Bayesian characterization of the young open cluster NGC 6383 using HDBSCAN and Gaia DR3

L.M. Pulgar-Escobar[1] & N.A. Henríquez-Salgado[1]

[1] *Departamento de Astronomía, Universidad de Concepción, Chile*

Received: 09 February 2024 / Accepted: 18 May 2024

Resumen / Este estudio se centra en determinar las características del joven cúmulo abierto NGC 6383. Para lograrlo, se utiliza el algoritmo de agrupamiento HDBSCAN para identificar los miembros potenciales del cúmulo basándose en los movimientos propios y los paralajes utilizando *Gaia Data Release 3*. Se obtienen varios parámetros de NGC 6383, como el *tidal radius*, el radio del núcleo, la distancia mediante paralaje y ajuste de isocronas, el movimiento propio promedio, la edad, la metalicidad y otros relevantes. Para realizar este análisis, utilizamos una extensión de Monte Carlo Hamiltoniano, el *No-U-Turn Sampler*. Los resultados de este análisis señalan que NGC 6383 es un cúmulo abierto muy joven ($\approx 1-4$ Myr), con una distancia de ≈ 1.1 kpc.

Abstract / This study focuses on determining the characteristics of the young open cluster NGC 6383. To achieve this, the HDBSCAN clustering algorithm is utilized to identify potential cluster members based on proper motions and parallaxes from *Gaia Data Release 3*. Various parameters of NGC 6383, such as tidal radius, core radius, distance through parallax and isochrone-fitting, proper motion, age, metallicity, and relevant others, are assessed. To perform this analysis, we utilize an extension of Hamiltonian Monte Carlo, the No-U-Turn Sampler. The results of this analysis point out that NGC 6383 is a very young open cluster ($\approx 1-4$ Myr), with a distance of ≈ 1.1 kpc.

Keywords / open clusters and associations: individual — galaxies: star clusters: general — stars: distances — techniques: photometric — parallaxes — proper motions

1. Introduction

Precise cluster's members identification is a crucial step when it comes to a correct identification and characterization of the cluster. Various methods have been used for this process (Lindoff, 1968; Fitzgerald et al., 1978; Pandey et al., 1989; Kharchenko et al., 2005; Paunzen et al., 2007), each offering different focuses and sometimes leading to consensus, or discrepancies. Recently, Bayesian analysis implementation on the membership identification, offers a different method to compare or compliment with the ones mentioned.

NGC 6383 * is a young open cluster situated in the Carina-Sagittarius arm, within the Sh 2-012 star formation region. It forms part of the larger Sagittarius OB1 association, along with NGC 6530 and NGC 6531. The galactic coordinates of the cluster are $\ell = 355.68°$ and $b = 0.05°$ (Rauw & De Becker, 2008).

The distance to the cluster has been estimated by various authors, with an upper limit of 2130 pc proposed by Trumpler (1930) and Zug (1937), and a lower limit of 760 pc and 834 pc suggested by Sanford (1949) and Aidelman et al. (2018), respectively.

The parameters of the cluster exhibit a wide range of results, contingent on the methodology utilized by the authors in their research. Consequently, the age of the cluster has been a topic of considerable debate.

The implementation of Bayesian analysis and Machine Learning techniques in this study, explores a different method to obtained various parameters for characterizing NGC 6383 and contribute to the debate.

2. Methodology

We acquired data from the *Gaia* third Data Release (DR3), executing a cone search of 25 arcmin radius on the *Gaia* archive, yielding 71847 sources. Initial selection filtered sources to distances between 500 pc and 2500 pc—equivalent to parallax ranges between 0.4 mas and 2 mas—based on limits established in prior research, resulting in 29092 sources.

To enhance data reliability, we applied astrometric fidelity parameters from Rybizki et al. (2022). This parameter, derived from a neural network analysis of 17 *Gaia* catalog metrics, assesses the astrometric solution's trustworthiness. Sources with an astrometric fidelity above 0.5 were retained, narrowing the selection to 20215 sources.

Subsequent refinement ensured inclusion of only sources with comprehensive parameters, finalizing the dataset at 19964 sources. Systematic parallax offsets identified in Lindegren et al. (2021) were corrected using the GAIADR3_ZEROPOINT package. Moreover, addressing the bias in proper motion for bright sources

*NGC 6383, also known as NGC 6374 in the New General Catalog and classified as Collinder 334 and Collinder 335 in the Collinder catalog, was initially misclassified in the original Collinder catalog (Collinder, 1931).

✉ *Contact* / lescobar2019@udec.cl

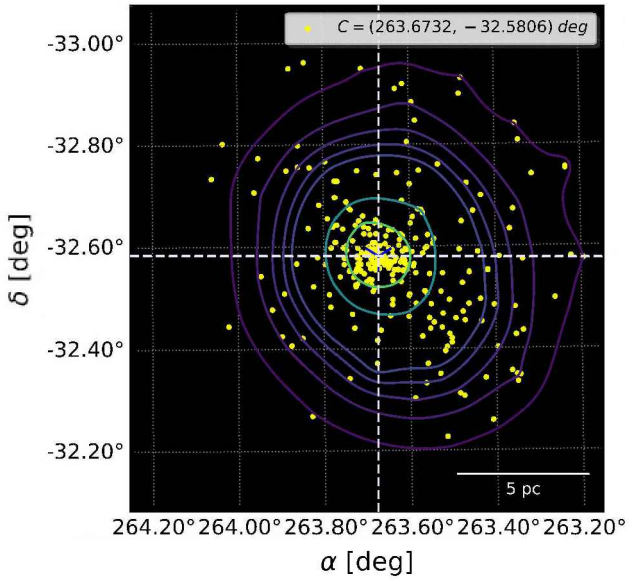

Fig. 1. Spatial distribution of sources in R.A. (α) and DEC. (δ), considering a probability of at least 60 percent. Concentric contour lines indicate levels of the KDE with exponential kernel, and the center of the distribution is marked by the white dashed lines.

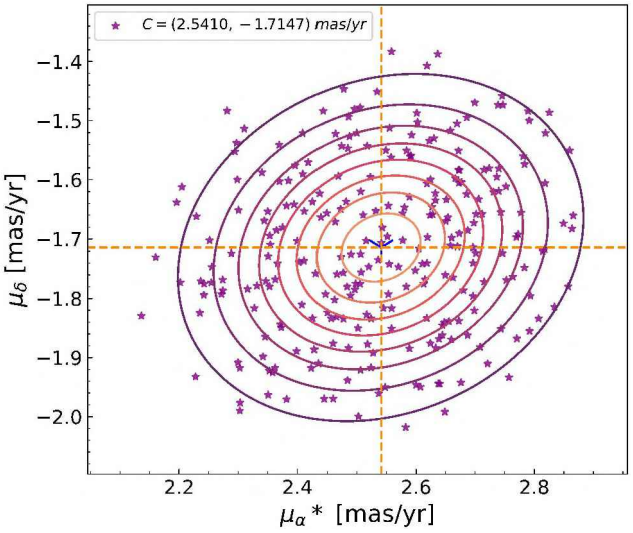

Fig. 2. Proper motions of sources within the cluster in R.A. (μ^*_α) and DEC. (μ_δ), considering a probability of at least 60 percent. Concentric contour lines indicate levels of the multivariate Gaussian, and the center of the proper motion distribution is marked by the orange dashed lines.

(G < 13 mag), as discussed by Cantat-Gaudin & Brandt (2021), we applied a magnitude-based correction for sources with G = 11 − 13 mag, compensating for up to 80 μas yr^{-1} discrepancy between the frames of reference for bright and faint sources.

2.1. COSMIC

Characterization Of Star clusters using Machine learning Inference and Clustering (COSMIC) developed by Lucas Pulgar-Escobar et al. (in prep), is a suite of functions designed for analyzing open clusters. Utilizing unsupervised machine learning algorithms, COSMIC processes extensive datasets, such as those from *Gaia*, to identify fundamental parameters of open clusters through clustering techniques and Bayesian estimation. As an open-source program, it is developed in Python 3.11 and integrates PyMC 5.10**, a Python library specialized in Bayesian analysis.

2.1.1. Membership determination

To identify potential members of NGC 6383, we employed the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm (Campello et al., 2013), known for its effectiveness in identifying clusters of varying densities and sizes without requiring a predetermined number of clusters (Hunt & Reffert, 2021). This density-based clustering method, which constructs a hierarchical representation of the data (McInnes et al., 2017), is particularly adept at handling different cluster shapes and sizes while distinguishing noise points. Its capacity to automatically determine the optimal cluster count makes HDBSCAN an ideal choice for this analysis, offering significant benefits over traditional clustering algorithms, especially in terms of robustness against noisy data and outliers.

HDBSCAN was applied*** to the 19964 sources using the proper motions as the clustering parameters, resulting in 544 sources with a probability over 0.5.

We applied the AstroPy Sigma Clipping utility (Astropy Collaboration et al., 2022) for a 2σ clipping around the median, yielding 399 probable members for parameters inference.

2.1.2. Parallax and Distance Estimation

Distance estimation from parallax is challenging due to measurement errors, and transforming parallax to distance is not straightforward and involves significant uncertainties, deviating from the simple *inverse of the parallax* approach (Bailer-Jones, 2015; Bailer-Jones et al., 2021). A crucial aspect is choosing an appropriate prior distribution. When analyzing the parallax measurements of stars within an open cluster and assuming a simplified one-dimensional perspective, we can see that the distribution of this parallax values follows a normal distribution. For accuracy, our analysis only includes members with fractional parallax errors under 0.1.

Our hierarchical model begins by calculating an initial average distance ($\mu_{\rm prior}$) from the observed parallaxes. The cluster's mean distance follows $\mathcal{U}(0.5\mu_{\rm prior}, 1.5\mu_{\rm prior})$ as a prior.

**www.pymc.io

***The algorithm's hyperparameters included an Euclidean distance metric and a minimum cluster size of 43.

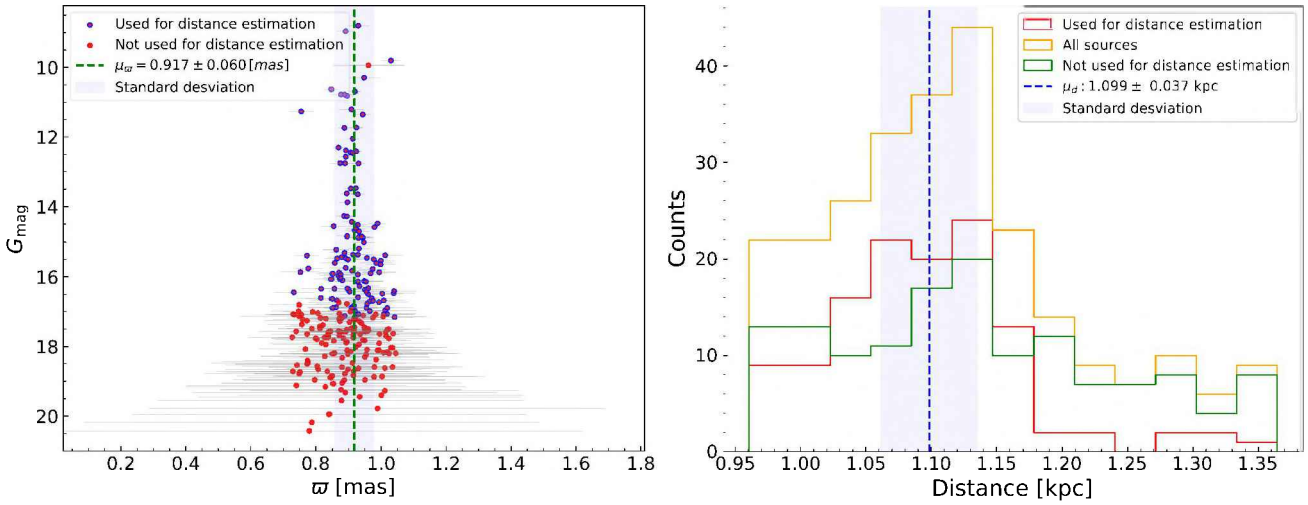

Fig. 3. *Left panel:* *Gaia* G-band magnitude ($G_{\rm mag}$) vs. parallax (ϖ), it's noticeable that fainter sources exhibit larger parallax errors, rendering them less reliable. Blue dots are sources with fractional parallax less than 0.1, and red ones are sources with values over 0.1, being excluded for the distance estimation. The green dashed line indicates mean parallax. *Right panel:* Histogram of the inverse values of the parallaxes as distances, with the same criteria as the left panel.

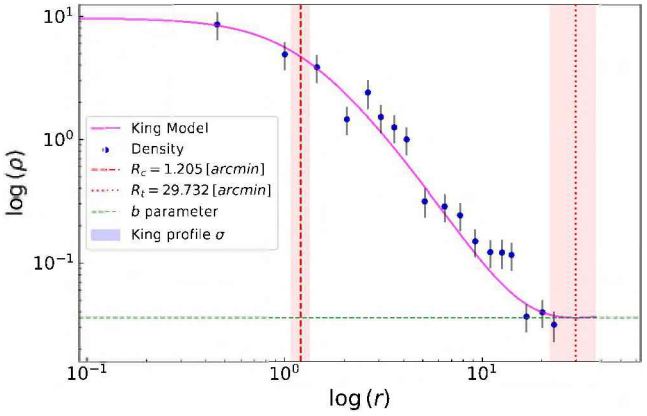

Fig. 4. Radial density profile of NGC 6383, with the blue dots representing the stellar density. The shaded blue zone outlines the deviation of the King's profiles, and the magenta line represents the median profile. The core radius R_c is shown as a red dashed line, the tidal radius R_t as a red dotted line, and the background density b as a green dashed line. The logarithmic axes denote the radius r in arcminutes and the stellar density ρ per arcminute squared.

2.1.3. Proper motions

To estimate the distribution of cluster members' proper motions, we model them using a two-dimensional normal distribution.

A normal prior distribution was assigned to the mean proper motions, based on frequentist means and standard deviations. The standard deviations are modeled using a HalfNormal distribution, while the correlation coefficient between the proper motions follows a $\mathcal{U}(-1, 1)$.

2.1.4. Center determination

To determine the cluster's center, we used a weighted Kernel Density Estimation (KDE) from SCI-KIT LEARN (Pedregosa et al., 2011), assigning weights inversely proportional to the distance from the mean proper motion. Optimal KDE parameters were determined through Grid Search Cross-Validation within the package, exploring a range of bandwidths—from the mean positional error to the search cone radius—and all kernel types. The location of maximum density was designated as the cluster's center.

2.1.5. Radial density profile

To determine structural parameters of the cluster, we utilized the King (1962) density profile. The numerical density ρ, was computed by dividing the cluster area into concentric annuli, each containing an equal number of stars. The number of annuli, K, adheres to the equiprobable bin rule ($K = 2n^{2/5}$), where n is the star count within the cluster.

Model priors were set as follows: background density $b \sim \mathcal{U}(0, 2\rho_{\rm min})$, scale factor $k \sim \mathcal{U}(0, 2\rho_{\rm max})$, core radius $R_c \sim \mathcal{U}(0, 0.8R_t)$, and tidal radius $R_t \sim \mathcal{U}(R_c, 1.5T_{\rm max})$. $T_{\rm max}$ is the maximum value among tidal potential radius, Hill radius, gravitational bound radius, and the maximum observed cluster member distance. This ensures the chosen parameter space for R_c and R_t is within astrophysical valid limits.

2.2. AsTeCA

AsTeCA, or Automated Stellar Cluster Analysis (Perren et al., 2015), is a suite of tools designed to automate the standard tests applied to stellar clusters to determine their basic parameters. In order to obtain accurate estimates for a cluster's metallicity, age and extinction

Table 1. Obtained results for the cluster, distance(ϖ) is the distance obtained by parallax analysis. On the other hand, D.M. indicates distance distance modulus.

Parameter	Value	Unit
Distance (ϖ)	1.099 ± 0.037	kpc
Distance (D.M.)	1.534 ± 0.258	kpc
Age	6.202 ± 0.036	log(age)
Metallicity (Z)	0.015 ± 0.014	-
Parallax (ϖ)	0.917 ± 0.060	mas
Number of members	266	stars
Absorption (A_V)	1.468 ± 0.081	mag
Core Radius (R_c)	1.205 ± 0.126	arcmin
Background (b)	0.035 ± 0.017	stars arcmin^{-2}
Distance Modulus	10.903 ± 0.327	mag
Tidal Radius (R_t)	29.732 ± 7.694	arcmin
Center Density (k)	10.508 ± 0.795	stars arcmin^{-2}
Cluster Center R.A.	263.673 ± 0.004	deg
Cluster Center DEC.	-32.580 ± 0.004	deg
Proper Motion R.A.	2.541 ± 0.007	mas yr^{-1}
Proper Motion DEC.	-1.714 ± 0.006	mas yr^{-1}

values, we use AsTeCA's isochrone fitting process. We used the PARSEC v1.2S isochrones (Bressan et al., 2012; Tang et al., 2014), with the Gaia EDR3 photometric system and Kroupa (2001, 2002) canonical two-part-power law IMF corrected for unresolved binaries, with a logarithmic age range between 6.0 and 7.9. All parameter priors are set as uniform distributions, and the median values of the resulting posterior distributions are informed.

3. Preliminary Results

The informed results and figures are obtained using sources with a membership probability exceeding 60 percent, ensuring a high level of confidence in the analysis. Table 1 presents a summary of the results, highlighting key parameters of NGC 6383. The cluster's center is accurately defined in Fig.1, while Fig.2 illustrates the proper motions. Insights into distance and parallax estimation are provided in Fig.3, and Fig.4 showcases the cluster's structural parameters alongside the fitted King's Profile. Additionally, Fig.5 displays the CMD of NGC 6383 alongside the best-fit isochrone.

4. Conclusion and future work

Our study employed Bayesian methods and HDBSCAN to refine parameters like distance, proper motion, and age for NGC 6383 using Gaia DR3.

Future efforts will broaden member analysis for a complete census and incorporate radial velocity for dynamic insights. Also, the use of a larger search cone to find possible stars in the tidal region of the cluster. Additionally, incorporating multi-wavelength data could enhance the accuracy of determining stellar properties. Finally, a focused investigation into the binary and multiple systems within NGC 6383 like the central binary HD 159176 could offer valuable information about its star formation history.

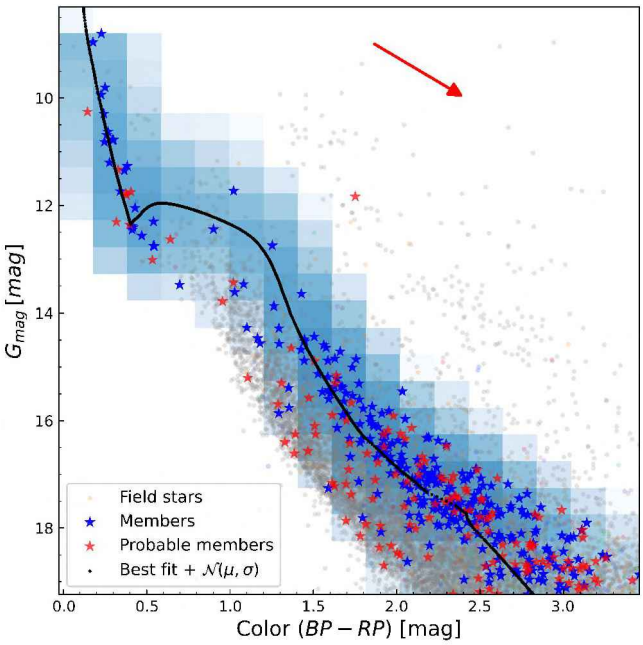

Fig. 5. Color-Magnitude Diagram of the cluster. The best isochrone fitting obtained by AsTeCa is indicated with a black line. Blue stars are sources with probabilities over 60 percent, while red stars are sources with probabilities under 60 percent. The red arrow is the reddening vector. The shaded area is a normal distribution around the best fit.

Acknowledgements: We gratefully acknowledge support by the ANID BASAL project FB210003.

References

Aidelman Y., et al., 2018, A&A, 610, A30
Astropy Collaboration, et al., 2022, apj, 935, 167
Bailer-Jones C.A.L., 2015, PASP, 127, 994
Bailer-Jones C.A.L., et al., 2021, AJ, 161, 147
Bressan A., et al., 2012, MNRAS, 427, 127
Campello R.J.G.B., Moulavi D., Sander J., 2013, 160–172, Springer Berlin Heidelberg, Berlin, Heidelberg
Cantat-Gaudin T., Brandt T.D., 2021, A&A, 649, A124
Collinder P., 1931, Annals of the Observatory of Lund, 2, B1
Fitzgerald M.P., et al., 1978, MNRAS, 182, 607
Hunt E.L., Reffert S., 2021, A&A, 646, A104
Kharchenko N.V., et al., 2005, A&A, 438, 1163
King I., 1962, AJ, 67, 471
Kroupa P., 2001, MNRAS, 322, 231
Kroupa P., 2002, Science, 295, 82
Lindegren L., et al., 2021, A&A, 649, A2
Lindoff U., 1968, Arkiv for Astronomi, 5, 1
McInnes L., Healy J., Astels S., 2017, The Journal of Open Source Software, 2
Pandey A.K., et al., 1989, MNRAS, 236, 263
Paunzen E., Netopil M., Zwintz K., 2007, A&A, 462, 157
Pedregosa F., et al., 2011, Journal of Machine Learning Research, 12, 2825
Perren G.I., Vázquez R.A., Piatti A.E., 2015, A&A, 576, A6
Rauw G., De Becker M., 2008, vol. 5, 497
Rybizki J., et al., 2022, MNRAS, 510, 2597
Sanford R.F., 1949, ApJ, 110, 117
Tang J., et al., 2014, MNRAS, 445, 4287
Trumpler R.J., 1930, Lick Observatory Bulletin, 420, 154
Zug R.S., 1937, Lick Observatory Bulletin, 489, 89