Machine Learning Methods used to Distinguish Open Clusters from Asterisms using Gaia DR3 data

G. Kapanadze*, T. Kvernadze, O. Kvaratskhelia, G. Kurkhuli, V. Kozlov

Georgian National Astrophysical Observatory, Abastumani, Georgia

Abstract

We present the use of the `scikit-learn` DBSCAN clustering code as a machine learning tool to test the membership and integrity of apparent open clusters to distinguish asterisms from real open clusters in the Gaia DR3 3D data space. For testing means, we studied known Open Clusters NGC 1798 and NGC 6633. In the field of NGC 1798 we accidentally confirmed an Open Cluster LP17. For final analyze, we processed the 11 open clusters of Dolidze-Jimsheleishvili as the most of them are small having low spatial density and are hard to confirm as an Open Clusters with other methods. As a result, we report that 3 of them show clustering tendency as the open clusters: DolidzeDzim 6, DolidzeDzim 7 and DolidzeDzim 10.

Keywords: *open clusters: machine learning methods: clustering algorithms*

1. Introduction

Astronomy generates a large amount of data, thanks to modern-day terrestrial and space telescopes and superior imaging technologies (Ball & Brunner, 2010; Pesenson et al., 2010). It is a result of the past and current long duration and multi-wavelength observational surveys: the Sloan Digital Sky Survey or SDSS (York et al., 2000), which provided with multi-color images of about 1/3 of sky and spectra of millions of Galactic and extra-galactic objects; Pan-STARRS (Kaiser et al., 2010), the Zwicky Transient Facility (Bellm, 2014), CoRot (Auvergne et al., 2009), ASASS (Pojmanski et al., 2005), SuperWASP (Pollacco et al., 2006), OGLE (Udalski et al., 2015), Kepler (Borucki, 2016) and TESS (Ricker et al., 2015) obtained the time-series observations of numerous asteroids, variable stars, supernovae, AGN and more.

But one of the most productive space mission is the ESA Gaia (Gaia Collaboration, et al., 2016) which has already cataloged accurate positions, proper motions, parallaxes, photometry and spectra for over a billion stars in our Galaxy. Such a rich data is useful for different applications: charting three-dimensional map of the Milky Way, analyze of stellar properties (temperatures, metallicity, etc.), study of Galaxy structure and dynamics and many more.

Special interest is raised for searching for unknown open clusters and Galactic spiral arm sub-structures (Cantat-Gaudin et al., 2020; Castro-Ginard et al., 2018; Piatti et al., 2022) using the Gaia multi-star and multi-parameter catalog. However, analyzing this massive archive of information is an impossible mission for astronomers if special automated computational methods or AI tools are not involved.

The goal of present work is to use the Machine Learning algorithms to check the spatial clustering probability of small and sparsely populated apparent open clusters using the Gaia DR3 data (Gaia Collaboration, et al., 2023), which sometimes hard to distinguish from asterisms.

2. Open Clusters vs Asterisms

As a sample of candidate clusters, we selected 11 Open Clusters (OC) of Dolidze and Jimsheleishvili (Dolidze & Jimsheleishvili, 1966). The main parameters of these OCs are listed in Table 1. According to the authors, the clusters consist of up to 15 bright stars, each having magnitude range between 7 - 12 magnitudes. The angular sizes of the clusters vary between 12 - 34 arcmin.

*info@astronomia.ge, Corresponding author

doi: https://doi.org/10.52526/25792776-2025.72.1-40

Table 1. Dolidze-Jimsheleishvili Open Clusters

Name	RA (J2000)	DEC (J2000)	Angular size, arcmin	Number of stars	Magnitude range
DolidzeDzim 1	02:47:30	+17:16:00	12.0	12	8.5 - 11
DolidzeDzim 2	05:23:09	+11:25:47	12.0	12	9 - 10.5
DolidzeDzim 3	05:33:31	+26:31:30	15.0	10	9 - 11.5
DolidzeDzim 4	05:35:54	+25:57:00	28.0	15	6.5 - 9.5
DolidzeDzim 5	16:27:20	+38:01:58	27.0	7	9 - 11
DolidzeDzim 6	16:45:24	+38:21:00	17.0	5	9 - 12
DolidzeDzim 7	17:11:20	+15:28:43	20.0	6	10
DolidzeDzim 8	17:26:22	+24:12:23	14.0	6	8.5 - 9.5
DolidzeDzim 9	18:08:47	+31:33:18	34.0	15	8.5 - 11
DolidzeDzim 10	20:05:21	+40:31:25	20.0	12	8.5 - 11
DolidzeDzim 11	20:51:09	+35:53:07	13.0	12	9.5 - 12

Two clusters from this list, particularly Nos. 5 and 8, were marked by the authors as "doubtful". The first 9 clusters in the list were identified using spectral classification of stars in the blue spectral region. These groups of stars were suspected as clusters due to their resemblance to open clusters of type 2 (relatively small number of stars in the giant branch) according to the classification scheme proposed by Trumpler (1930). Two clusters from this group, i.e. Nos. 1 and 8, are removed from the catalogue of Dias et al. (2002) on the basis of the study by Archinal & Hynes (2003).

The Dolidze and Jimsheleishvili clusters were also analyzed by Tadross (2009) using the color-magnitude diagrams (CMD) of 2MASS photometry. The ages, metallicities, distances, and values of reddening were derived by fitting color-magnitude diagrams to the theoretical isochrones for all 11 clusters. The diameters of the clusters were also estimated. In a CMD fitting procedure, the exact location of the giant branch plays a crucial role. In this work, however, the giant branch was determined only by the colors of a few giants, without any discussion about their membership status.

Kazlauskas et al. (2013) examined the reality of the open Dol-Dzim 5 cluster using available photometric, spectroscopic, and astrometric information. They have shown that this concentration of stars could hardly be regarded as a physical ensemble and is most likely an asterism of several bright stars.

To study this problem with alternative method, we collected equatorial coordinates and parallaxes of stars down to 19th magnitudes from the Gaia DR3 catalogue (Gaia Collaboration, et al., 2023) in the fields with centers equal to Dolidze and Jimsheleishvili Open Clusters' centers and with diameters equal to 1 degree.

To find the possible grouping in these star lists, we used the Density-Based Spatial Clustering of Applications with Noise or DBSCAN (Ester et al., 1996) method from the `scikit-learn` Machine Learning Python package (Pedregosa et al., 2011) which finds core samples in regions of high density and expands the clusters from them in a 3D space. This method was used successfully by many authors to find spatial overdensities or cluster membership determinations (e.g. Caballero & Dinis, 2008; Gao et al., 2014, 2017).

The DBSCAN algorithm views clusters as areas of high density separated by areas of low density. Due to this rather generic view, clusters found by DBSCAN can be any shape, as opposed to k-means, which assumes that clusters are convex shaped. The central component of the DBSCAN is the concept of core samples, which are samples that are in areas of high density. A cluster is therefore a set of core samples, each close to each other (measured by some distance measure) and a set of non-core samples that are close to a core sample (but are not themselves core samples). There are two parameters to the algorithm, *MinPts* and ϵ, which formally define what we mean when we say dense. Higher *MinPts* or lower ϵ indicate a higher density necessary to form a cluster.

While the parameter *MinPts* primarily controls how tolerant the algorithm is towards noise (on noisy and large data sets it may be desirable to increase this parameter), the parameter ϵ is crucial to choose appropriately for the data set and distance function and usually cannot be left at the default value. It controls the local neighborhood of the points. When chosen too small, most of the data will not be clustered at all (and labeled as -1 for "noise"). When chosen too large, it causes close clusters to be merged into one cluster, and eventually the entire data set to be returned as a single cluster. Some heuristics for choosing this parameter have been discussed in the literature, for example, based on a knee in the nearest neighbor distances plot.

doi: https://doi.org/10.52526/25792776-2025.72.1-40

Figure 1 shows DBSCAN algorithm logical flow chart.

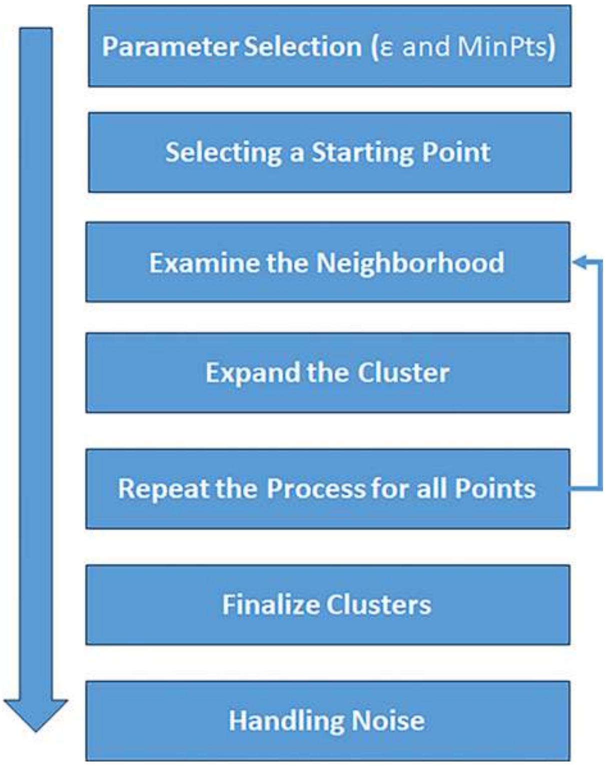

Figure 1. Logical flow chart of the DBSCAN algorithm.

3. Test Analyze

To test the potential of the DBSCAN method, we used it for well-known Open Clusters, NGC 1798 discovered by Edward Barnard in 1885 and NGC 6633 discovered in 1745 by Jean-Philippe de Chéseaux.

Figure 2. The CDS star maps of well-known Open Clusters, NGC 1798 (left) discovered by Edward Barnard in 1885 and NGC 6633 (right) discovered in 1745 by Jean-Philippe de Chéseaux.

Table 2 shows the main parameters of the Open Clusters NGC 1798 and NGC 6633.

Figure 3 and Figure 4 show the DBSCAN clustering results plotted in 3D graph (top right) with RA along X axes, DEC along Y axes and distance along Z axes. The subplot in the top left part of the figures show apparent map drawing of all the stars. The bottom left and right subplots show distribution of stars RA vs Distance and DEC vs Distance 2D spaces. The clusters found by the DBSCAN are marked with dark colors, while field stars classified by the DBSCAN not connected with the potential clusters, are marked with a light brown color.

The figures clearly show that both clusters are found by the DBSCAN in the correct distances, though there are groups of field stars overlaid on these clusters, but are located far from the clusters itself and are not connected with them.

doi: https://doi.org/10.52526/25792776-2025.72.1-40

Table 2. main parameters of the Open Clusters NGC 1798 and NGC 6633.

	NGC 1798	NGC 6633
RA, DEC (J2000)	05:11:39, +47:41:30	18:27:31, +06:33:59
Galactic long. & lat.	160.70, +4.85	36.09, +8.29
Angular size, arcmin	8.3	27
Distance, pc	3550	385
Size, pc	10	5
Members	161	90

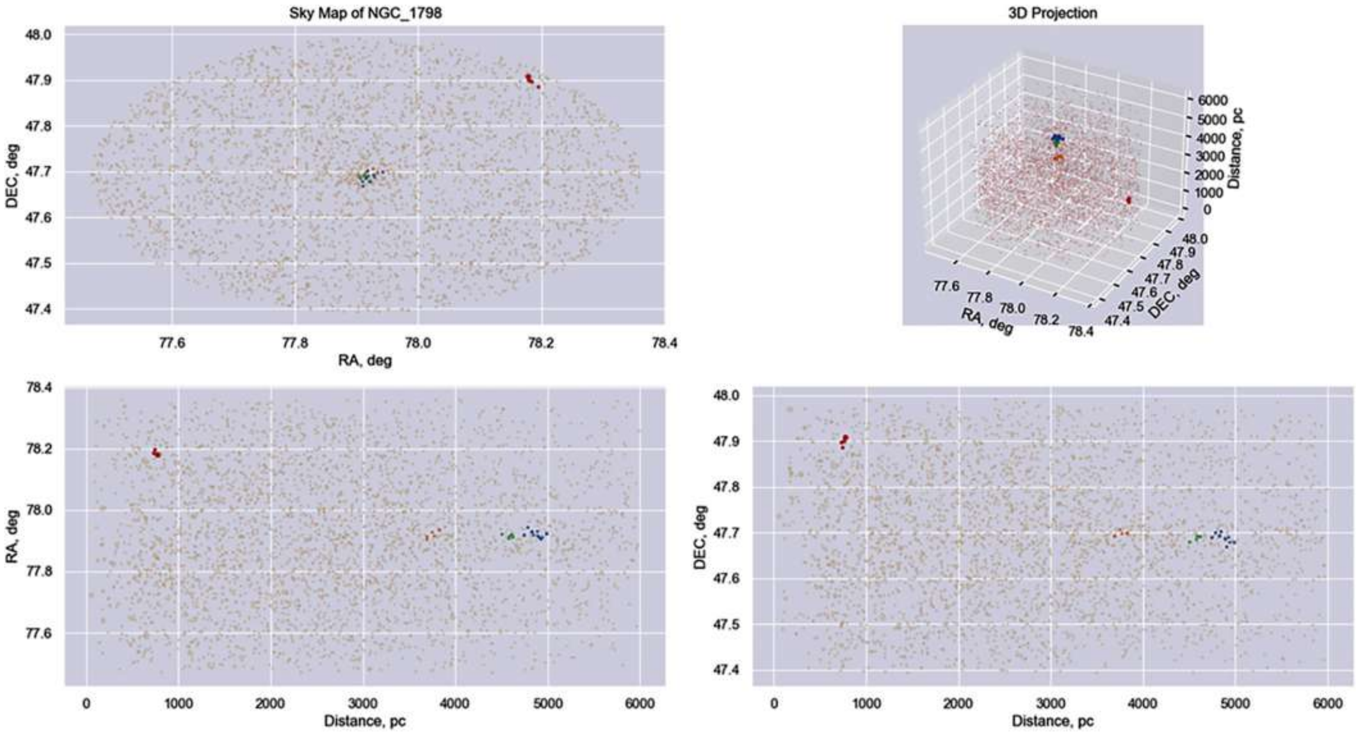

Figure 3. The DBSCAN clustering result plot for NGC 1798. The small group of stars marked by dark-red color centered at about RA = 78.2°and DEC=+47.9°and at a distance of around 800 pc exactly fits with the Open Cluster LP17 (Loktin & Popova, 2017).

In the field of the OC NGC 1798 we see another group of stars marked by the DBSCAN with dark red color and centered at about RA = 78.2°and DEC=+47.9°and at a distance of around 800 pc. This group fits perfectly with Open Cluster LP17 (Loktin & Popova, 2017) and did not intend to find it at all from the beginning. In addition, this finding clearly emphasizes the perfect potential of DBSCAN in clustering research.

4. Results and Discussion

We analyzed with the same DBSCAN algorithm all Dolidze-Jimsheleishvili Open Clusters. First of all, it should be emphasized that no one of the clusters clearly shows the high probability clustering tendency, though 3 of them could be considered as potential open clusters. In this work, we selected the following intial parameters: $MinPts = 6$ and $\epsilon = 0.3$. Figures 5 and Figure 6 show the results of the DBSCAN analyze for the clusters Dolidze-Dzim 2 (Figure 5) and Dolidze-Dzim 3 (Figure 6) which we consider to classify rather as asterisms, but not open clusters.

The analyzing of the DBSCAN results for all 11 Dolidze-Dzim clusters, we consider that 3 of them show potential clustering tendency: Dolidze-Dzim 6, Dolidze-Dzim 7 and Dolidze-Dzim 10. The figures below show corresponding plots from the DBSCAN. Further analyze is needed to find if other parameters including color-magnitude diagrams, metalicities, and especially proper motions of the stars showing positive clustering also tend to the same grouping.

doi: https://doi.org/10.52526/25792776-2025.72.1-40

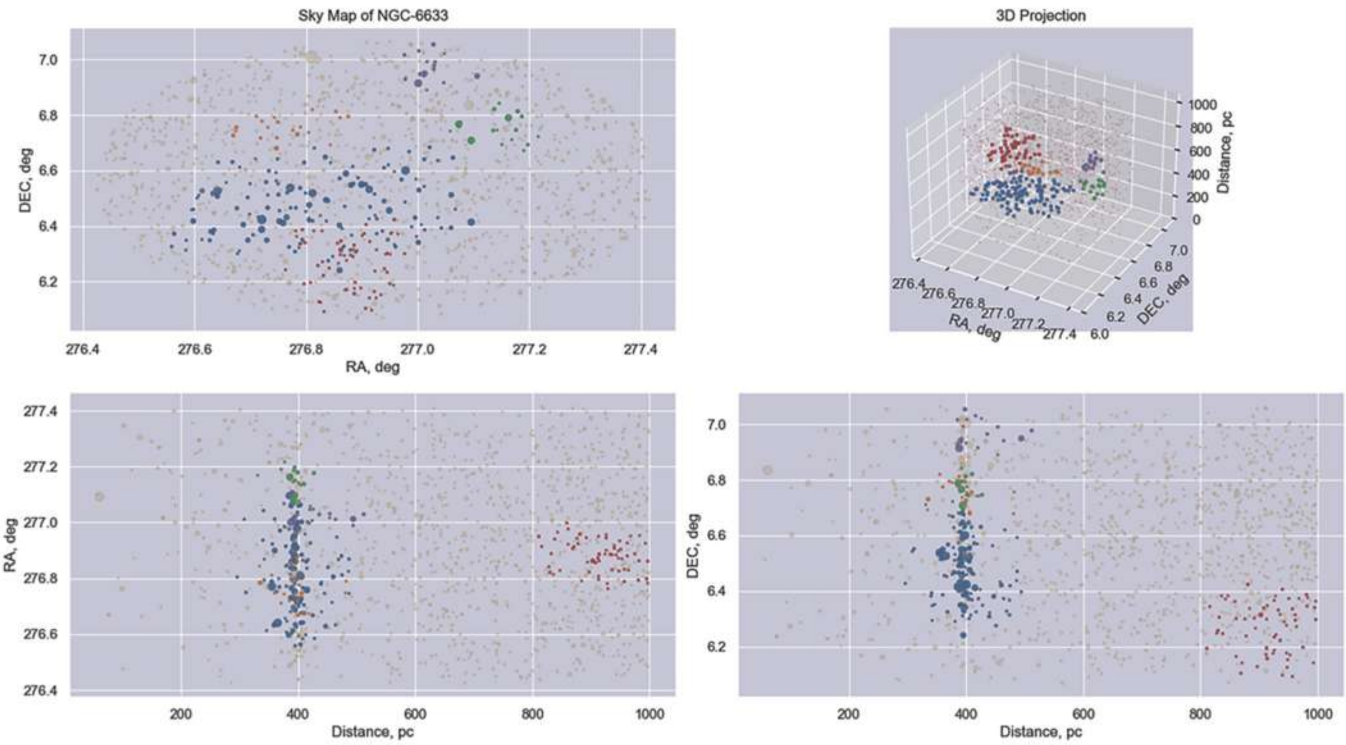

Figure 4. The DBSCAN clustering result plot for NGC 6633.

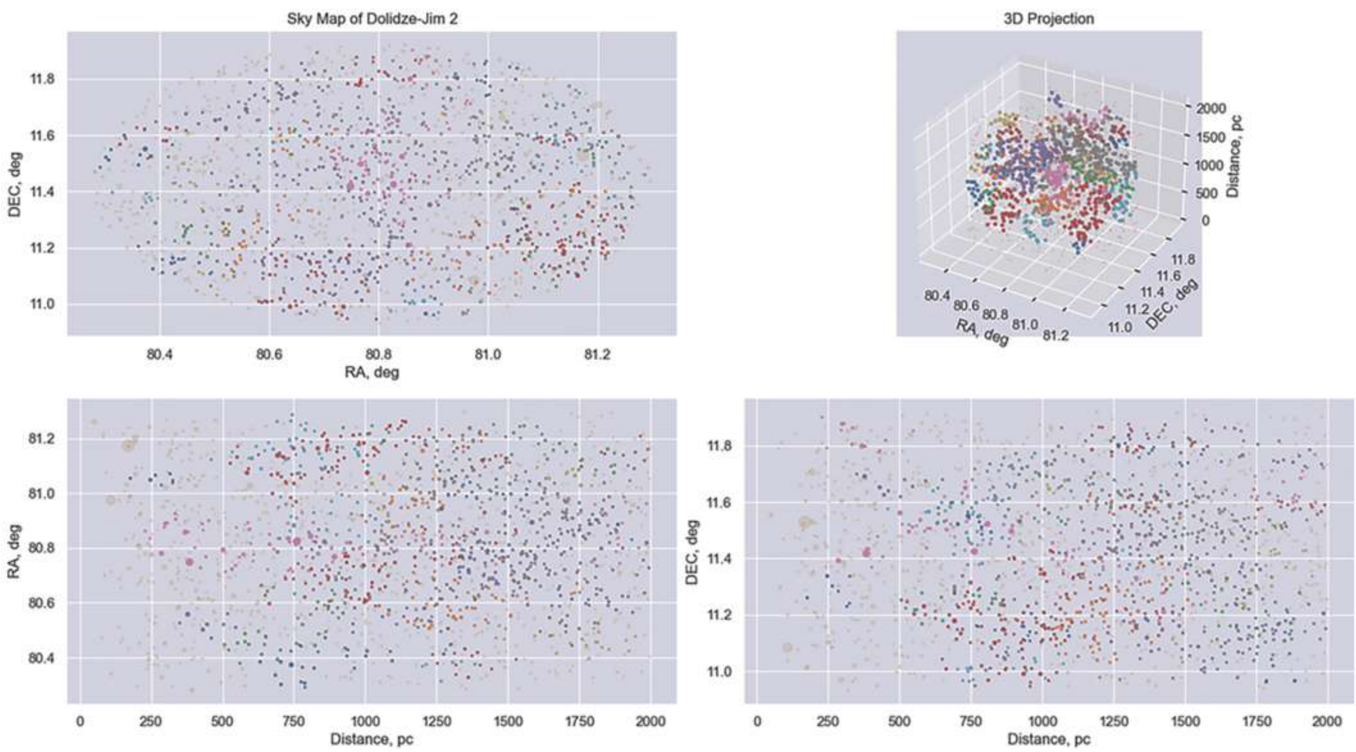

Figure 5. The DBSCAN clustering result plot for Dolidze-Dzim 2.

doi: https://doi.org/10.52526/25792776-2025.72.1-40

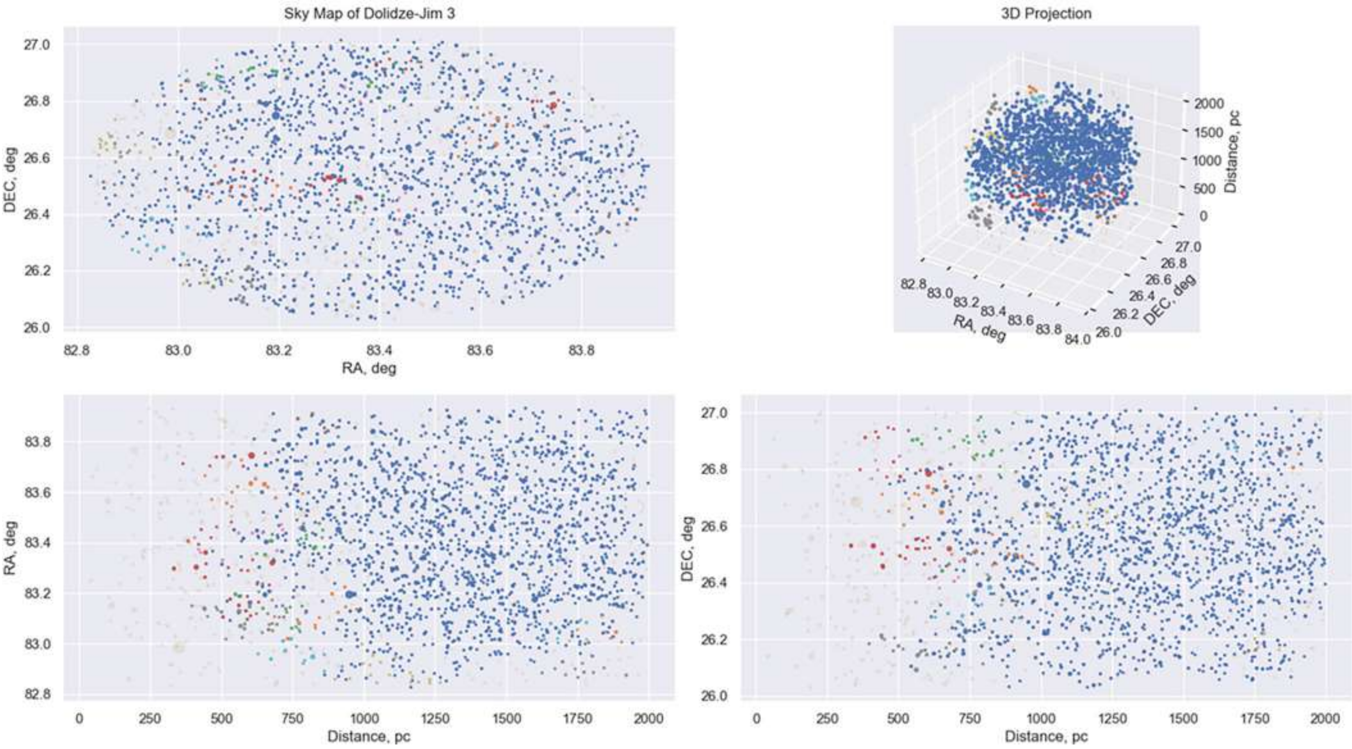

Figure 6. The DBSCAN clustering result plot for Dolidze-Dzim 3.

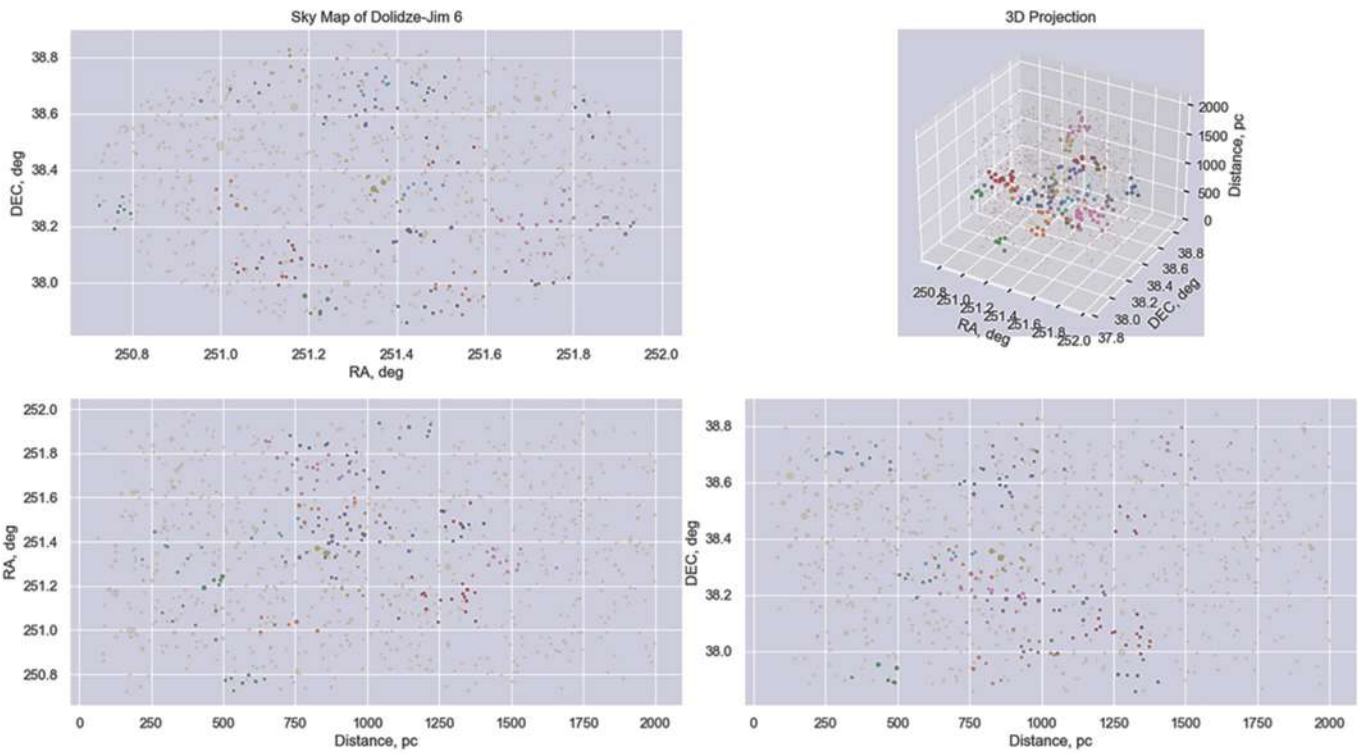

Figure 7. The DBSCAN clustering result plot for Dolidze-Dzim 6.

doi: https://doi.org/10.52526/25792776-2025.72.1-40

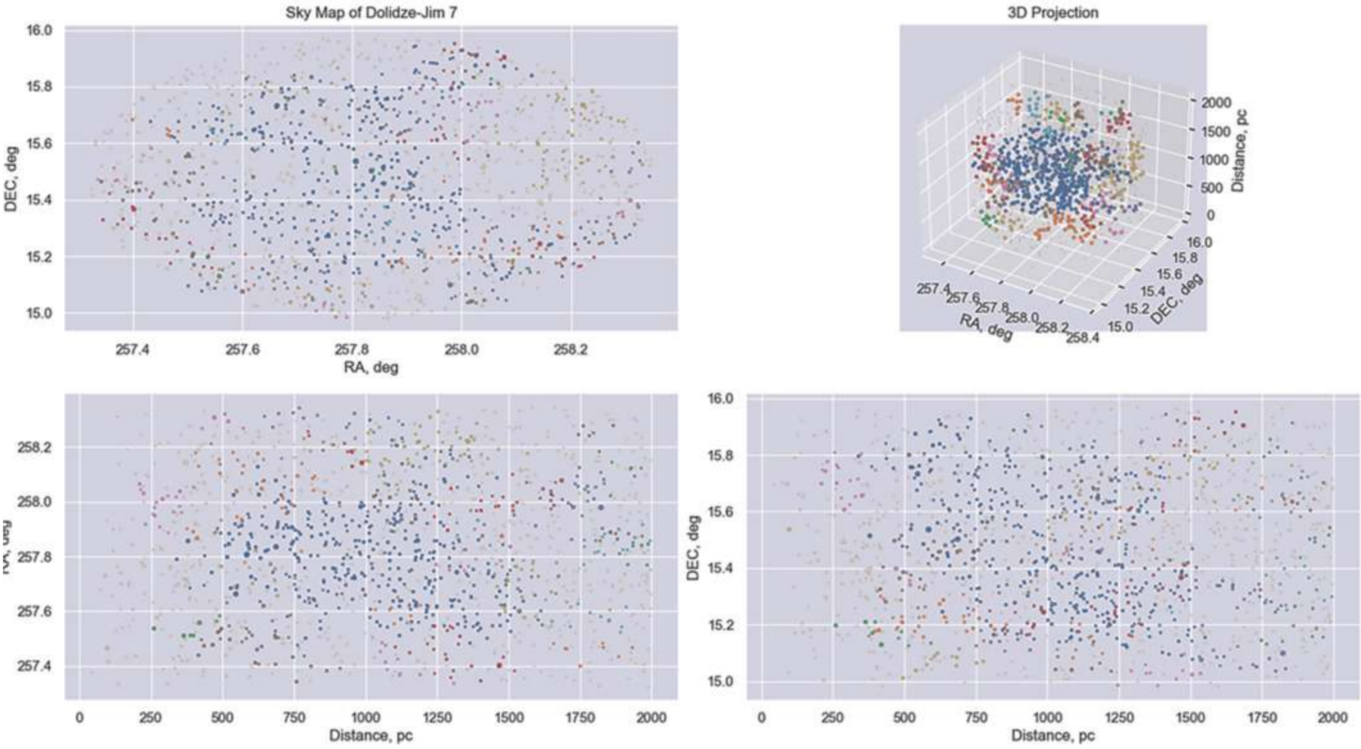

Figure 8. The DBSCAN clustering result plot for Dolidze-Dzim 7.

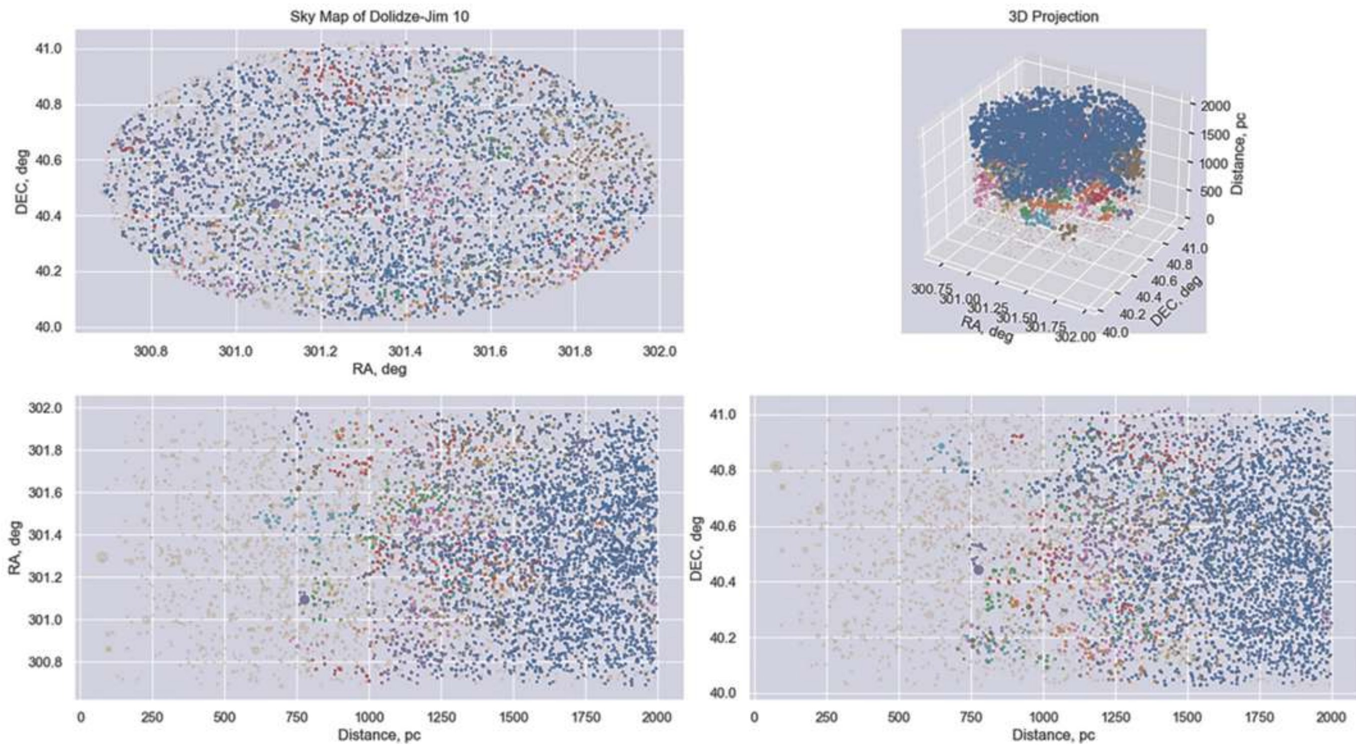

Figure 9. The DBSCAN clustering result plot for Dolidze-Dzim 10.

doi: https://doi.org/10.52526/25792776-2025.72.1-40

References

Archinal B., Hynes S., 2003, Star Clusters, Willmann-Bell Inc., Richmond, VA

Auvergne M., Bodin P., Boisnard L., et al. 2009, Astron. Astrophys. , 506, 411

Ball N. M., Brunner R. J., 2010, Journal of Modern Physics D, 19, 1049

Bellm E., 2014, in The Third Hot-wiring the Transient Universe Workshop, ed. P. R. Wozniak, M. J. Graham, A. A. Mahabal, R. Seaman, 27, 33

Borucki W. J., 2016, Reports on Progress in Physics, 79, 036901

Caballero J., Dinis L., 2008, Astron. Nach. , 329, 801

Cantat-Gaudin T., Anders F., Castro-Ginard A., et al. 2020, Astron. Astrophys. , 640, A1

Castro-Ginard A., Jordi C., Luri X., et al. 2018, Astron. Astrophys. , 618, A59

Chambers K., Magnier E., Metcalfe N., et al. 2016, eprint arXiv, 1612.05560

Dias W., Alesi B., Mointinho A., Lépine J., 2002, Astron. Astrophys. , 389, 871

Dolidze M., Jimsheleishvili G., 1966, Astronomicheskji Tsirkular, 382, 7

Ester M., Kriegel H.-P., Sander J., Xu X., 1996, in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. KDD'96 (AAAI Press), 382, 226

Gaia Collaboration, Prusti T., de Bruijne J. H. J., et al. 2016, Astron. Astrophys. , 595, 36

Gaia Collaboration, Vallenari A., Brown A., Prusti T., et al. 2023, Astron. Astrophys. , 674, 22

Gao X.-h., Chen L., Hou Z.-j., 2014, Chinese Astron. Astrophys., 38, 257

Gao X., Wang C., Gu X., Xu S., 2017, Acta Astronomica Sinica, 58, 46

Kaiser N., Burgett W., Chambers K., et al. 2010, Proc. SPIE: Ground-based and Airborne Telescopes III, 7733, 77330E

Kazlauskas A., Sperauskas J., Boyle R., 2013, New Astronomy, 19, 34

Loktin A., Popova M., 2017, Astrophys. Bull., 72, 257

Pedregosa F., et al., 2011, Journal of Machine Learning Research, 12, 2825

Pesenson M. Z., Pesenson I. Z., McCollum B., 2010, Advances in Astronomy, 2010, 350891

Piatti A., Illesca D., Massara A., et al. 2022, Mon. Not. R. Astron. Soc. , 518, 6216

Pojmanski G., Pilecki B., Szczygiel D., 2005, Acta Astronomica, 55, 275

Pollacco D. L., Skillen I., Collier Cameron A. e. a., 2006, The Publications of the Astronomical Society of the Pacific, 118, 1407

Ricker G. R., Winn J. N., Vanderspek R., et al. 2015, Journal of Astronomical Telescopes, Instruments, and Systems, 1, 014003

Tadross A., 2009, Astrophys. Space. Sci. , 323, 383

Trumpler R., 1930, Lick Obs. Bull., 420, 154

Udalski A., Szymański M. K., Szymański G., 2015, Acta Astronomica, 65, 1

York D. G., Adelman J., Anderson J., E. J., et al. 2000, Astron. J. , 120, 1579

doi: https://doi.org/10.52526/25792776-2025.72.1-40