Membership determination in open clusters using DBSCAN Clustering Algorithm

Mudasir Raja[1]**, Md Mahmudunnobe**[2]**, Priya Hasan**[1] **and S N Hasan**[1]

1. Maulana Azad National Urdu University Hyderabad,500032 2. Minerva University, California, USA

Correspondence: mudasirraja@gmail.com

Abstract

In this work, we aim to study membership of four open clusters (NGC 1893, NGC 581, NGC 2264 and NGC 2354) using the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering algorithm on Gaia EDR3 Data. We select stars from the Gaia EDR3 catalog, construct a five-dimensional phase space (three-dimensional spatial position and two-dimensional proper motion) and obtain reliable cluster members using machine learning (DBSCAN). We obtain cluster parameters for our sample and compare it with the catalog values. The technique demonstrates the effectiveness of machine learning in membership determination of clusters.

Introduction

Star clusters are the building blocks of galaxies and are the key to understanding the formation and evolution of stars and galaxies([1], [2]).They are an over-density of sample of stars in the same region of the sky formed from the same molecular cloud and hence all member stars are approximately at the same distance, of the same age and only differ in mass. Cluster members move with a common velocity, which is an imprint of their formation process. Hence, to identify members, we look for stars in the same region of the sky, at the same distance and with a common velocity. However, there will always be contamination from field stars with similar velocities and at similar distances in the same region of the sky.

This animation

In order to detect the members of open clusters, we need homogeneous, precise, astrometric, photometric and kinematic data of stars and a highly efficient member determination method. Earlier, because of the lack of this kind of stellar data, many scholars determined membership for a particular region and proposed various methods like the Vasilevskis-Sanders (VS) method. However, these methods have certain limitations, for example, when the number of member stars is far less than the number of field stars, the effectiveness may not be good. Also, the VS method is not suitable for large datasets with complex distributions.

Machine Learning (ML) provides alternative methods of membership determination. Random Forest (RF) is a supervised classification method and was applied to the Gaia DR2 data in [3, 4]. [5] used the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering algorithm for the determination of members for two open star clusters that lie within 100 pc. In this paper, we want to check the effectiveness of DBSCAN for more distant clusters. Thus we run DBSCAN for a sample of four much distant open clusters to increase our membership.

1 Data and Sample Selection

We selected a sample of four young clusters *viz.* NGC 581, NGC 1893, NGC 2264, NGC 2354 with a spread in age and distance. We also added additional criteria to remove noisy data, where we only include stars whose $parallax\_over\_error > 3$, $pmra\_error < 0.3$ and $pmdec\_error < 0.3$. In addition, in EDR3 data, we have used $ruwe\_filter < 1.4$ and have taken radius equal to 20 arcmin for all these clusters.

Table 1: Basic cluster parameters for our sample [6]

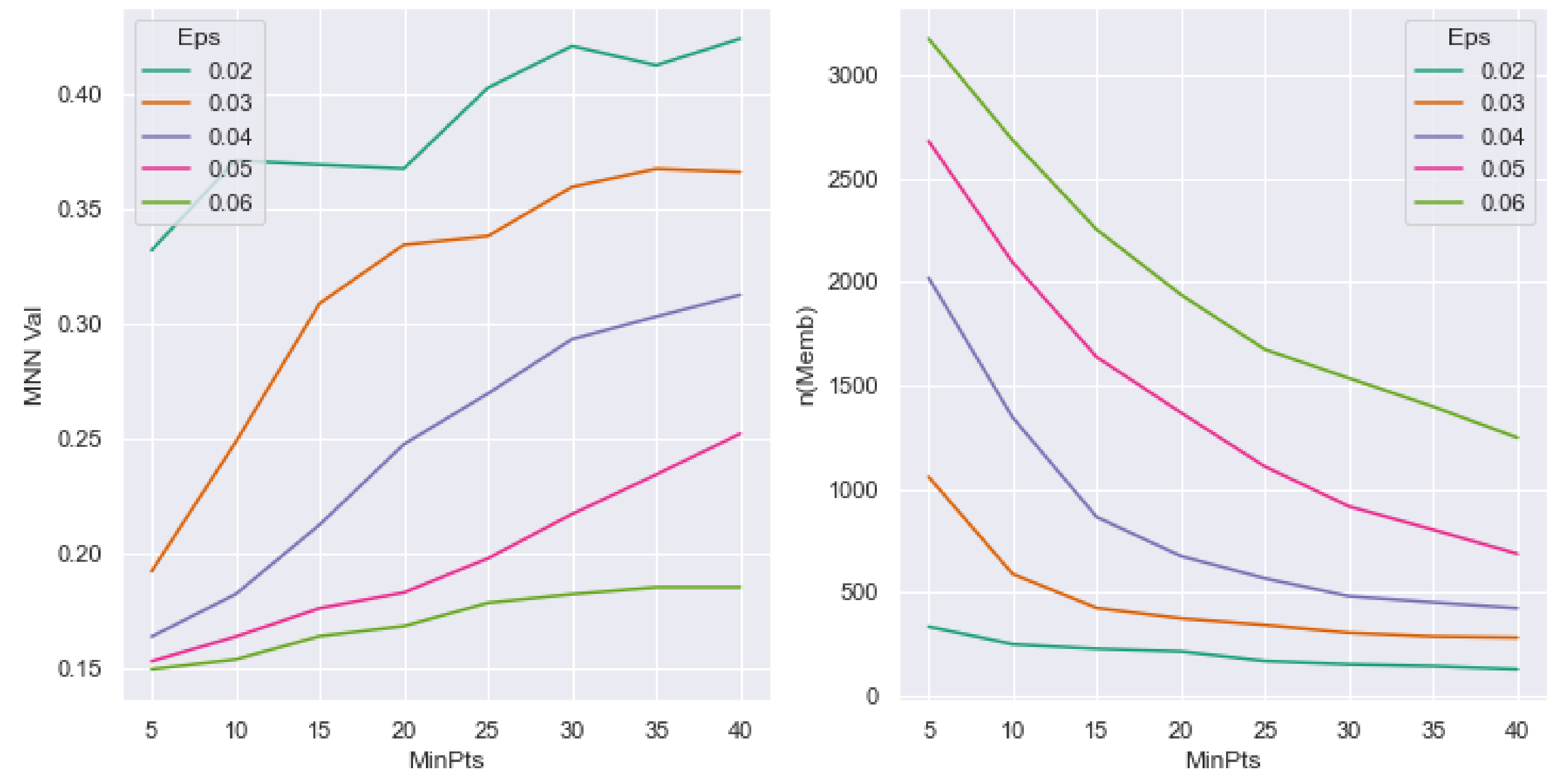
Cluster	l	b	Ang.Dia	Distance	$E(B-V)$	log t_{age}	R_{GC}
	deg	deg	arc min	pc	mag	log(yr)	kpc
NGC 581	128.05	-01.80	5.0	2194	0.38	7.3	10.0
NGC 1893	173.59	-1.68	25	6000	0.45	6.5	14.5
NGC 2264	202.94	+02.2	39.0	667 1	0.05	6.9	9.1
NGC 2354	238.368	-6.792		4085	0.37	8.126	

2 The DBSCAN Method

DBSCAN is an unsupervised density based clustering technique that can discover clusters of non-spherical shapes and hence is highly suitable for open clusters. The DBSCAN algorithm needs an averaged density of stars in the region; in smaller regions this average is more representative than if we take the whole sky, where the density can significantly vary from one region to another. Before applying DBSCAN, we first normalized the data to ensure that the difference in range of the variables does not influence the analysis. We chose a search radius of 20 arcmin around the cluster center. After running DBSCAN, we get one or more groups of stars along with many noise samples. Noise samples are the points which does not have enough points around them. We chose the DBSCAN group with the larger number of stars as the member group for the cluster.

Epsilon ϵ and MinPoints ($Minpts$) are the two input hyperparameters of the DBSCAN algorithm and whose selection directly affects the clustering effectiveness. As suggested by [5], we used the elbow method using k-dist graphs to find the optimal range of values for ϵ and $Minpts$. Once we identify an optimal range of ϵ and $Minpts$, we did a more detailed hyperparameter search using our chosen metric *mean nearest neighbor (MNN) distance*. *MNN distance* is defined as the mean distance of the nearest neighbor of each star in the group. As the cluster are compact and members close to each other in the parameter space, their *MNN distance* would be small for clusters. We finalized the value of ϵ and $Minpts$, where we get the lowest value for *MNN distance*. We also determined the number of retrieved members for different pairs of ϵ and $Minpts$. When the *MNN distance* for two or more pairs are very close to each other, we chose the one with highest number of members as shown in Fig. 1. For all four clusters, we found the value of ϵ to be 0.06 and that of $Minpts$ to be 25 as the most optimal value.

Figure 1: Parameters (ϵ and $Minpts$) optimization using the *MNN distance* and no of members.



3 Results

Using the DBSCAN method, the results we have obtained are compared with the results obtained by [4] and the variation in distance and number of member stars have been calculated as depicted in the table. Table 2 shows the results of DBSCAN for our sample.

Table 2: Results from DBSCAN for our sample and Comparison with the results from RF

Cluster	RF Members	DBSCAN Members	Ratio of DBSCAN to RF	Parallax	Distance
					pc
NGC 581	815	1674	2.05	0.30 ± 0.08	3333
NGC 1893	992	1144	1.15	0.37±0.10	2702
NGC 2264	693	543	0.78	1.38±0.09	724
NGC 2354	747	244	0.32	0.77±0.03	1298

Figure 2: Parallax distribution of the four clusters

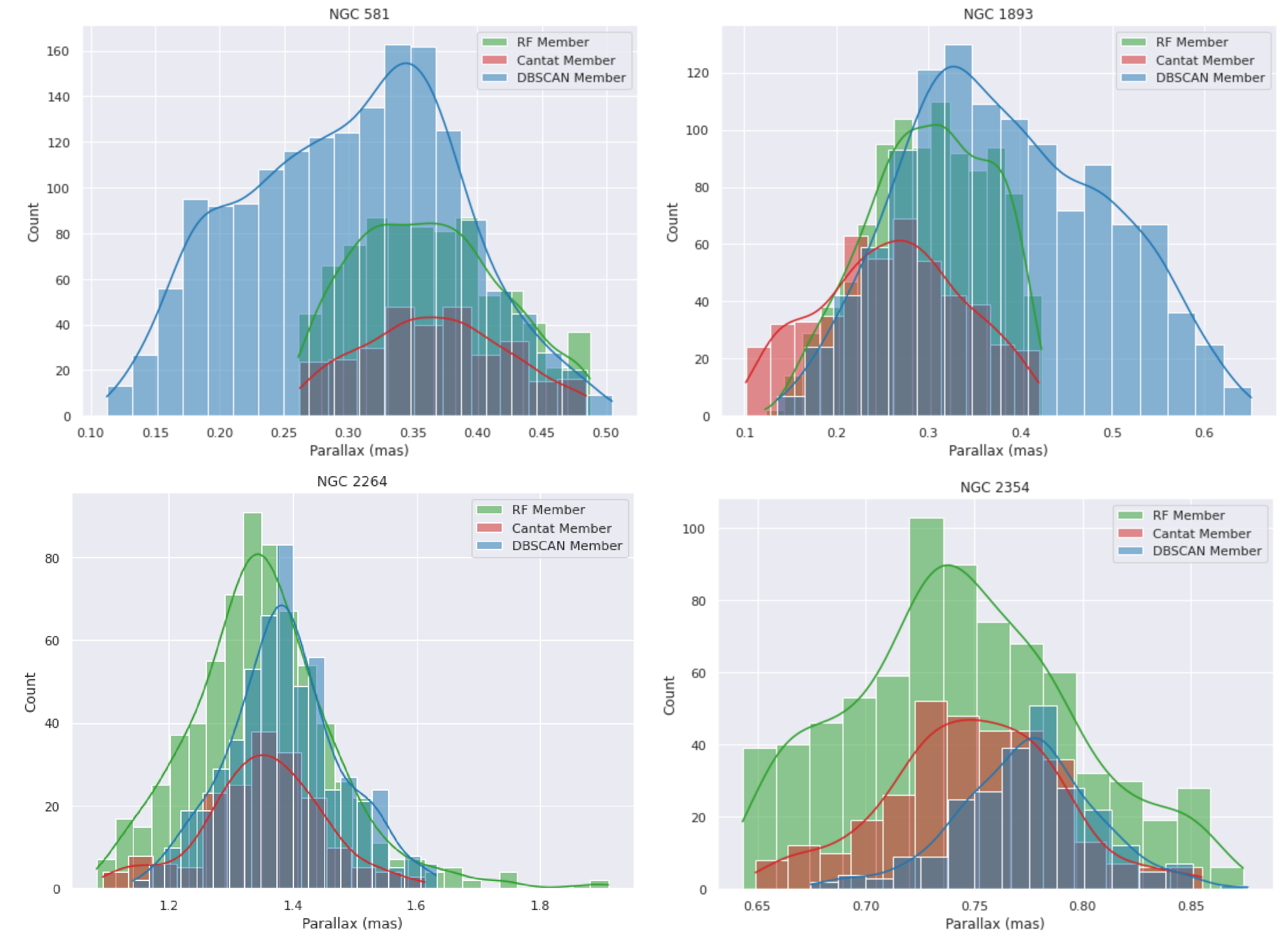
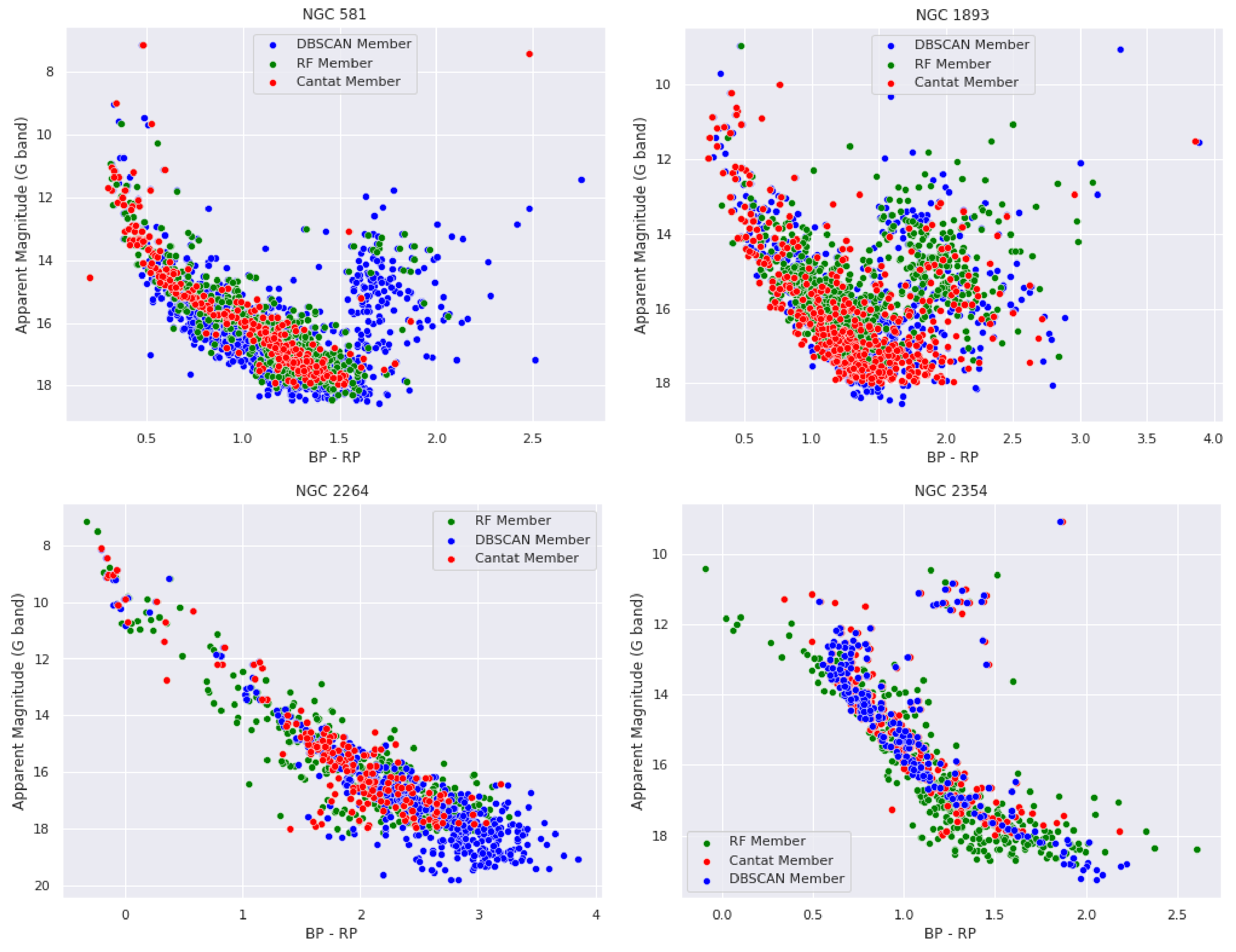


Figure 3: Color-Magnitude Diagram of these four Clusters clearly showing the non-main sequence members in the four clusters



Conclusion

We have used DBSCAN method to find the membership of four open clusters with Gaia EDR3 data.
Our Results are as followed:

- The sample of stars in clusters can be increased by a large factor, almost 1–2 times. This improves our accuracy in determining various parameters of a star cluster ranging from distance, extinction and mass function. The sizes of the studied clusters also increased with the increase in membership and we can study the outer regions of clusters.
- As we have not used photometric data while estimating membership, we can identify variables, premain sequence stars (NGC 1893), as well as all other possible non main-sequence members of the cluster.

This work indicates that the DBSCAN method has good potential applications in star cluster studies. We plan to use this data to make a detailed study of membership of these clusters. A complete paper will be published elsewhere.

References

[1] K. A. Janes and R. L. Phelps. The galactic system of old star clusters: The development of the galactic disk. , 108:1773–1785, November 1994.

[2] E. D. Friel. The Old Open Clusters Of The Milky Way. , 33:381–414, 1995.

[3] Xinhua Gao. A Machine-learning-based Investigation of the Open Cluster M67. , 869(1):9, December 2018.

[4] Md Mahmudunnobe, Priya Hasan, Mudasir Raja, and S. N. Hasan. Membership of stars in open clusters using random forest with gaia data. *The European Physical Journal Special Topics*, 230:2177–2191, 2021.

[5] Xu Shou-kun, Wang Chao, Zhuang Li-hua, and Gao Xin-hua. Dbscan clustering algorithm for the detection of nearby open clusters based on gaia-dr2two. *Chinese Astronomy and Astrophysics*, 43:225–236, 04 2019.

[6] T. Cantat-Gaudin, C. Jordi, A. Vallenari, A. Bragaglia, L. Balaguer-Núñez, C. Soubiran, D. Bossini, A. Moitinho, A. Castro-Ginard, A. Krone-Martins, L. Casamiquela, R. Sordo, and R. Carrera. A Gaia DR2 view of the open cluster population in the Milky Way. , 618:A93, Oct 2018.