the relationship between $A(O)$ and R_{Gal} is approximately linear.

Where the Model Does Not Fit Well

- **Large Residuals and Outliers:** The largest residuals occur in regions where the relationship between $A(O)$ and R_{Gal} is likely more complex than a simple linear trend. For example, at larger distances from the galactic center (higher R_{Gal}), the model may not adequately capture the behavior of the data. This could be due to non-linear effects or physical processes that affect metallicity, such as gas accretion, outflows, or radial migration, which are not accounted for in a simple linear model.
- **Non-linear Trends:** In regions with larger residuals, a non-linear model might perform better. For example, the presence of inflection points in the metallicity gradient may indicate that a polynomial or more complex model would be better suited to capturing the true relationship between $A(O)$ and R_{Gal} .

Interpretation of Statistical Metrics

- **RMSE:** The RMSE value of 0.088 shows that, on average, the linear model makes predictions that are relatively close to the observed values. However, the maximum residual of 0.770 suggests that there are individual points where the model's performance breaks down.
- **R-squared:** The R^2 value of 0.873 is strong and shows that most of the variation in the data is explained by the linear model, but the presence of unexplained variance suggests that there are factors not captured by the model.

Conclusion and Recommendations

The **linear model fits well in most regions**, as evidenced by the high R^2 and low RMSE values. However, **in regions with large residuals**, especially at larger distances from the galactic center, the model struggles to capture the true behavior of the data.

A possible improvement would be to explore **non-linear models**, such as a polynomial regression or a piecewise linear model, which could better capture the complexity of the relationship between $A(O)$ and R_{Gal} , particularly in the outer regions of the galaxy.

Plot

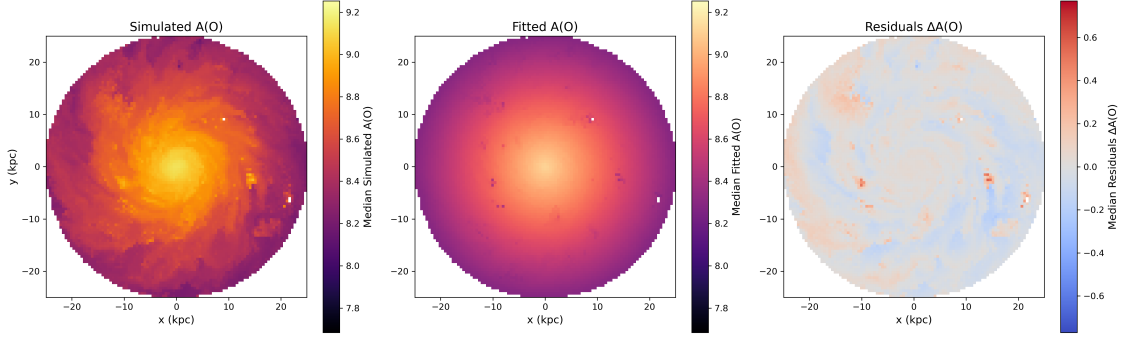


Figure 3: 2D-histogram of the median simulated $A(O)$ (left), 2D-histogram of the median fitted $A(O)$ (middle) and 2D-histogram of the median residuals $\Delta A(O)$ (right)

Choice of 2D Bins

We selected 100 bins along both the x and y axes to achieve a balance between resolution and readability. This bin size provides sufficient detail to capture spatial variations in gas-phase metallicity $A(O)$, revealing localized features such as the spiral structure of the galaxy and variations near the galactic center. Fewer bins (e.g., 20-50) would result in less precision, smoothing out critical details like the spiral arms and finer metallicity gradients. This would blur important features, making it harder to detect discrepancies between the model and the data. On the other hand, using more bins (e.g., 200 or more) could overcomplicate the plot, introducing noise that obscures the major trends. With more bins, each bin would contain fewer data points, reducing the statistical significance of the median values and increasing random variation, which could lead to misinterpretation of the data. Therefore, 100 bins offer an optimal compromise to reveal significant patterns while maintaining clarity.

Analyze the residuals

The residuals in the spiral structure reveal that the linear model struggles to capture the complex variations in metallicity along the spiral arms of the galaxy. These regions are often sites of heightened star formation, where gas dynamics, such as inflows and outflows, lead to uneven distributions of metals. Consequently, the linear model falls short of accurately predicting $A(O)$ in these areas. By contrast, near the galactic center, the model performs well, as the residuals are minimal. This suggests that the central region has a smoother metallicity gradient, likely due to the well-mixed gas present in these areas. In the outer regions of the galaxy, particularly along the spiral arms, we observe larger residuals, both positive and negative, highlighting the model's inability to account for localized metallicity variations. Such variations may be driven by radial migration of stars, which can transport metal-rich stars between regions, further complicating the metallicity distribution. Additionally, regions with positive residuals (red) reflect areas of recent star formation, where the model underpredicts metallicity, while negative residuals (blue) indicate areas where the model overpredicts, likely due to inflows of low-metallicity gas or the presence of older stars. The spiral arms and outer regions exhibit complex dynamics that a simple linear model cannot fully capture, thus necessitating more sophisticated modeling techniques to address the non-linear metallicity variations driven by processes such as star formation, gas flows, and stellar migration.