A&A, 689, A18 (2024)
https://doi.org/10.1051/0004-6361/202450901

Astronomy & Astrophysics

Exploring Galactic open clusters with *Gaia*

I. An examination in the first kiloparsec★

Jeison Alfonso[1], Alejandro García-Varela[1,★★], and Katherine Vieira[2]

[1] Universidad de los Andes, Departamento de Física, Cra. 1 No. 18A-10, Bloque Ip, A.A. 4976 Bogotá, Colombia
[2] Instituto de Astronomía y Ciencias Planetarias, Universidad de Atacama, Copayapu 485, Copiapó 1531772, Chile

Received 28 May 2024 / Accepted 17 July 2024

ABSTRACT

Context. Since the first publication of the *Gaia* catalogue, a new view of our Galaxy has arrived. Its astrometric and photometric information has improved the precision of the physical parameters of open star clusters obtained from them.
Aims. Using the *Gaia* Data Release 3 (DR3) catalogue, our aim was to find physical stellar members including faint stars for 370 Galactic open clusters located within 1 kpc. We also estimated the age, metallicity, distance modulus, and extinction of these clusters.
Methods. We employed the HDBSCAN algorithm on both astrometric and photometric data to identify members in the open clusters. Subsequently, we refined the samples by eliminating outliers through the application of the Mahalanobis metric utilizing the χ^2 distribution at a confidence level of 95%. Furthermore, we characterized the stellar parameters with the PARSEC isochrones.
Results. We obtained reliable star members for 370 open clusters with an average parallax error of $\sigma_{\varpi} = 0.16$ mas. We identified about ∼40% more stars in these clusters compared to previous work using the *Gaia* DR2 catalogue, including faint stars as new members with $G \geq 17$. Before the clustering application we corrected the parallax zero-point bias to avoid spatial distribution stretching that may affect clustering results. Our membership lists include merging stars identified by HDBSCAN with astrometry and photometry. We note that the use of photometry in clustering can recover up to 10% more stars in the fainter limit than clustering based on astrometry only; this combined with the selection of stars filtered out by quality cuts significantly reduces the number of stars with huge σ_{ϖ}. After clustering, we estimated age, Z, and A_V from the photometry of the membership lists.
Conclusions. We carried out a search to extend the membership list for 370 open clusters mainly on the Galactic plane in a neighbourhood of 1 kpc. Our methodology provides a robust estimator for the identification of outliers and also extends the membership lists to fainter stars in most of the clusters. Our findings suggest the need to carefully identify spurious sources that may affect clustering results.

Key words. open clusters and associations: general – methods: data analysis – Galaxy: disk

1. Introduction

The Milky Way disk has been the birthplace of open clusters that continually evolve over time (Lada & Lada 2003; McKee & Ostriker 2007). The interaction between these stellar systems and the Galactic gravitational potential shapes their collective dynamics allowing us to understand the evolutionary state of the Galaxy. As clusters age, their members gradually disperse into the surrounding increasing the field population. This phenomenon presents an opportunity to test stellar and gravitational models (Küpper et al. 2015). The loss of stellar members takes place through the process of relaxation, sometimes also known as dissolution or evaporation (Krumholz et al. 2019), resulting in the formation of elongated structures characterized by over- and underdensities attributed to gravitational interactions with giant molecular clouds or with the spiral arms or the bulge of the Milky Way (Binney & Tremaine 2008). These stars leave the cluster through the Lagrangian points at velocities that slightly exceed the escape velocity (Küpper et al. 2008).

Currently, the open cluster census is increasing, largely attributed to the wealth of data provided by the *Gaia* Data Release 3 (DR3; Gaia Collaboration 2016; Gaia Collaboration 2023; Babusiaux et al. 2023). According to Hunt & Reffert (2023), approximately seven thousand clusters have been identified. To determine the star members of these systems, one approach can be the application of unsupervised machine learning techniques based on density hierarchical algorithms, as done by Castro-Ginard et al. (2018); Castro-Ginard et al. (2022) and Cantat-Gaudin et al. (2020) for Galactic open cluster.

Age and metallicity play a fundamental role in understanding processes such as mass segregation and the expansion of open clusters (Della Croce et al. 2024). Recent studies based on spectroscopic and photometric data have found different distributions of these parameters (Yong et al. 2012; Anders et al. 2017; Spina et al. 2017). Once cluster members have been identified using *Gaia* data, their age and metallicity can be determined by isochrone fitting.

For this work our aim was to obtain reliable stellar members and cluster parameters for 370 nearby open clusters, constrained within 1 kpc, using Cantat-Gaudin et al. (2020) as a reference catalogue. Before the clustering implementation, we corrected the zero-point parallax bias for all sources based on the functions presented by Lindegren et al. (2021). Then, we utilized the HDBSCAN algorithm, computing distances with the Mahalanobis metric in two feature spaces: one comprised of equatorial coordinates, proper motions and parallaxes, and the other with two additional photometric colours compared to the previous set of features. The cluster members were selected as the union of the

★ The tables with cluster parameters and their stellar members will be made available as online material at the CDS.
★★ Corresponding author; je.alfonso1@uniandes.edu.co

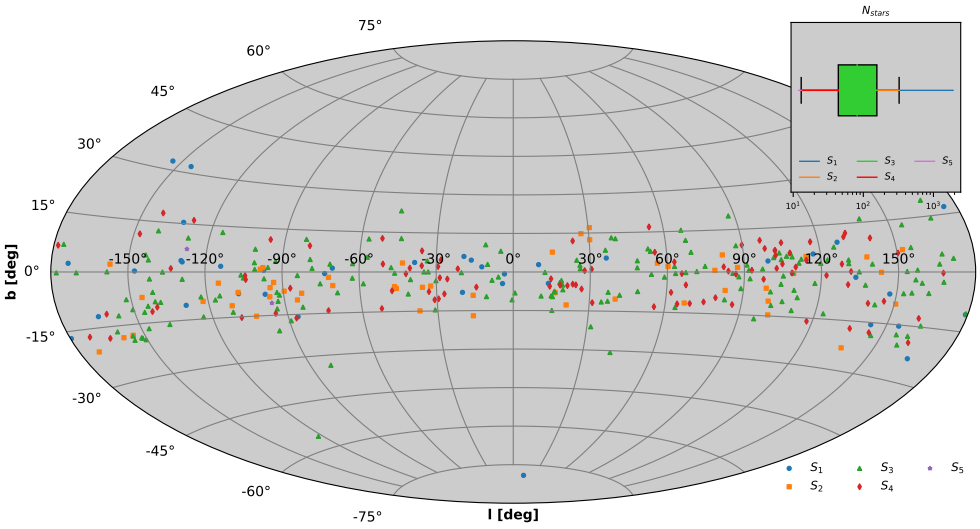

Fig. 1. Celestial map in Galactic coordinates for 370 open clusters selected from the Cantat-Gaudin et al. (2020) sample within a 1 kpc neighbourhood of the Sun. The colours represent open clusters belonging to some of the five sub-samples (S_i, for $i = 1, \ldots, 5$) defined from the box plot of their number of stars ($N_{\rm stars}$). The limits between sub-samples correspond to $N_{\rm stars} = 324, 156, 44, 13$, respectively.

members found in each feature space. Next, we removed outliers once again using the Mahalanobis metric considering a 95% confidence interval with the χ^2 distribution. Finally, we estimated the age, metallicity, extinction, and distance modulus to each cluster fitting the PARSEC tracks (Bressan et al. 2012; Marigo et al. 2013) to the colour–magnitude diagrams (CMDs) of the stellar members.

This document is structured in six sections, which are summarized below. Section 2 details the data selection within each open cluster region, based on equatorial coordinates and the criteria to ensure the inclusion of high-quality astrometric sources. Section 3 comprises three subsections: a detailed explanation of the `HDBSCAN` algorithm, the methodology for determining membership, and the procedures for identifying and removing outliers. Section 4 presents the outcomes of our analysis, including the determination of stellar membership within each cluster, the challenges encountered in dealing with parallax uncertainties for faint stars, and the methodology employed for inferring cluster parameters. Section 5 addresses the implications and significance of our results. Finally, our key conclusions and insights are summarized in Section 6.

2. Data

The identification of star members in open clusters using *Gaia* data, typically involves employing a magnitude cutoff around $G \sim 17$ or $G \sim 18$ to mitigate the impact of increased astrometric errors, particularly pronounced for distant clusters (Castro-Ginard et al. 2018; Cantat-Gaudin et al. 2020). Nevertheless, with the *Gaia* DR3 data, it is now feasible to detect star members in open clusters even in the faintest regions of the CMD, while still maintaining reliable astrometric measurements.

Among the 2017 open clusters in Cantat-Gaudin et al. (2020), we selected 370 clusters within 1 kpc. The Hyades and Coma clusters were excluded from our analysis due to their large apparent size caused by their proximity to the Sun. As the distribution of tidal radii of open clusters reported in the literature does not exceed 50 pc, we adopted this value as the maximum physical size of any Galactic open cluster. Taking into account the reported distance and coordinates of each cluster given by Cantat-Gaudin et al. (2020), we computed the maximum angular size covered in the sky for each of these systems. We took this apparent size as the search radius in the *Gaia* query, centred on their equatorial coordinates to download data. We also made a distance cutoff, centred on the mean cluster parallax, of ±400 pc in each region by converting it into parallax (mas) to avoid including field stars too distant from the cluster. In addition, to remove spurious sources and ensure high-quality data, we filtered out data requiring:

1. `ruwe` < 1.4,
2. `parallax_over_error` > 10,
3. `visibility_periods_used` > 6,
4. `astrometric_excess_noise` < 1,
5. `phot_{g,bp,rp}_mean_flux_over_error` > 10.

Making the box plot of the number of members ($N_{\rm stars}$) in each cluster reported by Cantat-Gaudin et al. (2020), our sample of 370 open clusters has been categorized into five distinct subsamples. Sub-sample S_1 has the highest star counts clusters (38 clusters, $N_{\rm stars} > 324$) while S_5 has the lowest ones (2 clusters, $N_{\rm stars} \leq 13$). Sub-sample S_2 comprises 53 clusters with $156 < N_{\rm stars} \leq 324$, while S_4 has 95 clusters with $13 < N_{\rm stars} \leq 44$. The central interquartile sub-sample is S_3 with 182 clusters having $44 < N_{\rm stars} \leq 156$. All of these sub-samples and the box plot are shown in Fig. 1.

3. Methods

In this section, we provide an overview of the clustering algorithm used to perform membership and the methodology to obtain star members in each open cluster, remove outliers, and infer their cluster parameters.

3.1. HDBSCAN

The Hierarchical Density-Based Spatial Clustering of Applications with Noise `HDBSCAN` (McInnes et al. 2017) is suitable for application to tabular *Gaia* data. This algorithm is based on implementing `DBSCAN` Ester et al. (1996) varying the epsilon value. `HDBSCAN` builds a hierarchy based on the minimum spanning tree graph from a mutual reachability distance computed with an input pairwise metric. Then, it begins to remove edges from the dense graph and look for cluster stability: the most stable ones are classified as clusters (Campello et al. 2013, and reference therein). Recently, Hunt & Reffert (2021) showed that this algorithm has the best performance on *Gaia* data in reducing false positives compared to DBSCAN and Gaussian Mixture models, it can recover clusters with many shapes and sizes which

allows to explore the substructures and halo populations in many clusters (Zhong et al. 2019). HDBSCAN has also significantly enriched the census of Galactic open clusters, enabling the discovery of new clusters and recovering the established ones in the literature. The main hyperparameters in HDBSCAN are the minimum number of samples in a cluster (min_cluster_size) and the number of samples in a neighbourhood (min_samples). These hyperparameters need to be carefully selected to have the best efficiency running on *Gaia* data.

To obtain clusters from the data, the HDBSCAN algorithm requires a metric to compute distances between points in the dataset. The Euclidean metric is the most commonly used, however, it requires that the correlation matrix has unit-variance (Feigelson & Babu 2012). This is true for independent data, but real-world data has different kinds of direct correlations. Indeed, in the *Gaia* catalogue there are correlations between the astrometric data at different levels. As mentioned in Gaia Collaboration (2023, see Sect. 4.5.7), correlations exist between astrometric parameters for the same source, between different sources for the same astrometric parameter, and between arbitrary astrometric parameters for different sources. Therefore, these issues may affect the results of clustering algorithms. One approach to deal with correlations between data is using the Mahalanobis metric to compute distances in the dataset (Mahalanobis 2018). To our knowledge, this metric has not been used with *Gaia* data. It is useful in multivariate analysis and can also be employed to identify outliers (see Sect. 3.3).

We suppose any two vectors $\boldsymbol{r_i}$ and $\boldsymbol{r_j}$ in the dimensional space with a positive-definite covariance matrix $\boldsymbol{S}$, and thus the Mahalanobis distance $d_M(\boldsymbol{r_i}, \boldsymbol{r_j})$ between $\boldsymbol{r_i}$ and $\boldsymbol{r_j}$ is

$$d_M(\boldsymbol{r_i}, \boldsymbol{r_j}) = \sqrt{(\boldsymbol{r_i} - \boldsymbol{r_j})^T \boldsymbol{S}^{-1} (\boldsymbol{r_i} - \boldsymbol{r_j})}. \quad (1)$$

This metric is passed to HDBSCAN as the pairwise metric, it takes into account any level of correlation in data and is closely to Hotelling's T^2 based on multivariate normal distribution (Feigelson & Babu 2012). It is worth mentioning that if the covariance matrix has unit-variance, e.i., the covariance matrix is the unity, it recovers to the well known standard Euclidean metric.

3.2. Membership determination

The current membership lists of Galactic open clusters include few or no faint stars due to catalogue limitations. Recent works (e.g. van Groeningen et al. 2023; Hunt & Reffert 2023) provide updated membership lists reaching the *Gaia* limit up to $G \sim 21$. At this level, the low-mass stars are strongly scattered from the main sequence, and this may affect cluster parameter estimates showing a clear need to be cautious at fainter magnitudes. Nevertheless, once we deal with error and correlation problems, it is thus possible to extend the star membership lists with reliable statistics even in the faint region of the CMDs.

Before the clustering implementation, we corrected the parallax zero-point bias given the ecliptic latitude, magnitude and colour of any source throughout the Python library gaiadr3_zeropoint[1], this allows us to obtain a corrected parallax (ϖ_P) for each star based on functions presented by Lindegren et al. (2021). To recover the stellar members in each open cluster we adopted two different approaches to apply HDBSCAN using the Mahalanobis metric given by Eq. (1). First, we used the common astrometric space with five features, we name it set 5F: (α, δ, $\mu_{\alpha*}$[2], μ_δ, ϖ_P). Then we used a combination of astrometry and photometry with seven features, 7F: (α, δ, $\mu_{\alpha*}$, μ_δ, ϖ_P, $G - G_{\rm RP}$, $G_{\rm BP} - G_{\rm RP}$), where we included two colour indexes. We note that adding photometry to perform clustering increases the number of fainter stars up to $G \sim 19.8$ in most of the clusters, but decreases the number of the brightest ones. Once the clustering algorithm has been performed on 5F and 7F, we merged star members from the two sets (5F∪7F) in order to have the larger amount of them. This produced a final list that includes reliable stellar members up to $G \sim 19.8$ that were undetected by the astrometric space 5F alone.

3.3. Outlier removal

Despite the quality cuts applied on *Gaia* DR3 data explained in Sect. 2, and that HDBSCAN is very efficient and robust in detecting clusters with different shapes, the membership candidates are not exempt from selecting outliers (Feigelson & Babu 2012; Hunt & Reffert 2021). They can be bona fide extreme objects in a single population, the result of systematic errors or contamination by spurious data. To remove these problematic data, we also employed the Mahalanobis distance. This metric cannot only be used as a pairwise metric as done in Sect. 3.1, but also to remove outliers. In this case, the Mahalanobis metric can be used as a measure of the distance between a point $\boldsymbol{r_i}$ in the dataset and a probability distribution Q. Following Eq. (1), for a vector $\boldsymbol{r_i}$ the Mahalanobis distance is

$$d_M(\boldsymbol{r_i}) = \sqrt{(\boldsymbol{r_i} - \boldsymbol{\mu})^T \boldsymbol{S}^{-1} (\boldsymbol{r_i} - \boldsymbol{\mu})}, \quad (2)$$

where $\boldsymbol{\mu}$ is the sample mean and $\boldsymbol{S}$ is the positive-definite covariance matrix. In other words, the distance given by Eq. (2) is a measure of the distance between a data point and the sample mean, thus, we can remove points that are far from $\boldsymbol{\mu}$ (Feigelson & Babu 2012).

We used a robust estimator for $\boldsymbol{\mu}$ and $\boldsymbol{S}$ to reduce the effect of the outliers on these values. For this, we implemented the Minimum Covariance Determinant (MCD) estimator developed by Rousseeuw & van Driessen (1999). This method uses a sub-sample of the dataset that minimizes the determinant of the covariance matrix. Then, based on the χ^2 distribution and assuming a multivariate normal distribution we can define an outlier threshold, in our case at $c = \sqrt{\chi^2_{p=5}}$ for the 95% confidence interval with p degrees of freedom: 5 for the astrometric space 5F. Once the threshold is defined, those data points with extreme values (i.e. $d_M(\boldsymbol{r_i}) > c$) are removed from the sample. Figures 2 and 3 shows the distribution of the Mahalanobis distances for 486 samples in the Roslund 6 open cluster selected by the clustering algorithm. By following the procedure described above, an outlier threshold of $c = 3.33$ was found using the χ^2 distribution. Those points in Fig. 3 with values above the cutoff represented by the horizontal dashed blue line are considered outliers. We found 57 stars identified as spurious data in this cluster. This procedure is applied only on the astrometric space 5F with equatorial coordinates, proper motions and parallax corrected by the zero-point bias. We opted not to include photometry (7F) because white dwarf and red giant stars may be classified as outliers due to their clear separation from the main sequence.

[1] https://gitlab.com/icc-ub/public/gaiadr3_zeropoint

[2] $\mu_{\alpha*} = \mu_\alpha \cos\delta$.

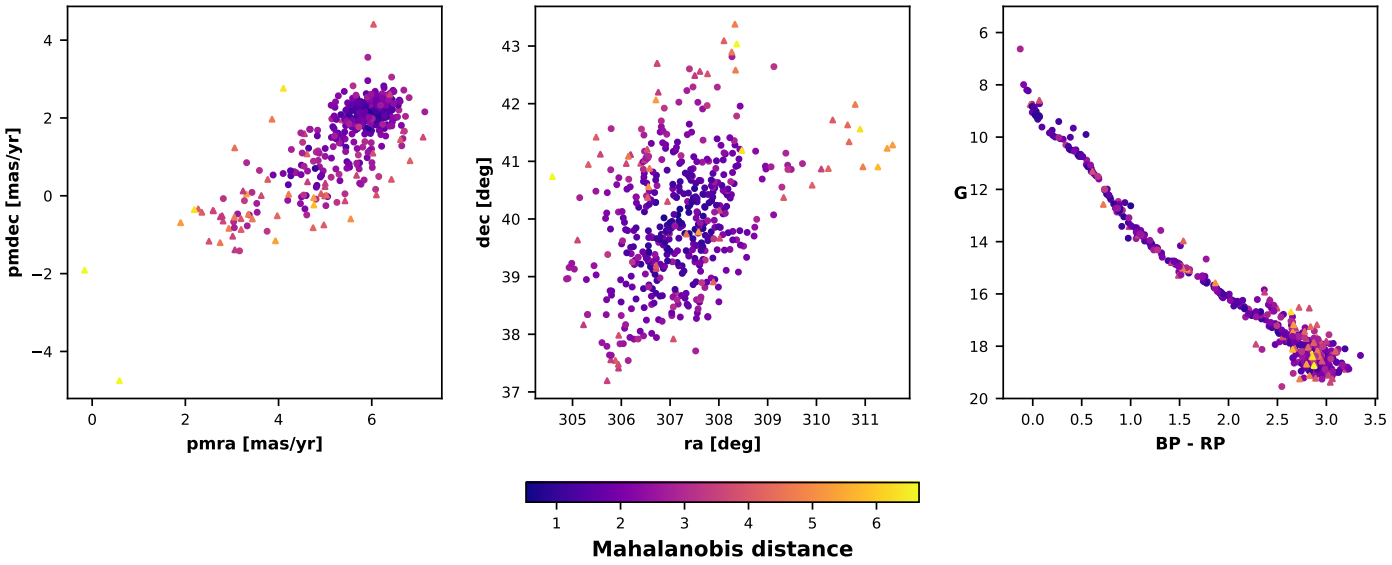

Fig. 2. Vector point diagram (left), equatorial coordinates (middle), and CMD (right) of the 486 members selected by HDBSCAN for the Roslund 6 open cluster. After an outlier threshold corresponding to 3.33 following the methodology presented in Sect. 3.3, the sample reduces to 429 with 57 stars identified as outliers. The cluster members and outliers are marked by circles and triangles, respectively. The colour bar indicates the Mahalanobis distances.

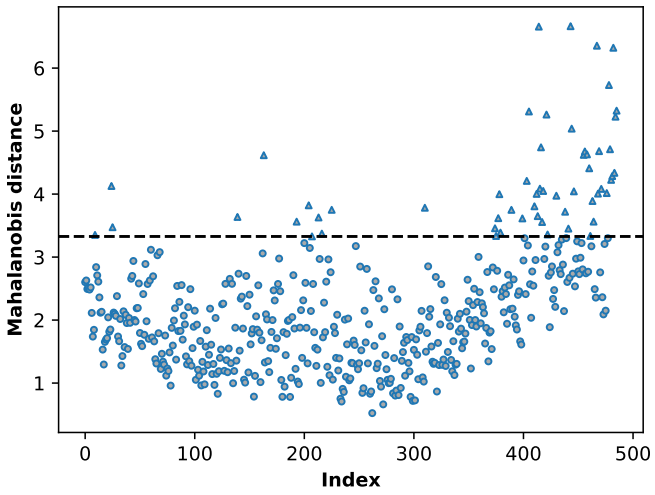

Fig. 3. Mahalanobis distance distribution for the Roslund 6 members selected by the HDBSCAN algorithm. The horizontal dashed black line is the outlier threshold with a value of $c = 3.33$. There are 57 stars identified as outliers in this open cluster, indicated by triangles above the dashed line. The horizontal axis represents an arbitrary star index assigned by HDBSCAN to the selected members.

4. Results

In this study we aimed to obtain reliable star members for 370 nearby open clusters, mainly on the Galactic disk, using the *Gaia* DR3 catalogue. We employed the HDBSCAN algorithm applied on two sets: astrometry with five features and a combination of astrometry and photometry with seven features (see Sect. 3.2). After clustering application, we cleaned up the samples implementing an outlier removal process with the Mahalanobis metric. The final stellar members are determined by merging results from 5F and 7F. Figure 4 shows the comparison between the number of stars ($N_{\rm stars}$) found in this work and those found in the reference sample by Cantat-Gaudin et al. (2020). Table 1 also shows the number of stars and parameters for some clusters found in this work (tables available at the CDS). The age, metallicity and extinction were estimated by fitting the PARSEC tracks (Bressan et al. 2012; Marigo et al. 2013) to the CMDs through the BASE-9 software.

4.1. Membership results

The star membership was performed on *Gaia* DR3 using the HDBSCAN algorithm through the `scikit-learn` Python package (Pedregosa et al. 2011). To select the open clusters and avoid hyperparameters issues that can arise due to the different size populations, we divided the whole sample into five sub-samples as explained in Sect. 2. In order to improve the HDBSCAN efficiency, Hunt & Reffert (2023) chose `min_cluster_size` $\in \{10, 20, 40, 80\}$, which are reasonable sizes for a star cluster. Following their methodology, we found that due to the wide population sizes, the same `min_cluster_size` does not work for all sub-samples. For that matter, the algorithm was unable to detect open clusters, particularly those with a reduced number of stars (S_3, S_4, S_5). Therefore, we chose a different `min_cluster_size`: for S_1, S_2, S_3, S_4 and S_5 we select 60, 50, 40, 30 and 20, respectively. Moreover, we use `min_samples` = 5 instead of 10 chose by Hunt & Reffert (2023) for all regions. This further improved the efficiency of recovering the clusters when HDBSCAN was running. Before clustering implementation, we standardized the sub-samples data through the `RobustScaler` on the `scikit-learn` Python package (Pedregosa et al. 2011) scaling features using statistics that are robust to outliers. Then, the algorithm was applied using the Mahalanobis metric given by Eq. (1) as the pairwise metric. This increased considerably the computational time due to covariance matrix calculations. We ran the entire routines using GPU cores in the supercomputer *Hypathia* at Universidad de los Andes, Colombia. All tasks lasted about eight days per sub-sample.

After the clustering has been applied, we recovered the stellar members merging results from 5F and 7F implementations. Then, we removed outliers for the whole sub-samples using once again the Mahalanobis distance given by Eq. (2). For instance,

Table 1. Cluster parameters for selected open clusters.

Cluster	$N_{\rm stars}$	α	δ	$\mu_{\alpha*}$	μ_δ	ϖ_P	log(age)	Z	$A_{\rm V}$	$m-M$
		(deg)		(mas yr^{-1})		(mas)		(dex)	(mag)	
					...					
BH 99	544	159.49	−59.15	−14.46	1.00	2.207	7.66	0.0672	0.34	8.35
Collinder 463	644	27.06	71.74	−1.74	−0.37	1.123	8.51	0.1943	0.79	9.65
IC 1396	577	324.80	57.53	−2.26	−4.58	1.040	6.98	−0.0225	1.59	9.79
Melotte 22	1130	56.61	24.10	19.92	−45.40	7.322	8.01	0.0287	0.17	5.55
NGC 1342	455	52.91	37.39	0.41	−1.66	1.485	9.05	−0.1134	0.71	9.18
					...					

Notes. Equatorial coordinates (α, δ), proper motions ($\mu_{\alpha*}$, μ_δ), and corrected parallax (ϖ_P) are computed as the median values in each cluster sample. The cluster parameters log(age), Z, $A_{\rm V}$, and $m-M$ are estimated through the BASE-9 suite software. The full version of this table is available at the CDS.

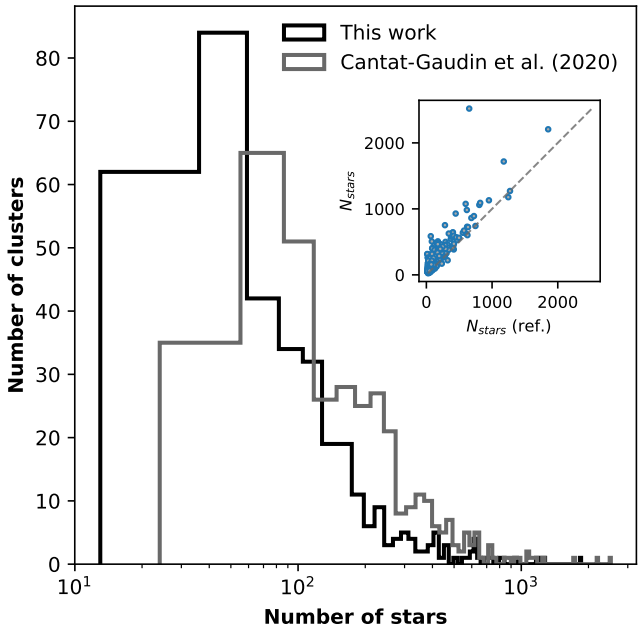

Fig. 4. Histogram of the number of stars ($N_{\rm stars}$) found in this work and those reported in Cantat-Gaudin et al. (2020). The inset shows a plot of the comparison in number of stars between this work (y-axis) and Cantat-Gaudin et al. (2020) (x-axis). The dashed line in the inset represents the identity.

Fig. 2 shows the Mahalanobis distance distribution in the vector point diagram, equatorial coordinates and the CMD for the Roslund 6 open cluster. Following the methodology described in Sect. 3.3 we set an outlier threshold of 3.33 (dashed horizontal line in Fig. 3), then, all data with $d_M(\boldsymbol{r_i})$ greater than the threshold are classified as outliers. From the Roslund 6 members found by `HDBSCAN` we obtained a clean list of 429 stars considering a 95% confidence interval, which classifies 57 stars as spurious data.

It can be seen in Fig. 2 that, as expected, outliers are located at the outskirts of the cluster and/or at the faint end of the main sequence, where more contamination occurs, about 70% of outliers found have $G > 16$. Nonetheless, on average for all clusters, about 86% of all initial `HDBSCAN`-selected members with G > 16 are kept as members after the outlier cleaning step, the lowest percentage being 66% and the largest one 100% which occurs for 11 clusters. As an example, for cluster Roslund 6 depicted in Fig. 2, only 15% of the faint stars were rejected. This indicates that the methodology applied can properly select bona fide faint cluster members, fulfilling the main goal of this investigation.

4.2. Approaching the faintest limit

The recent catalogues of open clusters using *Gaia* provide valuable information on the census of stars and properties of these Galactic objects. The efforts are focused on extending the number of stellar members and detecting new clusters based on the astrometric information provided by the catalogues. It can be seen how the number of clusters has been growing due to new data collected, from a few thousands in the last decade (Kharchenko et al. 2013) to more than seven thousand in one of the latest catalogues made by Hunt & Reffert (2023). This would not have been possible without *Gaia*, its content and its precise parallaxes have changed our view of the Milky Way.

However, no catalogue is completely exempt from systematic errors and dealing with these issues may be tedious. Many ways have been used to remove spurious data in the *Gaia* catalogues, quality cuts on statistics based on data being the most commonly implemented. For instance, Cantat-Gaudin et al. (2020) opted to choose sources up to $G = 18$ in *Gaia* DR2, which lowers the number of stars in the faintest region of the CMD. On the other hand, Rybizki et al. (2022) has trained neural networks on a diverse set of features for stars with very high signal-to-noise ratio but negative parallax (i.e. `parallax_over_error < -4.5`) in *Gaia* EDR3, which amount to a non-negligible population (Luri et al. 2018). In addition, to improve the open cluster census, Hunt & Reffert (2023) selected data by cutting on the re-normalized unit weight error (`ruwe`) and with a quality value of at least 0.5 in the statistic computed by Rybizki et al. (2022), before clustering implementation. These cuts on *Gaia* DR3 data considerably increased the number of star members in known open clusters and allowed the detection of new ones. However, at the faint end in the CMD, parallax uncertainties (σ_ϖ) increase and if not treated with caution they may distort the spatial distributions (Smith & Eichhorn 1996; Bailer-Jones 2015; Luri et al. 2018). Rybizki et al. (2022) also conclude that to obtain good astrometric samples, data can be

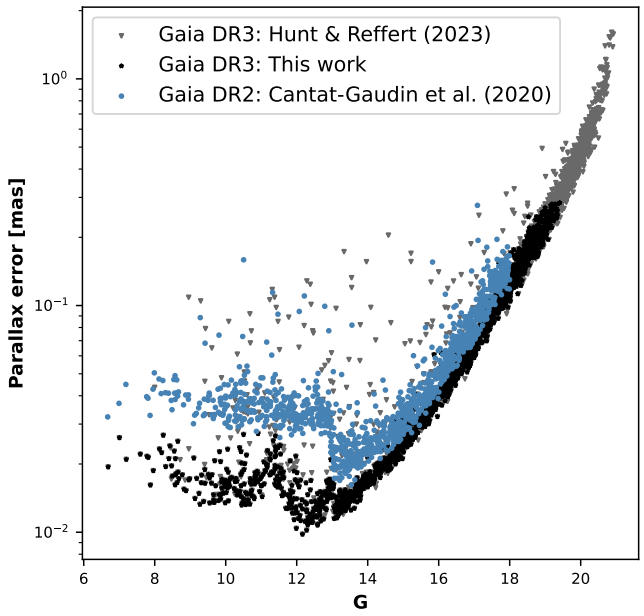

Fig. 5. G magnitude vs parallax error (σ_ϖ) for the Stock 2 open cluster. There are 1178, 1718, and 3011 stars classified by Cantat-Gaudin et al. (2020) (blue points), this work (black points), and Hunt & Reffert (2023) (grey points). The y-axis is in logarithmic scale.

filtered out by `ruwe` and `astrometric_excess_noise`, which in addition to the `parallax_over_error` may remove spurious sources as done in this study. Nevertheless, the selection function is far from trivial and distance estimates inverting the parallax for sources with high parallax uncertainties can bias the results. In conclusion, dealing with these issues is far from straightforward and further research is required.

In this work, we have aimed to reach a proper balance between extending the faint limit of cluster members while keeping contamination at reduced levels, providing valuable members samples for further meaningful studies of these clusters. Our first step to reduce spurious data was to restrict the *Gaia* DR3 data by σ_ϖ, re-normalised unit weight error and astrometric excess noise (see Sect. 2). As mentioned by Lindegren et al. (2021), there are random and systematic errors in the *Gaia* catalogue, particularly in parallaxes, sources at both bright and faint limits of *Gaia* have increasing uncertainties that in case of ϖ will produce poorer estimates of the real parallax $\varpi_{\rm true}$ (Luri et al. 2018).

Figure 5 shows the star members for the Stock 2 cluster classified by Cantat-Gaudin et al. (2020) in blue, this work in black and Hunt & Reffert (2023) in grey. Parallax improvement between *Gaia* DR2 and DR3 is most remarkable between $G \sim 6$ and $G \sim 15$ but beyond $G \sim 15$ towards the faintest limit, parallax uncertainties still increase considerably. Yet, we have been able to go deeper without introducing significant noise. For example, for the Stock 2 cluster, we found 1374 stars with $G > 18$ for which the average σ_ϖ is ∼0.16 mas, while the Hunt & Reffert (2023) sample reaches σ_ϖ up to ∼1.6 mas, an order of magnitude larger errors. Such high uncertainties may affect the real spatial distribution in the clusters, for example, at a distance of 379 pc reported for Stock 2 by Cantat-Gaudin et al. (2020), a $\sigma_\varpi \sim 0.5$ mas would significantly distort its real distance range to 318 − 414 pc, affecting estimates on tidal radius or mass segregation process. On the other hand, a ∼0.16 mas error corresponds to a difference of ±25 pc, below the largest accepted radius for open clusters (Portegies Zwart et al. 2010).

In addition to this, the spatial distributions for many open clusters with star members from the *Gaia* catalogue have line-of-sight stretching. This effect persists even after correcting for bias parallax and estimating the distance, using the Bailer-Jones (2015) method, for example. This could lead to overestimation of tail-like structures that may be an effect produced by the dynamic evolution of the clusters or merely the product of an optical effect in the parallax measurement. This problem deserves further investigation for clusters with this effect in their spatial distributions.

4.3. Inferring the cluster parameters

The age and metallicity are the cornerstone of stellar evolution, they allow us to know about the chemical composition of the stars in the Milky Way. These parameters provide insights into the evolutionary state of stellar systems and also supply an overview of the matter distribution across the Galaxy, which is crucial to understand processes such as mass segregation, radial migration, and dynamical heating (Mackereth et al. 2019). Estimating such parameters is a non-trivial task, and the common ways to compute them is, for example, by lithium depletion boundary and the classical isochrone fitting (Dinnbier et al. 2022). Furthermore, artificial neural networks have gained popularity and have been successfully implemented in *Gaia* data to estimate those parameters bringing us another way to tackle this problem (Kounkel & Covey 2019; Cantat-Gaudin et al. 2020).

To infer age, metallicity, extinction, and distance modulus for each cluster, we opted to fit the PARSEC isochrones (Bressan et al. 2012; Marigo et al. 2013) to the CMDs through the Bayesian Analysis for Stellar Evolution with nine variables BASE-9 software (von Hippel et al. 2006; Robinson et al. 2016). This suite software obtains the best cluster parameters by robust statistical principles that compute random walks to sample the marginal posterior probability function. BASE-9 requires photometry (G, G_{BP}, G_{RP}) and their uncertainties as inputs to start sampling. Since *Gaia* does not provide photometry errors we used parallax error σ_ϖ as a proxy, as done by Kounkel & Covey (2019). In other words, these errors are a measure of relative uncertainties of the data used. The binaries option in BASE-9 is turned off, therefore all stars are treated as single. We chose the initial guess for age, extinction, and distance modulus from Cantat-Gaudin et al. (2020). For metallicity we used the values reported in Netopil et al. (2016), Dias et al. (2021) and Fu et al. (2022) for a total of 338 open clusters, the remaining ones have an initial value of $Z = 0.01$ dex, which is a reasonable value for open clusters in the solar neighbourhood (Bossini et al. 2019). BASE-9 was executed on blocks of 40 clusters in parallel with a total of one hundred thousand iterations. After sampling, we obtained a flat chain with a length of ten thousand iterations, thus the estimated cluster parameters are the 50th percentiles or the second quartiles Q_2. Figure 6 shows the corner plot of the marginal posterior probability function sampling for the BH 99 cluster with the upper and lower uncertainties in each parameter as the 16th and 84th percentiles. Some isochrones for Alessi 9, NGC 2516, Roslund 6 and UPK 640 open clusters with parameters estimated through BASE-9 are depicted in Fig. 7.

The left panel of Fig. 8 presents the age distribution of the studied open clusters. The log(age) covers a range from 6.97 with Collinder 69 as the youngest, to 9.56 with UBC 21 as the oldest cluster in the sample. Among the 370 clusters, 250 open clusters were found between the 16th and 84th percentiles with a median value of log(age)= 8.03, similar to the median of log(age)= 8.2 in Bossini et al. (2019). The right panel of Fig. 8

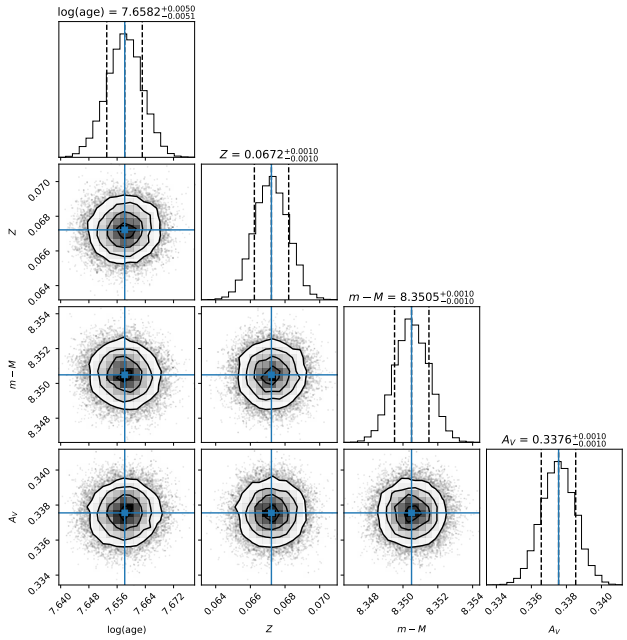

Fig. 6. Corner plot showing the marginal posterior probability distribution for the BH 99 open cluster computed using BASE-9 with *Gaia* photometry. The blue solid vertical line indicates the estimated parameter, which corresponds to the 50th percentile.

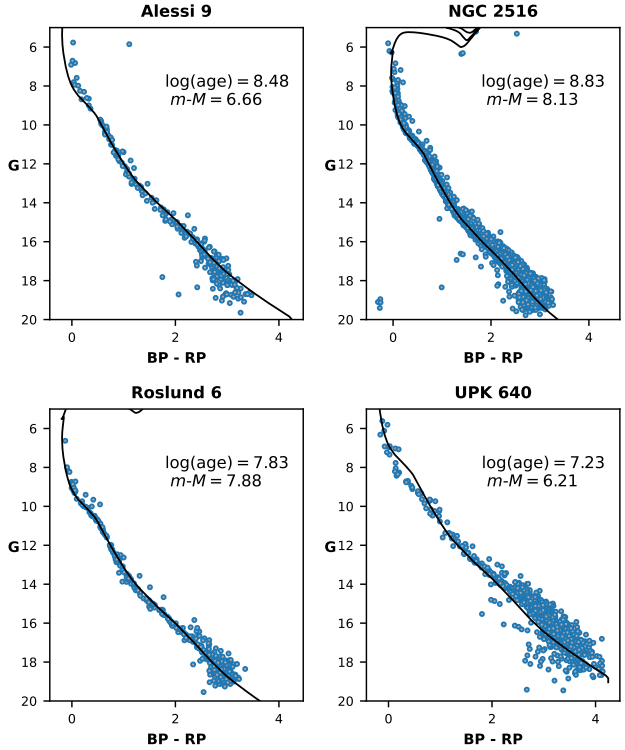

Fig. 7. CMDs for Alessi 9, NGC 2516, Roslund 6, and UPK 640 open clusters. The solid black curve is the PARSEC isochrone with parameters estimated through BASE-9.

shows the extinction versus distance modulus, which shows that most of the clusters have low extinction with values less than 1.0 mag. The UPK 201 cluster presents the highest extinction (A_V = 3.39) in our sample. Figure 9 shows the comparison for age (*left top panel*), distance modulus (*left bottom panel*) and extinction (*right bottom panel*) estimated in this work against Cantat-Gaudin et al. (2020) (for all clusters), and for metallicity (*right top panel*) against Dias et al. (2021) (for 324 open clusters in common). Cantat-Gaudin et al. (2020) values were obtained by an artificial neural network trained using the PARSEC isochrones. Dias et al. (2021) metallicities were obtained using the *Gaia* DR2 photometry and an isochrone fitting code also with the PARSEC tracks.

As explained before, Cantat-Gaudin et al. (2020) opted to perform a cutoff in $G \sim 18$ in order to avoid issues previously stated (see Sect. 4.2) in the fainter region of the CMDs, while we include stellar members with magnitudes up to $G \sim 19.8$ in most of the clusters using quality cuts discussed in Sect. 2. We found slight differences in ages compared to Cantat-Gaudin et al. (2020) that may have different causes, such as different methods to estimate age and our work including fainter members. Nevertheless, the discrepancies are limited to a few values and only 18 clusters have $|\Delta\log(\text{age})| > 0.5$. Particularly, the three open clusters with the highest discrepancies in $\Delta\log(\text{age})$ are UBC 21, UPK 18 and UPK 542. Nonetheless, these clusters agree in the other parameters, suggesting that age estimates are sensitive to large scatter in the CMD. In addition, most of the clusters have similarities in extinction with Cantat-Gaudin et al. (2020) and only 10 of them have $|\Delta A_V| > 0.005$. In case of distance modulus, the discrepancies are small and only 14 open clusters have $|\Delta(m-M)| > 0.05$, the vast majority of them with distance estimates 10% or less closer than those of Cantat-Gaudin et al. (2020).

We noted that only 27 among the 324 clusters have $|\Delta Z| > 0.002$ dex with a median value of 0.0001 dex, which indicates similarities in the metallicities derived in this work with Dias et al. (2021). Further research is required to quantify the impact of the inclusion of faint and low-mass stars on age and metallicity estimates through the standard isochrone fitting process. This will be discussed in the second series of this paper (Alfonso et al. in prep.).

5. Discussion

We have updated the cluster parameters with the *Gaia* DR3 catalogue for 370 nearby open clusters for which we obtained reliable stellar members using the `HDBSCAN` clustering algorithm with the Mahalanobis distance as the pairwise metric. Among the cluster sample we covered a distance range from ∼ 104 pc with Alessi 13 depicted in the bottom panel of Fig. 10 as the closest, to ∼998 pc with NGC 7762 as the farthest. We show in Fig. 1 that these clusters are mainly in the Galactic plane, which in fact, the colours for the five sub-samples highlights the wide variety of cluster size populations, as mentioned in Sect. 2. We also cover metallicities ranging from −0.23 dex to 0.42 dex with UBC 31 as the most metal-rich open cluster. Moreover, the ages calculated in this work are in the range of 6.98–9.56 with Collinder 69 and UBC 21 as the youngest and oldest open clusters, respectively. The latter with the highest age increment of +1.08 log(age) compared to Cantat-Gaudin et al. (2020).

The cluster parameters, shown in Table 1, were estimated with the PARSEC isochrones through the BASE-9 implementation. Figure 10 shows the obtained metallicity, age, and extinction of the open clusters, colour-coded in a sky map. A total of 53 890 of our 87 708 stellar members are in the catalogue provided by Cantat-Gaudin et al. (2020), representing an increment of about ∼40%. It is worth mentioning that there are 1738 repeated stars in the full sample listed as members of two or three clusters. This may be caused by very close binary open clusters

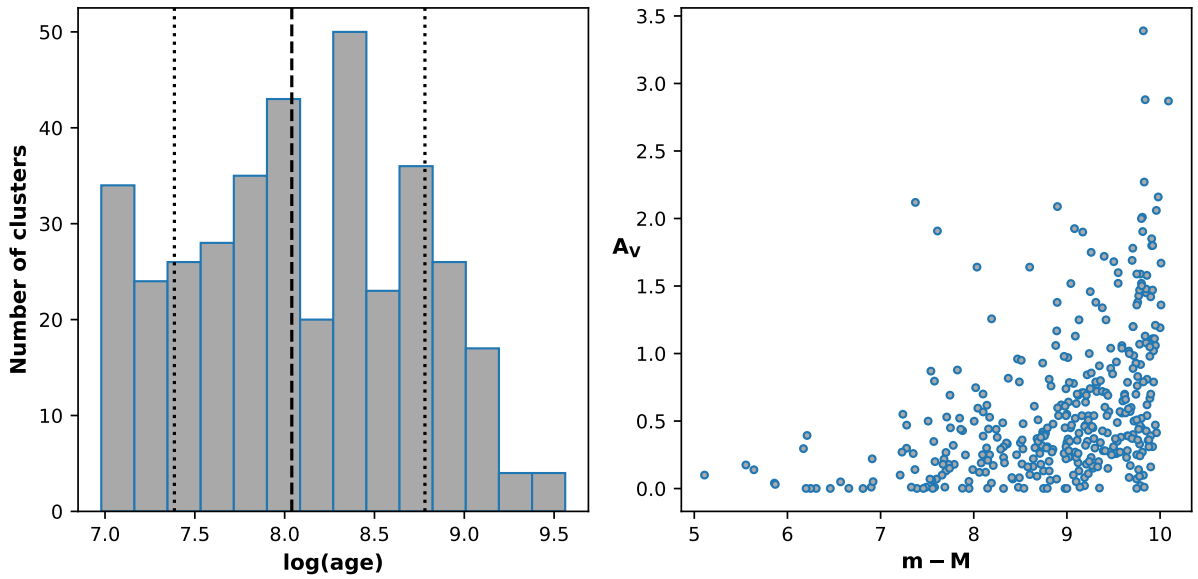

Fig. 8. Cluster parameter results. Left: age distribution of the studied open clusters. The dashed vertical line represents the median value (log(age) = 8.03), while the dotted lines are the 16th (log(age) = 7.38) and 84th (log(age) = 8.78) percentiles, respectively. Right: extinction vs distance modulus estimated in this work.

that the algorithm overlap as one and requires further analysis case by case to reach a conclusion.

On the other hand, the current census in Hunt & Reffert (2023) has about seven thousand clusters with almost one thousand five hundred within 1 kpc. We opted to focus on open clusters not including comoving groups and streams as Theias (Kounkel & Covey 2019), their wide variety of stellar content not sharing a common origin (Zucker et al. 2022) may be challenging for cluster parameter estimation. A total of 56 and 46 open clusters have now parameters not reported previously by Hunt & Reffert (2023) and Dias et al. (2021), respectively.

In our sample, we found four pairs of truly binary open clusters reported by de La Fuente Marcos & de La Fuente Marcos (2009) and Song et al. (2022). These pairs have distances about 30 pc between the components and also similar ages and metallicities. They are: Collinder 135 - UBC 7, Alessi 43 - Collinder 197, NGC 6716 - Collinder 394 and ASCC 16 - ASCC 21. In addition, we also found six system pairs with distances less than about 20 pc: ASCC 123 - Stock 12, BDSB91 - vdBergh 80, Collinder 394 - NGC 6716, Gulliver 6 - UBC 17b, RSG 7 - RSG 8 and UBC 392 - UPK 194. These objects deserve further research to confirm or discard a common origin or physical association (de La Fuente Marcos & de La Fuente Marcos 2009).

We noted that including photometry to `HDBSCAN` allows to obtain faint stars that were not detected using only astrometry, with parallax uncertainties up to $\sigma_{\varpi} = 0.16$ mas for the open clusters studied in this work. These faint stars are crucial to understand dynamical properties of these clusters, such as the mass segregation effect, which may be due to dynamical evolution from a non-mass-segregated cluster or primordial as a product of star formation process (Allison et al. 2009). Nevertheless, including faint stars with the *Gaia* data is a cautious task due to the astrometric uncertainties: the fainter in the CMD, the larger the σ_{ϖ} because of the photon limit in the *Gaia* CCD detector (see Fig. 5). At the faint regime, the main sequence has high dispersion, which can make it difficult to estimate cluster parameters. Further investigation is require to quantify how well or poorly age and metallicity are estimated through traditional isochrone fitting including these stars beyond $G \geq 20$. In addition, using quality cuts such as those adopted in this work (see Sect. 2) can deal with this problem (Lindegren et al. 2021; Rybizki et al. 2022), which reduces the number of spurious data that may affect cluster parameter estimates and clustering results.

This study is mainly focused on obtaining star members for open clusters near the Galactic plane within 1 kpc with the novel *Gaia* DR3 catalogue using the unsupervised machine learning algorithm `HDBSCAN` applied on astrometry and photometry. The algorithm was applied taken into account the correlations among the different kind of informations provided in the catalogue through the Mahalanobis metric and also used the latter to remove outliers from the cluster samples. We found stars up to $G \sim 19.8$ that were not included in the reference Cantat-Gaudin et al. (2020) sample with *Gaia* DR2. The exquisite information in *Gaia* DR3 allows us to compute basic kinematic parameters and also update age, extinction, distance modulus, and determine metallicity which was not reported in the reference sample. Additionally, spectroscopic surveys such as LAMOST (Cui et al. 2012) and WEAVE (Dalton et al. 2012) with *Gaia* will provide an extended and full characterization of the dynamical and stellar process in these open clusters, which for the lack of radial velocities in *Gaia* will bring full details about the kinematic of these clusters in their velocity spaces.

6. Summary and conclusions

In this work, we obtained reliable star members for 370 nearby open clusters using the novel *Gaia* DR3 and the `HDBSCAN` algorithm applied in two approaches. First, we performed the algorithm on the common astrometric space `5F` with equatorial coordinates, proper motions, and the parallax corrected by the zero-point bias. Then, we applied once again `HDBSCAN` by adding two colour indexes `7F`. We note that by including photometry to clustering, the number of faint stars with $G > 17$ increases by about 10% compared to clustering based solely on astrometry, which allows us to obtain additional stars in the faintest region of the CMD. The Mahalanobis metric was used as the pairwise

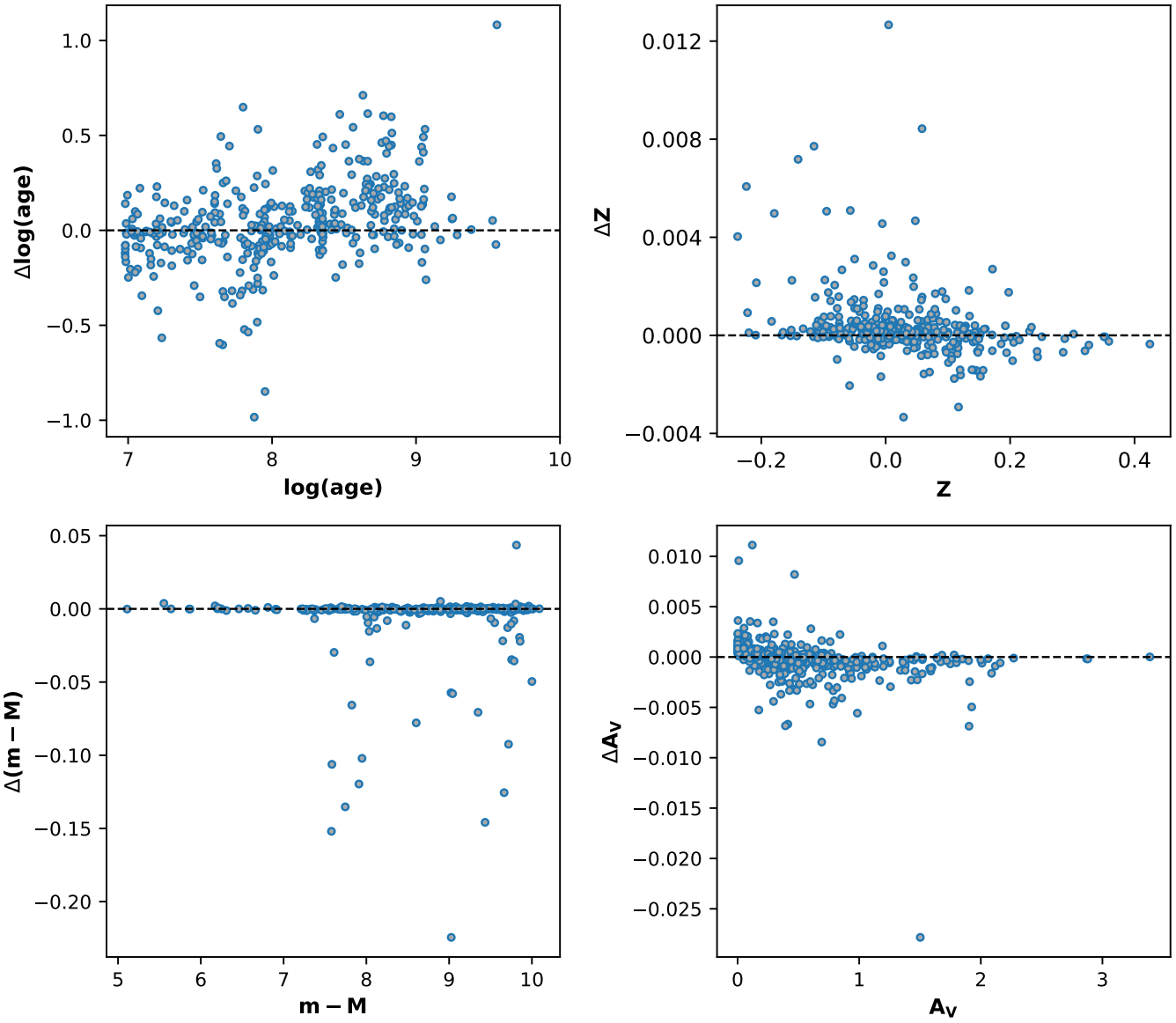

Fig. 9. Comparison of cluster parameters derived in this work. The log(age) (top left), distance modulus (bottom left) and extinction (bottom right) are compared against Cantat-Gaudin et al. (2020) for 370 clusters. The metallicity (top right) is compared against Dias et al. (2021) for 324 clusters in common: $\Delta\log(\text{age}) = \log(\text{age})_{\text{this work}} - \log(\text{age})_{\text{literature}}$ vs. $\log(\text{age})_{\text{this work}}$. $\Delta Z = Z_{\text{this work}} - Z_{\text{literature}}$ vs. $Z_{\text{this work}}$. $\Delta(m-M) = (m-M)_{\text{this work}} - (m-M)_{\text{literature}}$ vs. $(m-M)_{\text{this work}}$. $\Delta A_V = A_{V\,\text{this work}} - A_{V\,\text{literature}}$ vs. $A_{V\,\text{this work}}$.

metric to compute distances for clustering to take into account any level of correlations between astrometric and photometric data. This metric was also used to remove outliers from the membership lists; we computed the covariance matrix through the minimum covariance determinant estimator to avoid influences of extreme values. Then, based on the χ^2 distribution and considering a 95% of confidence interval with five degrees of freedom for the astrometric space 5F, we removed spurious data from the membership lists in each cluster.

We used the *Gaia* DR3 photometry with the stellar members classified by the clustering algorithm to estimate the age, metallicity, extinction, and distance modulus fitting the PARSEC isochrones through the BASE-9 suite software. We found that to establish the metallicity distribution in open clusters, it is required to extend the membership list to clusters beyond the solar neighbourhood. This provides details about the structure of the Galaxy and also the chemical distribution in the thin and thick disk. In this work we showed the need to carefully select stars in open clusters, which, due to astrometric and photometric uncertainties in *Gaia*, may affect clustering results, produce an overestimation of the cluster parameters because of scattering at faint magnitudes, and also identify real or fake stretching in the line of sight of the clusters. These issues will be discussed in the second series of this paper (Alfonso et al., in prep.).

The catalogue presented in this paper exploits the *Gaia* DR3 data to extend stellar members in nearby clusters without including stars with high parallax uncertainties. We found a total of 87 708 stars for the 370 open clusters studied in this work. This first paper also provides a new technique to tackle the membership problem in star clusters including astrometry and photometry data. Our approach relies on dealing with correlations in the data and uses a robust statistical estimator to avoid any level of affectation in the final membership lists.

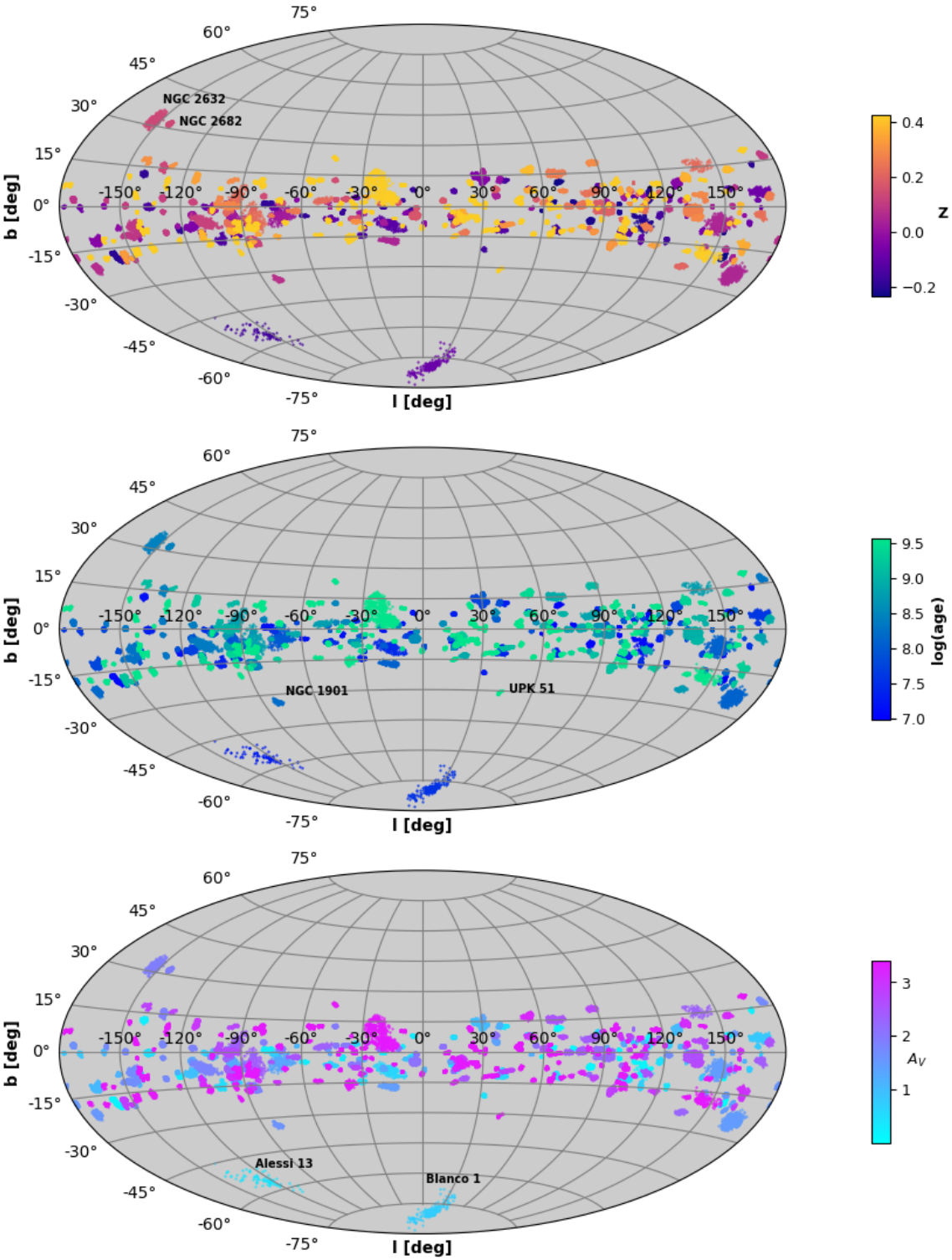

Fig. 10. Celestial maps in Galactic coordinates for the 370 open clusters with a total of 87 708 stars. The colour bar indicates the metallicity (top), log(age) (middle), and extinction (bottom) distribution estimated in this work. The open clusters NGC 2632, NGC 2682, UPK 51, NGC 1901, Alessi 13, and Blanco 1 have the highest and lowest Galactic latitude.

Data availability

The Full Table 1 is available at the CDS via anonymous ftp to cdsarc.cds.unistra.fr (130.79.128.5) or via https://cdsarc.cds.unistra.fr/viz-bin/cat/J/A+A/689/A18

Acknowledgements. The authors thank the anonymous referee for his/her/their very valuable and constructive comments and suggestions that improve the quality of this manuscript. The authors would like to thank the Vice Presidency of Research & Creation's Publication Fund at Universidad de los Andes for its financial support and also the Fondo de Investigaciones de la Facultad de Ciencias de la Universidad de los Andes, Colombia, through its Programa de Investigación

código INV-2023-162-2853. Jeison Alfonso acknowledges doctoral fellowship support from the Departamento de Física de la Universidad de los Andes, Colombia, in form of *Asistencia Graduada Doctoral Docencia*. This work has made use of data from the European Space Agency (ESA) mission *Gaia* (https://www.cosmos.esa.int/gaia), processed by DPAC, (https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement.

References

Allison, R. J., Goodwin, S. P., Parker, R. J., et al. 2009, MNRAS, 395, 1449
Anders, F., Chiappini, C., Minchev, I., et al. 2017, A&A, 600, A70
Babusiaux, C., Fabricius, C., Khanna, S., et al. 2023, A&A, 674, A32
Bailer-Jones, C. A. L. 2015, PASP, 127, 994
Binney, J., & Tremaine, S. 2008, Galactic Dynamics, 2nd edn. (Princeton: Princeton University Press)
Bossini, D., Vallenari, A., Bragaglia, A., et al. 2019, A&A, 623, A108
Bressan, A., Marigo, P., Girardi, L., et al. 2012, MNRAS, 427, 127
Campello, R. J. G. B., Moulavi, D., & Sander, J. 2013, in Advances in Knowledge Discovery and Data Mining, eds. J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (Berlin, Heidelberg: Springer Berlin Heidelberg), 160
Cantat-Gaudin, T., Anders, F., Castro-Ginard, A., et al. 2020, A&A, 640, A1
Castro-Ginard, A., Jordi, C., Luri, X., et al. 2018, A&A, 618, A59
Castro-Ginard, A., Jordi, C., Luri, X., et al. 2022, A&A, 661, A118
Cui, X.-Q., Zhao, Y.-H., Chu, Y.-Q., et al. 2012, Res. Astron. Astrophys. 12, 1197
Dalton, G., Trager, S. C., Abrams, D. C., et al. 2012, SPIE Conf. Ser., 8446, 84460P
Della Croce, A., Dalessandro, E., Livernois, A. R., & Vesperini, E. 2024, A&A, 683, A10
de La Fuente Marcos, R., & de La Fuente Marcos, C. 2009, A&A, 500, L13
Dias, W. S., Monteiro, H., Moitinho, A., et al. 2021, MNRAS, 504, 356
Dinnbier, F., Kroupa, P., Šubr, L., & Jeřábková, T. 2022, ApJ, 925, 214
Ester, M., Kriegel, H.-P., Sander, J., et al. 1996, in Second International Conference on Knowledge Discovery and Data Mining (KDD'96), eds. E. Simoudis, J. Han & U. Fayyad (AAAI Press), 226
Feigelson, E. D., & Babu, G. J. 2012, Modern Statistical Methods for Astronomy (Cambridge: Cambridge University Press)
Fu, X., Bragaglia, A., Liu, C., et al. 2022, A&A, 668, A4
Gaia Collaboration (Prusti, T., et al.) 2016, A&A, 595, A1
Gaia Collaboration (Vallenari, A., et al.) 2023, A&A, 674, A1
Hunt, E. L., & Reffert, S. 2021, A&A, 646, A104
Hunt, E. L., & Reffert, S. 2023, A&A, 673, A114
Kharchenko, N. V., Piskunov, A. E., Schilbach, E., Röser, S., & Scholz, R. D. 2013, A&A, 558, A53
Kounkel, M., & Covey, K. 2019, AJ, 158, 122
Krumholz, M. R., McKee, C. F., & Bland-Hawthorn, J. 2019, ARA&A, 57, 227
Küpper, A. H. W., MacLeod, A., & Heggie, D. C. 2008, MNRAS, 387, 1248
Küpper, A. H. W., Balbinot, E., Bonaca, A., et al. 2015, ApJ, 803, 80
Lada, C. J., & Lada, E. A. 2003, ARA&A, 41, 57
Lindegren, L., Bastian, U., Biermann, M., et al. 2021, A&A, 649, A4
Luri, X., Brown, A. G. A., Sarro, L. M., et al. 2018, A&A, 616, A9
Mackereth, J. T., Bovy, J., Leung, H. W., et al. 2019, MNRAS, 489, 176
Mahalanobis, P. C. 2018, Sankhyā: Indian J. Statis. Ser. A, 80, S1
Marigo, P., Bressan, A., Nanni, A., Girardi, L., & Pumo, M. L. 2013, MNRAS, 434, 488
McInnes, L., Healy, J., & Astels, S. 2017, J. Open Source Softw., 2, 205
McKee, C. F., & Ostriker, E. C. 2007, ARA&A, 45, 565
Netopil, M., Paunzen, E., Heiter, U., & Soubiran, C. 2016, A&A, 585, A150
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, J. Mach. Learn. Res., 12, 2825
Portegies Zwart, S. F., McMillan, S. L. W., & Gieles, M. 2010, ARA&A, 48, 431
Robinson, E., von Hippel, T., Stein, N., et al. 2016, Astrophysics Source Code Library [record ascl:1608.007]
Rousseeuw, P. J., & van Driessen, K. 1999, Technometrics, 41, 212
Rybizki, J., Green, G. M., Rix, H.-W., et al. 2022, MNRAS, 510, 2597
Smith, H., J., & Eichhorn, H. 1996, MNRAS, 281, 211
Song, F., Esamdin, A., Hu, Q., & Zhang, M. 2022, A&A, 666, A75
Spina, L., Randich, S., Magrini, L., et al. 2017, A&A, 601, A70
van Groeningen, M. G. J., Castro-Ginard, A., Brown, A. G. A., Casamiquela, L., & Jordi, C. 2023, A&A, 675, A68
von Hippel, T., Jefferys, W. H., Scott, J., et al. 2006, ApJ, 645, 1436
Yong, D., Carney, B. W., & Friel, E. D. 2012, AJ, 144, 95
Zhong, J., Chen, L., Kouwenhoven, M. B. N., et al. 2019, A&A, 624, A34
Zucker, C., Peek, J. E. G., & Loebman, S. 2022, ApJ, 936, 160