

Universidad Internacional de la Rioja (UNIR)

ESIT

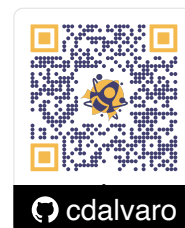
Máster Universitario en Inteligencia Artificial

Machine Learning Tools for Open Cluster Characteriza- tion with Gaia DR2 Data

Master Thesis

Author: Carlos David Álvaro Yunta

Advisor: Dr. César Augusto Guzmán Álvarez



City: Madrid, Spain

Date: 31st of December, 2020

Contents

List of Figures	iii
List of Tables	vi
Abstract	vii
Resumen	viii
Acknowledgements	ix
1 Introduction	1
2 Aims	5
2.1 General	5
2.2 Specific	5
3 State of the Art	6
3.1 Gaia Mission	7
3.2 Current Methods	8
3.3 Clustering Algorithms	11
4 Method	13
4.1 Data Mining	14
4.1.1 Download Process	16
4.2 Feature Selection	19
4.3 Soft Clustering with K-Means	21
4.4 Deep Embedded Clustering (DEC)	24

5	Results	28
5.1	Cluster Characterization with VO Tools	28
5.2	Comparing and Validating	34
5.2.1	NGC 2516	35
5.2.2	NGC 2632	37
5.2.3	NGC 2682	39
5.2.4	Melotte 25	41
5.3	Discussion	43
6	Conclusions and Future Work	45
6.1	Conclusions	45
6.2	Further Research	46
	References	47
	Paper	51

List of Figures

1.1	Examples of theoretical isochrones in the H-R diagram taken from Bressan et al. (2012, p. 16)	2
1.2	Examples of typical profiles in H-R diagrams for members of open clusters. On the left Melotte 22, on the right NGC 2682.	3
1.3	NGC 2682 configuration spaces.	3
1.4	Melotte 22 Parallax histogram, clearly differentiated around the value $\approx 7.3mas$ and the corresponding diagram from the photometric magnitudes.	4
3.1	Open Cluster Melotte 22 (Messier 45)	6
3.2	Open Cluster NGC 6494	7
3.3	Typical Clusterix 2.0 region selection panel.	9
3.4	Probability distribution in the proper motion space provided by Clusterix for NGC 2682.	10
4.1	Melotte 22 proper motions	13
4.2	Melotte 22 parallax and H-R diagram	14
4.3	OpenClust Catalogue Distribution	15
4.4	OpenClust Catalogue Selection Distribution	16
4.5	Data retrieval and analysis	17
4.6	Custom DB diagram	18
4.7	Pairwise relationships among variables using Melotte 22 data	20
4.8	K-Means comparisons with Melotte 22	22
4.9	K-Means model applied to Melotte 22	23
4.10	Melotte 22 H-R diagram with K-Means characterization	23
4.11	DEC model layer setup	25
4.12	DEC model applied to Melotte 22	25

4.13 H-R diagram comparison between DEC clustering and Clusterix+TOPCAT method.	26
4.14 Melotte 22 H-R diagram and parallax histogram with stars outside 0.25 and 0.75 quantiles filtered.	26
5.1 Clusterix 2.0 selection panel with Melotte 22 settings. Coordinates correspond to the center of Melotte 22 cluster, while the radius is the value registered for this cluster in the OpenClust catalogue multiplied by 1.5 (as explained in 4.1)	29
5.2 Aladin screenshot with equatorial coordinates for Melotte 22 stars referred to J2000.	30
5.3 Clusterix 2.0 Step 2/3: Region selection.	31
5.4 Kinematic probability associated to the membership of the open cluster. This probability is derived from the proper motion frequency analysis. . . .	31
5.5 Melotte 22 first selection using validation method. Parallax is $7.304mas$ with 0.797 of standard deviation.	32
5.6 Melotte 22 second selection. The estimated number of cluster members is 709.	32
5.7 Comparison between first and second selection H-R diagrams.	33
5.8 Comparison between first and second selection H-R diagrams.	33
5.9 NGC 2516 Clusterix+TOPCAT characterization.	35
5.10 NGC 2516 K-Means characterization. Identified as cluster $g3$	35
5.11 NGC 2516 DEC characterization. Identified as cluster $g3$	35
5.12 NGC 2516 DEC (filtered) characterization. Identified as cluster $g3$	35
5.13 NGC 2632 Clusterix+TOPCAT characterization.	37
5.14 NGC 2632 K-Means characterization. Identified as cluster $g1$	37
5.15 NGC 2632 DEC characterization. Identified as cluster $g1$	37
5.16 NGC 2632 DEC (filtered) characterization. Identified as cluster $g1$	37
5.17 NGC 2682 Clusterix+TOPCAT characterization.	39
5.18 NGC 2682 K-Means characterization. Identified as cluster $g4$	39
5.19 NGC 2682 DEC characterization. Identified as cluster $g2$	39
5.20 NGC 2682 DEC (filtered) characterization. Identified as cluster $g2$	39
5.21 Melotte 25 Clusterix+TOPCAT characterization.	41

5.22	Melotte 25 K-Means characterization. Identified as cluster <i>g4</i>	41
5.23	Melotte 25 DEC characterization. Identified as cluster <i>g5</i>	41
5.24	Melotte 25 DEC (filtered) characterization. Identified as cluster <i>g5</i>	41

List of Tables

4.1	Properties and descriptions of the fields used in this work.	19
4.2	Parameters shown are proper motion in right ascension and declination, parallax with their respective deviations and number of stars corresponding to Melotte 22 data.	27
5.1	NGC 2516 DEC model hyperparameters.	36
5.2	NGC 2516 results.	36
5.3	NGC 2632 DEC model hyperparameters.	38
5.4	NGC 2632 results.	38
5.5	NGC 2682 DEC model hyperparameters.	40
5.6	NGC 2682 results.	40
5.7	Melotte 25 DEC model hyperparameters.	42
5.8	Melotte 25 results.	42
5.9	Right ascension, declination, radius and number of stars of studied clusters. The number of stars corresponds to those stars contained within a cone of center (α, δ) and radius the cluster's radius multiplied by a factor of 1.5. . .	43

Abstract

The characterization and understanding of *Open Clusters* (OCs) allow us to understand better properties and mechanisms about the Universe such as stellar formation and the regions where these events occur. They also provide information about stellar processes and the evolution of the galactic disk.

In this work, we present a novel method to characterize OCs. Our method employs a model built on *Artificial Neural Networks* (ANNs). More specifically, we adapted a state of the art model, the *Deep Embedded Clustering* (DEC) model for our purpose. The developed method aims to improve classical state of the arts techniques. We improved not only in terms of computational efficiency (with lower computational requirements), but in usability (reducing the number of hyperparameters to get a good characterization of the analyzed clusters). For our experiments, we used the *Gaia DR2 database* as the data source, and compared our model with the clustering technique *K-Means*. Our method achieves good results, becoming even better (in some of the cases) than current techniques.

Key words. data analysis, deep embedded clustering, gaia, machine learning, open clusters characterization

Resumen

La caracterización y conocimiento de *Cúmulos Abiertos* permite conocer mejor propiedades y mecanismos del Universo tales como la formación de estrellas y las regiones donde se dan estos procesos. También permiten obtener información sobre procesos estelares y la evolución del disco galáctico.

En este trabajo presentamos un método novedoso para caracterizar estos cúmulos. Nuestro método hace uso de un modelo basado en *Redes Neuronales Artificiales*. Más concretamente, adaptamos un modelo del estado del arte, el *Deep Embedded Clustering* (DEC), a nuestro problema. El método desarrollado tiene como objetivo mejorar las técnicas clásicas del estado del arte. Con nuestro método, no sólo mejoramos en términos de eficiencia de cálculo (consiguiendo menores requisitos computacionales), también mejoramos en usabilidad (reduciendo el número de hiperparámetros para conseguir una buena caracterización de los cúmulos analizados). Para nuestros experimentos, usamos la base de datos *Gaia DR2* como fuente de datos, y comparamos nuestros modelos con el algoritmo de clustering *K-Medias*. Nuestro método consigue buenos resultados, siendo incluso mejor (en algunos casos) que las técnicas actuales.

Palabras Clave: análisis de datos, caracterización de cúmulos abiertos, gaia, inteligencia artificial

Acknowledgements

I would like to thank my thesis advisor Dr. César Augusto Guzmán Álvarez. He has guided me through the process of writing this work and has reviewed the whole work suggesting me improvements.

I would also like to acknowledge my father who always gives me its advices and helps me by giving me its priceless point of view. I would also like to thank my mother and girlfriend, who have supported me through the entire process of making and writing this project.

Finally, I would like to mention the acknowledgments that the developers of the main tools used in this project politely indicate on their web pages.

This work has made use of data from the European Space Agency (ESA) mission *Gaia* (<https://www.cosmos.esa.int/gaia>), processed by the *Gaia* Data Processing and Analysis Consortium (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.

This publication makes use of VOSA, developed under the Spanish Virtual Observatory project supported by the Spanish MINECO through grant AyA2017-84089. VOSA has been partially updated by using funding from the European Union's Horizon 2020 Research and Innovation Programme, under Grant Agreement n^o 776403 (EXOPLANETS-A)

This research has made use of the VizieR catalogue access tool, CDS, Strasbourg, France (DOI : 10.26093/cds/vizier). The original description of the VizieR service was published in 2000 ([Ochsenbein, Bauer, & Marcout, 2000](#), A&AS 143, 23)

This research has made use of the SIMBAD database, operated at CDS, Strasbourg, France ([Wenger et al., 2000](#), 2000,A&AS,143,9), "The SIMBAD astronomical database", Wenger et al.

Chapter 1

Introduction

Stellar open clusters (OCs) ([Janes & Adler, 1982](#)) are groups of stars gravitationally bounded originated from a single molecular gas cloud. They share the same chemical composition and age. Moreover, they have similar relative positions and proper motion. Those astronomical objects are relevant to understand the spiral structure, the dynamics and the chemical evolution of our galaxy.

Although most stars in the Milky Way are presented isolated, it is considered that most of them (or even all) are formed in clustered environments and spend a period of time gravitationally bounded with their siblings embedded in their original molecular cloud ([Clarke, Bonnell, & Hillenbrand, 2000](#)) ([Portegies Zwart, McMillan, & Gieles, 2010](#)). The evolution of those systems tends to sparse them in a few million years by interacting gravitationally with other systems. Galactic tidal forces and mechanisms that involve gas loss driven by stellar feedback are other causes of disruption ([Brinkmann, Banerjee, Motwani, & Kroupa, 2017](#)). Nevertheless, a small fraction of those systems will survive in the initial state and persist bounded in larger timescales.

Young OCs allow us to research star formation regions and improve our understanding about the mechanisms that create those stars. On the other hand, older open clusters give us information about stellar processes and how the galactic disk evolves. Some highly disturbed orbits could also provide evidence of recent merge events and accretion traces from outside the galaxy ([Cantat-Gaudin et al., 2016](#)).

The study of OCs has been pushed forward thanks to the huge and precise dataset from the Gaia mission ([Collaboration et al., 2016](#)) Gaia DR2 ([Gaia et al., 2018](#)), available since 2018. This dataset has helped to review already known open clusters and to find

new ones.

Stars that belong to the same OC share relative positions, inherited from their original gas cloud. This means that their distances to the Earth are similar for all of them and, therefore, they have a narrow dispersion in their parallax value. They also share similar values of proper motion, both in right ascension and declination. Another property they share is their chemical composition. Thus, their metallicity must be uniform, since these stars were born from the same gas cloud and at the same time stage. However, to take this last property into account, we are faced with the drawback that this parameter is poorly reported in Gaia DR2 database.

We will avoid this issue by looking at $E(b-r)/G_{mag}$ diagrams that show their connection with the isochrone curves. These curves are a tool used to determine how stars evolve in time (Bressan et al., 2012). The isochrones are derived from theoretical models. These models are mainly based on metallicity and mass/brightness ratio of stars. They also have a direct relation with that presented by the Hertzsprung–Russell (H-R) diagram of the stars belonging to the cluster. Examples of these curves are shown in Figure 1.1.

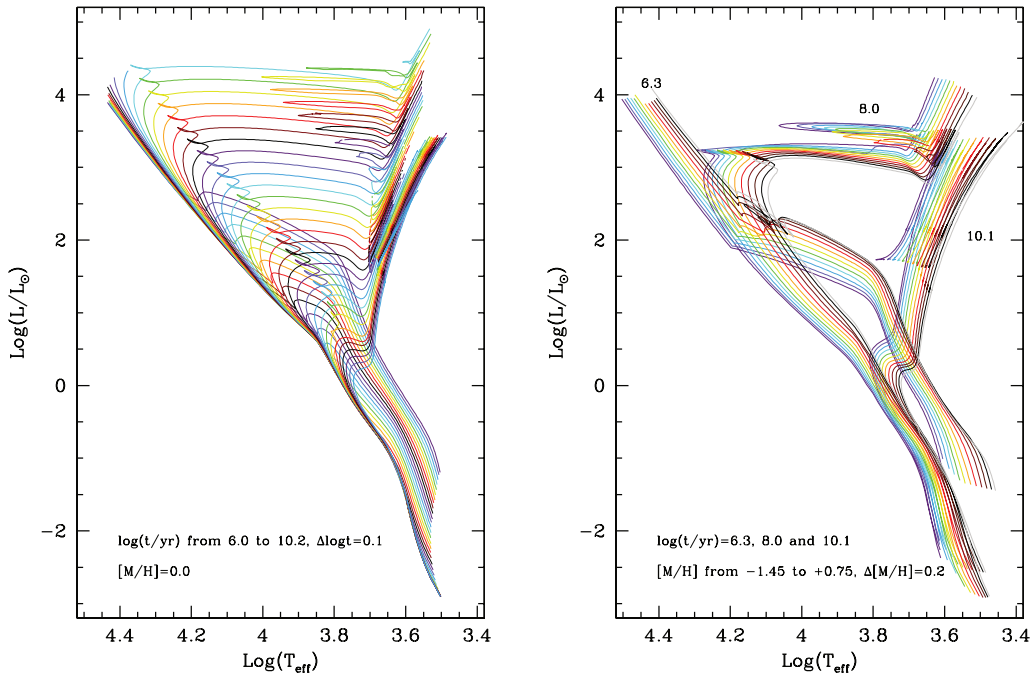


Figure 1.1: Examples of theoretical isochrones in the H-R diagram taken from Bressan et al. (2012, p. 16)

Since these stars were born close in time, they should have a sharp and well-defined profile, with low scattering through the main sequence in the H-R diagram. Examples for

this kind of diagrams are presented in Figure 1.2.

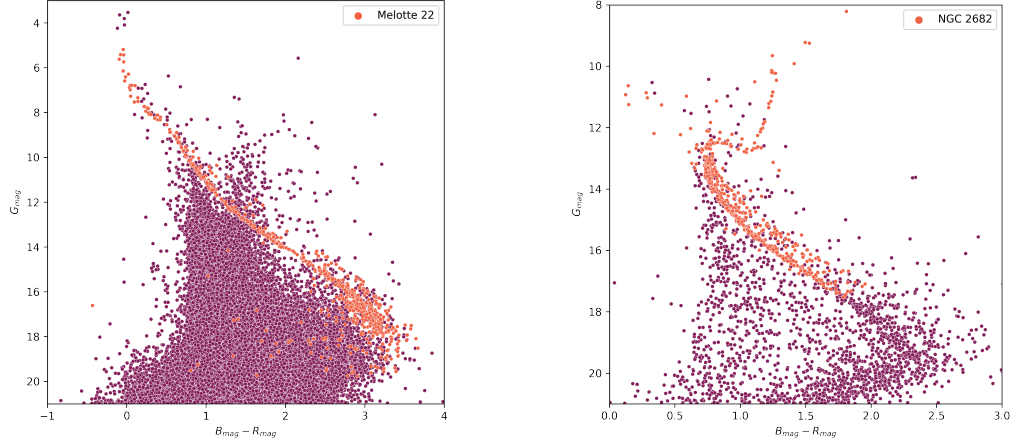


Figure 1.2: Examples of typical profiles in H-R diagrams for members of open clusters. On the left Melotte 22, on the right NGC 2682.

We will take advantage of these properties only to validate our characterization of OCs, but not to determine what stars belong to them. As we will explain later, we will only use dynamic properties such as proper motion and parallax to characterize open clusters.

Although the members of a cluster appear in the observational visual field as an overdensity in positions, as shown in Figure 1.3a, these coordinates are not useful to separate those stars that belong to the cluster from the other that do not. However, if we look for overdensities in the proper motion configuration spaces, it is possible, at least at first instance, to assume a possible membership cut (see Figure 1.3b).

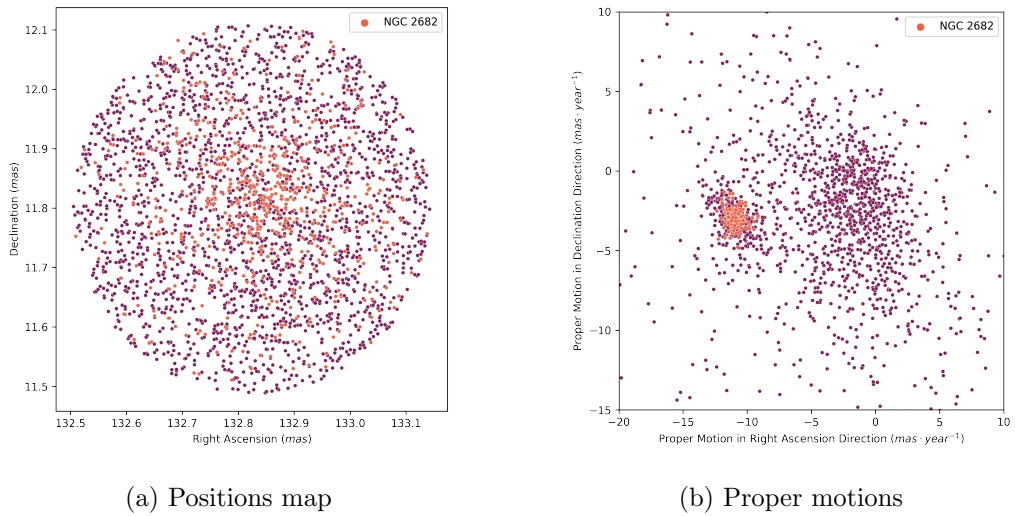


Figure 1.3: NGC 2682 configuration spaces.

Overdensities in proper motion are not always so evident. In order to improve the characterization capabilities of our model, we will consider the parallax distribution as well.

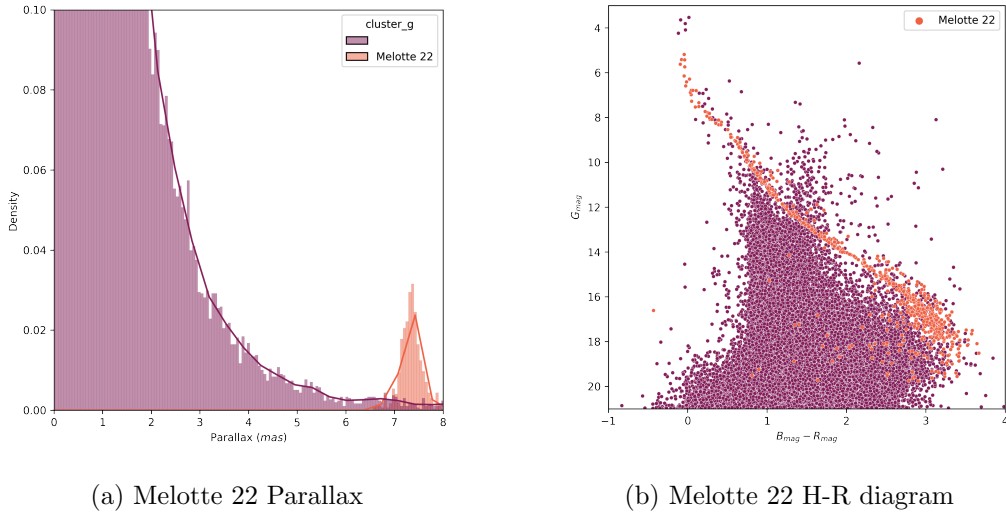


Figure 1.4: Melotte 22 Parallax histogram, clearly differentiated around the value $\approx 7.3\text{mas}$ and the corresponding diagram from the photometric magnitudes.

In many other cases things are not so simple, and the characterization process can turn very difficult and tedious using tools such as the available ones at the Virtual Observatory (VO). Furthermore, it frequently requires a parametrization based on prior knowledge regarding the studied region. The aim of this work is to test an unsupervised and non-parameterized model based on Machine Learning (ML) tools.

We start by enumerating the aims we try to achieve with this work. Then, we expose the current *state of the art*, mentioning the *Gaia Mission* used as data source in this work and describing other related works that try to solve this problem. The state of the art ends with a brief summary for some of the clustering algorithms available today. In the *Method* chapter, we explain our model to improve open cluster characterization. Also, we show an alternative method based on current procedures. We have used this alternative technique as a validation method to test ours. Then, a comparison between a set of results obtained with our method and the results obtained for the same clusters with the validation method is presented. Finally, we extract some conclusions based on the obtained results and present further researches that can be done in light of the results.

Chapter 2

Aims

2.1 General

The primary aim of this thesis is to *build an unsupervised clustering model for open cluster characterization*. The model must be *non-supervised and non-parameterized* to fit a wide range of clusters without the need for fine-tuning a high number of hyperparameters.

2.2 Specific

To achieve the general goal, we will set the following milestones:

- Gather information on the *state of the art*.
- Research unsupervised clustering algorithms suitable for grouping stars previously recovered.
- Recover data from Gaia DR2 database. This data will be taken as source for the machine learning model to characterize open clusters by grouping stars into clusters.
- Select and implement unsupervised clustering algorithms from the previous study.
- Use chosen algorithms with different datasets to find OCs.
- Look for an independent technique to use it as a validation method.
- Validate OCs found with our custom model by making comparisons with the validation method.

Chapter 3

State of the Art

An initial approach for finding OCs is the search for overdensities along the galactic disk. In general, this is a good starting point but, although it seems simple, it presents a fundamental problem already discussed. The near field around the OC is filled with two types of distinct star populations: those who belong to the OC (tens or hundreds to a few thousand) and a background made up of thousands or millions of stars that are not. Finding out which stars belong to the first group is the problem faced in this work. This selection is crucial to properly characterize the fundamental properties of the cluster (dynamics, total mass, age, chemical composition, among other).

Sometimes, the problem is easy to solve, as we have seen, by studying astrometric parameters and looking for overdensities in the proper motion configuration space as well as in the parallax space.

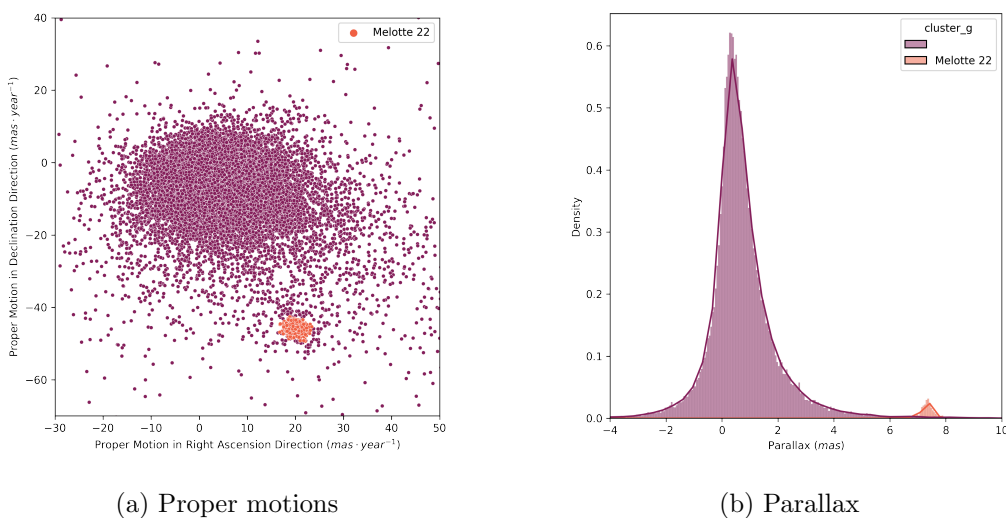


Figure 3.1: Open Cluster Melotte 22 (Messier 45)

Figure 3.1a shows the proper motion distribution in right ascension and declination. It is easy to see that a subgroup can be located at $[20, -45]$. Figure 3.1b shows the parallax distribution and confirms an overdensity at $\approx 7.3\text{mas}$ corresponding to Mellote 22 OC (See also Figure 1.4a).

However, as shown in Figures 3.2a and 3.2b, in general it is not as easy and becomes necessary to consider other parameters such as distances, or even metallicity and age (derived from isochrone curves). Sometimes even, photometric data may be required for the stars within the studied field.

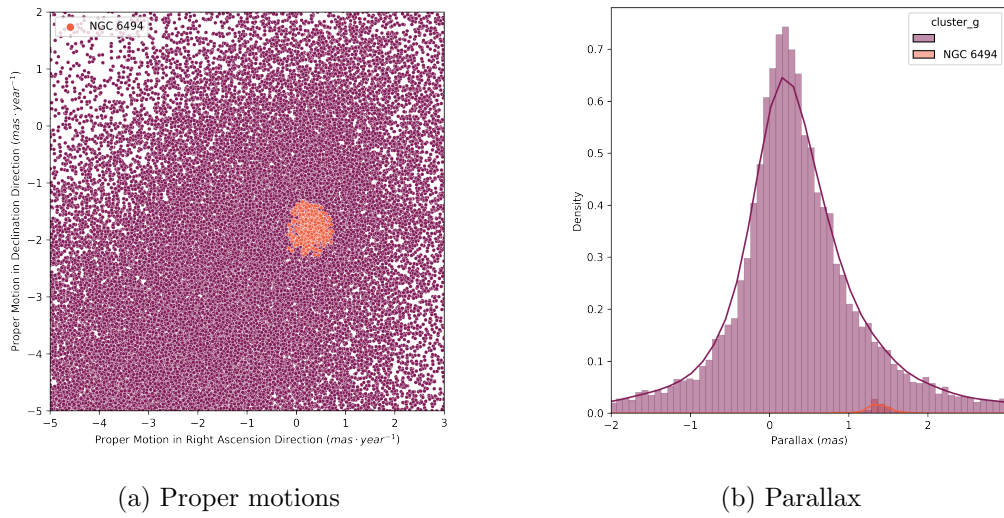


Figure 3.2: Open Cluster NGC 6494

At this point, it seems reasonable that the study of each individual group requires its own parametrization and technique to identify it. With Gaia DR2, which contains high-quality astrometric and photometric data for a huge number of stars, a new opportunity arises to simplify and optimize these processes. In this context, new approaches are being developed to improve current detection methods and automate OC characterization.

3.1 Gaia Mission

In this work we make use of Gaia DR2 since DR3 has not been released in time for us to include it. Gaia DR2 is a multidimensional dataset obtained by ESA's Gaia mission (located at L2, 1.5 million kilometers from Earth) and operational since 2014. The catalogue has high precision and accuracy astrometric data for more than 1.7 billion stellar sources, and magnitudes in three photometric filters (G, BP and RP) for more than 1,300 million

sources. 80% of Gaia DR2 sources are weaker than $G_{mag} \approx 18$. For magnitudes $G \approx 21$ nominal uncertainties reach $2mas$ for parallaxes and $5mas \cdot years^{-1}$ in proper motions. But for closer sources, (at the bright end, $G < 14$), the precision is of the order of $0.02mas$ and $0.05mas \cdot year^{-1}$ respectively (Cantat-Gaudin et al., 2018). In general, open clusters, which are the main subject of this work, are at optimal data thresholds.

In this context, and taking advantage of the opportunity provided by Gaia DR2, new approaches are being developed to improve current detection methods and automating OC characterization as much as possible. This is the case of tools such as Clusterix 2.0 (Balaguer-Núñez et al., 2020), TOPCAT (Taylor, 2005) and VOSA (Bayo et al., 2008), (all of them developed by the Virtual Observatory). These tools work together and offer good results, although it is not possible to use them in an unsupervised process and they require different parameterizations for each study case.

3.2 Current Methods

TOPCAT by itself does not offer an independent discrimination mechanisms to researchers. They have to try an initial selection, starting from the representation of proper motions, by delimiting the area where the cluster most probably appears. When this is not possible, it is necessary to have a previous knowledge about the cluster profile and to parametrize some approximated values. In both cases, the obtained selection must then be optimized step by step until a result that meets the confidence levels is achieved.

Another tool is Clusterix 2.0, which is an interactive web-based tool (Balaguer-Núñez et al., 2020). It takes the proper motion diagram without making any prior assumption about the membership of the candidate star and determines empirically the frequency functions.

Clusterix uses normal Gaussian kernel functions, defined as:

$$K(a, b) = \frac{1}{2\pi h^2} \exp \left[-\frac{1}{2} \frac{(a - a_i)^2 + (b - b_j)^2}{h^2} \right] \quad (3.1)$$

(a, b) are referred to the proper motion configuration space, while a point located in the center of the cell (i, j) provides the maximum contribution for calculating the local density. h is called the *smoothing* parameter and it is measured in the same units as proper motion.

Various clustering algorithms have been studied in the past with the purpose of finding

a valid non-parameterized method. UPMASK (Krone-Martins & Moitinho, 2014) was developed to only use photometry and positions, and then it was adapted to the Gaia data (Cantat-Gaudin et al., 2018) based only on proper motion and parallax. A similar approach has been carried out using DBSCAN and machine learning algorithms to discover new clusters.

In this sense, Clusterix also claims to be a non-parameterized method, but it critically depends on the initial selection for the field sizes to be analyzed.

Cluster info: 269.2667,-18.985_21.75_arcmin_GAIADR2

Selection of the "cluster+field" and "only field" regions

Click *Drawing Info* button if you need help on how to make the region selection

Area definition: ☒ Cluster+Field ☐ Void ☐ Field

Clear

Cluster+field:	Cluster+field area
269.27,-18.985,0.25;	0.19634954084936207
Void:	Void area
269.27,-18.985,0.32;	0.12534954687823274
Only field:	Field area
269.27,-18.985,0.3625;	0.09112582190818896

Membership determination parameters

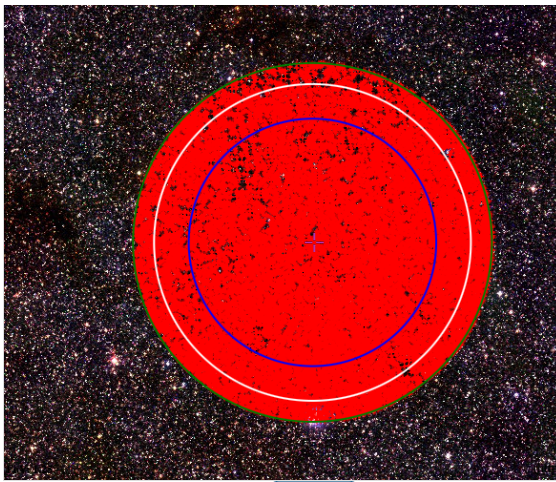
Proper motion limits (mas/yr)
Maximum μ : Maximum μ err:
8.0 10.0

Magnitude range \leq mag. \leq

Smooth param (mas/yr) (?):
0.4669142209204013

Fine tuning values
 γ threshold (?): 3.0
Empirical frequency function min value \longleftrightarrow Probability min value
0 0
 \leq pmRA \leq
 \leq pmDEC \leq

Matrix size ?
☒ Normal ☐ High precision
Total number of stars: 17898
Number of stars in the "cluster+field" region: 8453
Number of stars in the "field" region: 3837
Field sample size? 3837



Drawing info

Figure 3.3: Typical Clusterix 2.0 region selection panel.

Clusterix relies on several assumptions, among other:

The non-field population does not occupy the entire workspace, but is spatially concentrated, which makes it possible to distinguish two regions in the workspace: the only field region (label 'f'), dominated by star fields, and the cluster + field region (label 'c+f'), which includes both star fields and not star fields (Balaguer-Núñez et al., 2020).

This fact implies defining three areas or regions with different radius. The first 'c+f' corresponds to the one in which the cluster members are presumed to be contained together with other star fields that are not part of the cluster. The second region is the broadest

and assumes that it only contains stars in an extended visual field without components of the cluster. The third region is the intermediate one and is out of analysis (void area), since it would correspond to a possible transition zone between the other two.

The right choice of these radii, even having a previous estimation for the ‘c+f’ region, highly affects the execution of the algorithm and, in general, requires a considerable wide field ‘f’. There is no rule of thumb that defines relative proportions of these areas.

Finally, when an acceptable result is obtained, what Clusterix provides is a probability field associated to the set of stars under study. The recovered dataset contains all fields from Gaia database and an extra column with the probability value of each object to belong to the open cluster. This new dataset can be exported to TOPCAT for further optimization and final processing. Figure 3.4 shows an example of the probability distribution in the proper motion space for NGC 2682.

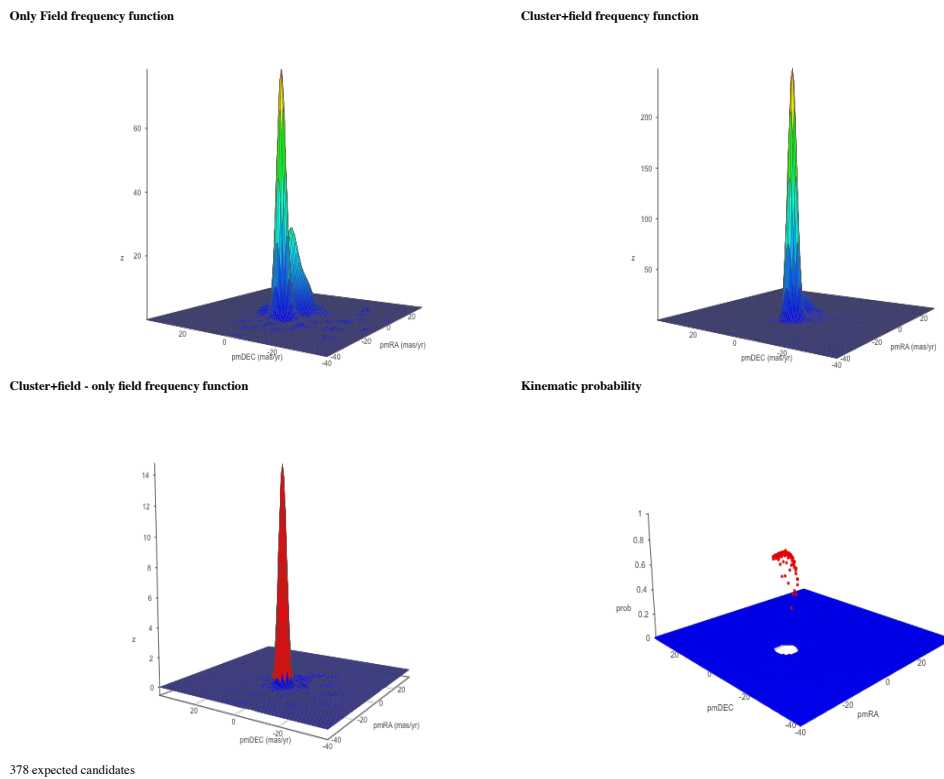


Figure 3.4: Probability distribution in the proper motion space provided by Clusterix for NGC 2682.

Finally, it is necessary to refer to the work developed on (Castro-Ginard et al., 2020). It presents a method based on machine learning techniques to make a systematic search for overdensities in the astrometric space of the galactic disk and a later identification of

OCs using photometric information, also from Gaia DR2.

The method includes two phases: the first one uses an unsupervised clustering algorithm, DBSCAN, to search for overdensities $(l, b, \pi, \mu_{\alpha*}, \mu_{\delta})$, and then applies a deep learning Artificial Neural Network (ANN), previously trained with magnitude diagrams, to identify isochrone patterns within the detected overdensities and thus proceed to confirm them as OC.

It should be noted that for the execution of this method, MareNostrum 42 (Barcelona Supercomputing Center) was used. So the neural network could handle the image recognition process with isochrone patterns and not applying theoretical models derived from values such as metallicity or masses, among other.

The result of this work is the recognition of 582 new open clusters distributed along the galactic disk for a galactic declination $b < 20^\circ$, which has meant an increase of 45% over previously ones.

Our aim in this work, linked to computational limitations compared to MareNostrum, is not to perform a blind search for clusters, but to obtain a method, within the machine learning environment, that allows the characterization of open clusters through unsupervised and non-parameterized procedures in a single step. This method should not require further refinements with other tools.

The result of Clusterix + TOPCAT, applied to the same clusters, will be used only for comparing and validating our results.

3.3 Clustering Algorithms

As we will explain later, what we need to accomplish our aim is an algorithm that manages to make data groups based on the dynamic properties of the stars.

There are several clustering algorithms: *K-Means*, *Mean-Shift Clustering*, *DBSCAN* (Ester, Kriegel, Sander, Xu, et al., 1996) among other. While each one behaves better according to the distribution of the objects to clusterize, we have chosen K-Means by its simplicity and good results.

K-Means requires a single parameter, the number of clusters to build (N). The algorithm looks for N center points which are vectors of the same dimension as the number of selected features. K-Means starts by making an initial groups configuration and then, it reassigns objects to other groups iteratively by minimizing the distance among points

inside the new group and maximizing the distance with the centers of the other groups.

The algorithm stops once a maximum number of iterations is reached, or when changes between iterations is lower than a minimum.

This algorithm works well and gives good results at first approach. However, we need to set a large number of clusters to find the open cluster we are looking for. This fact complicates the identification of the OC and many times the found cluster contains too many outliers. For that reason, we searched for a K-Means refinement based on an artificial neural network.

The *Unsupervised Deep Embedding for Clustering Analysis* model or *DEC* (Xie, Girshick, & Farhadi, 2016) takes K-Means as its starting point, but then, it trains an autoencoder to reduce the feature space and pass this transformed data through a Clustering Layer which refines the previous selection.

The mechanism of this model is explained in Section 4.4 and its implementation is available in `cdalvaro.ml.dec` package.

Chapter 4

Method

In this chapter we address the different steps followed to perform the creation of the unsupervised clustering model for open cluster characterization.

All code has been developed with the Python programming language ([Van Rossum & Drake, 2009](#)). Other auxiliary tools like *Docker* ([Merkel, 2014](#)), *PostgreSQL* ([PostgreSQL, 2020](#)), *Jupyter Notebooks* ([Kluyver et al., 2016](#)), and frameworks such as *Astropy* ([Astropy Collaboration et al., 2013](#)) ([Astropy Collaboration et al., 2018](#)), *Scikit-Learn* ([Pedregosa et al., 2011](#)), *Seaborn* ([Waskom et al., 2017](#)), *SQLAlchemy* ([Bayer, 2012](#)) and *Keras* ([Chollet et al., 2015](#)) have been used too.

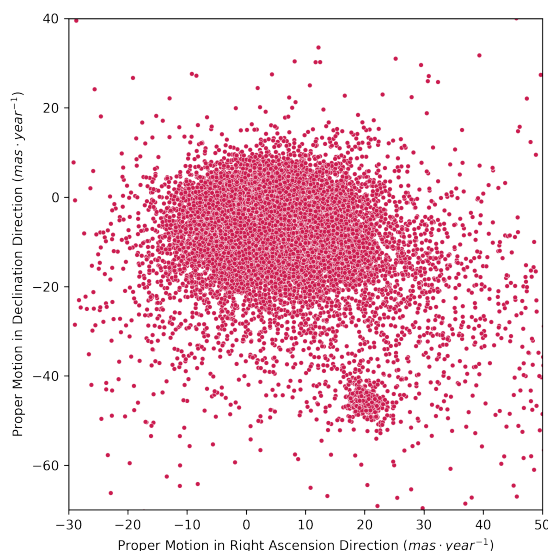


Figure 4.1: Melotte 22 proper motions

For the sake of simplicity, we are taking *Melotte 22* (*Messier 45*) ([Elsanhoury & Nough, 2019](#)) to illustrate the whole method. This cluster is well studied and its stars are not too

mixed with other stars that do not belong to the OC although they are contained in the same observation field, so images look clear.

Figure 4.1 shows *proper motion in right ascension and declination* for a sample of the downloaded dataset for Melotte 22. At first sight, two main clusters can be distinguished, one of them centered nearly at $[0, 0]$ and the second one with center at $[20, -45]$. This second cluster is the one we are looking for.

However, although the second cluster is almost isolated, there are stars that do not belong to the OC. Thus, we need more information to properly characterize the open cluster.

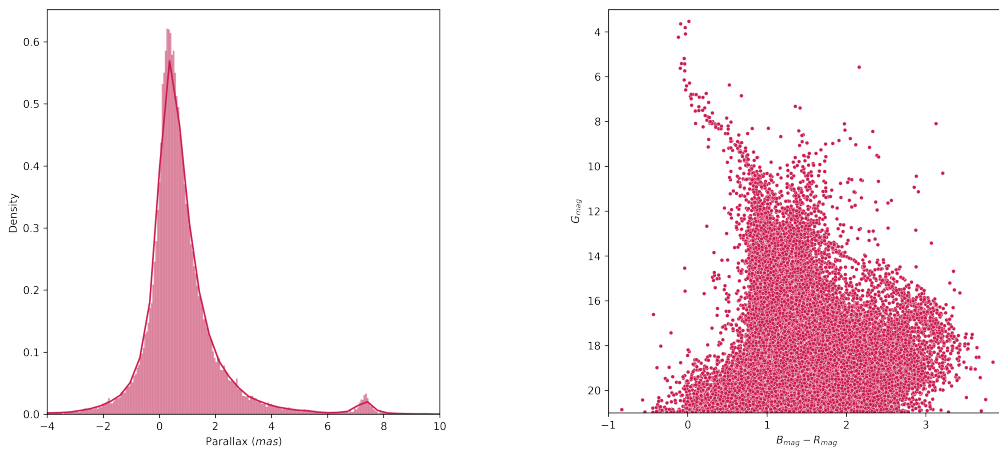


Figure 4.2: Melotte 22 parallax and H-R diagram

Figure 4.2 shows the parallax histogram and the H-R diagram for this cluster, respectively.

The figure on the left shows a resonance at $\approx 7.3mas$ belonging the OC. While the figure on the right would help us to look for isochrone curves and so, to determine the age of these stars.

4.1 Data Mining

Before being able to develop and test our clustering model, the first step is to download the required data from the Gaia repository and store it in a custom database for later access.

Due to the large amount of data available at Gaia, a complete download is not viable neither useful. In order to reduce the size of the dataset to be downloaded, the OpenClust

catalogue (Dias, Alessi, Moitinho, & Lépine, 2002) (Figure 4.3) has been used to restrict the sky regions to be explored.

This download is not limited to the size registered in the catalogue. Instead, a wider region (1.5 times the size of the cluster) is downloaded for each cluster to include stars outside the cluster. The idea is that the unsupervised model must be able to clusterize those stars that belong to the OC and to discard outsiders when characterizing the open cluster.

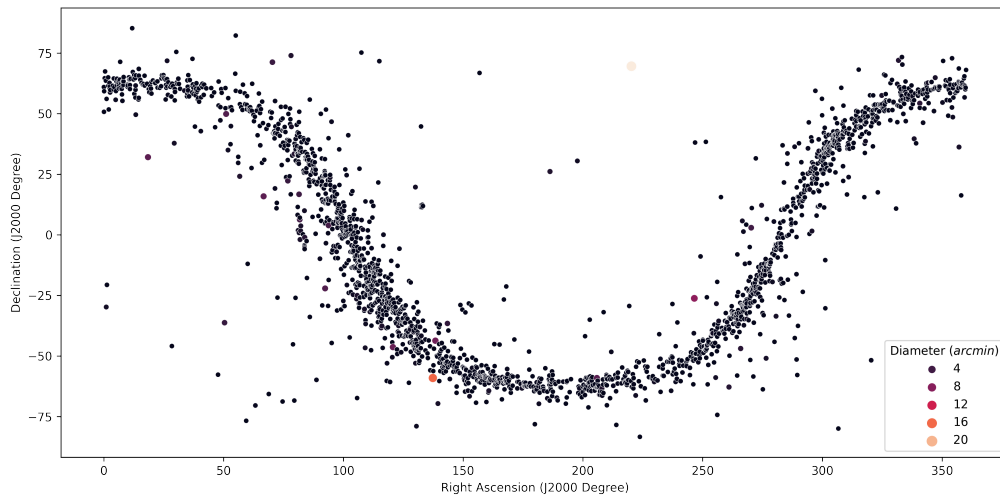


Figure 4.3: OpenClust Catalogue Distribution

Taking these considerations into account, the downloaded dataset covers nearly 114 million stars (around 42GB of compressed data). This dataset is significantly smaller than the whole Gaia DR2 dataset, which contains information for approximately 1,600 million stars. However, it is still too large.

Since not all the downloaded clusters are good enough, we will apply a series of filters to discard clusters which do not have enough stars or contain too many *null* values.

Therefore, a cluster must fulfil the following filters in order to be accepted:

- Cluster diameter above 25.0 arcmin
- Parallax absolute value greater than 0.0
- Number of stars¹ in the selected region above 40,000 stars

As shown in Figure 4.4, these constrains give us a smaller dataset but it is still a good

¹Every star must have all required features completely defined, i.e. without null values

representation of all clusters in the Milky Way since they are equally distributed around the galaxy disk.

After having applied these filters, the number of cluster to analyze is 169 with nearly 75 million stars.

Since we have set no upper limit to the number of stars inside a cluster, for the sake of simplicity and as a commitment to the project delivery dates and the available computing power, smaller clusters will be preferred over the greatest ones.

As an extension of this work, the designed model could be applied to a region not covered by these constraints for a later understanding of the results.

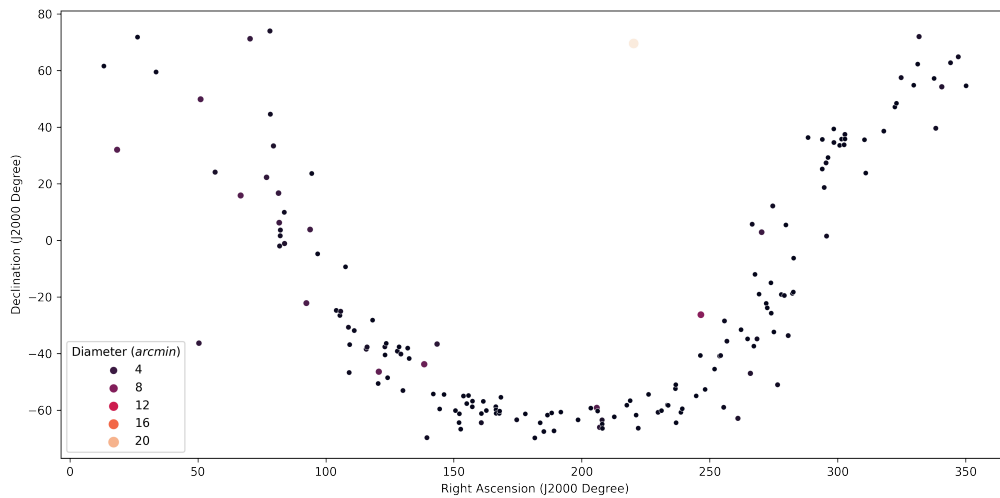


Figure 4.4: OpenClust Catalogue Selection Distribution

4.1.1 Download Process

The download process has been performed with two Docker containers: a *downloader* and the *database*.

The first one is a container built from a `python:3.8` image that contains the `cdalvaro` package and the `downloader.py` script. This script is prepared to load the OpenClust catalogue, connect to the Gaia DR2 database, download all stars contained inside each cluster (with a radius factor to increase the area to be downloaded) and save the downloaded data inside our custom database hosted in the second container.

`downloader.Dockerfile` contains the build instructions for the downloader image. We use [GitHub Actions](#) for automating the process for building a new image version with every new push made to the `main` branch.

This image can be pulled from the [GitHub Container Registry](#):

```
docker pull ghcr.io/cdalvaro/gaia-downloader:latest
```

The second container is just a `postgres:12.4` image which loads an initial script when the database is not yet initialized for creating the database schema. This container is used later as the main database for data analysis.

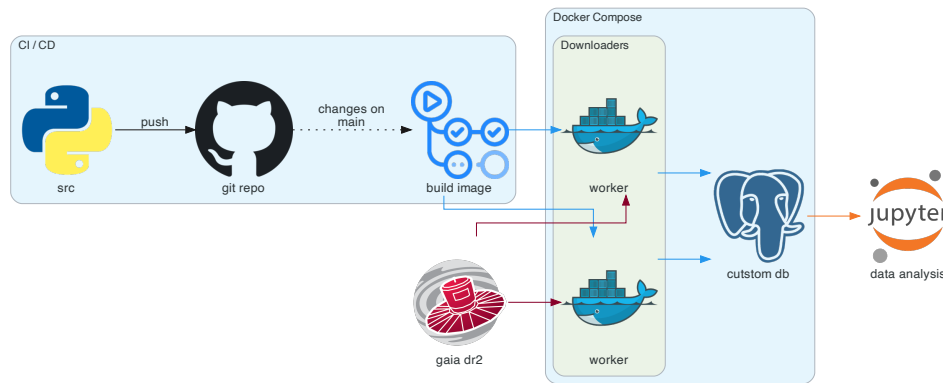


Figure 4.5: Data retrieval and analysis

These containers are orchestrated with a `docker-compose.yml` file which automatically launches one instance of each container. Figure 4.5 shows a diagram describing the services architecture for this work.

The database has two main tables in `public` schema:

- `public.regions` for storing cluster properties such as *location*, *diameter* and other properties
- `public.gaiadr2_source` which contains data for all downloaded stars from Gaia

The second table is partitioned by region, storing all stars related to one region inside an independent table for optimized access performance. Partition tables are located inside `gaiadr2` schema. See Figure 4.6.

Although `public.gaiadr2_source` is partitioned, there is no need for us to directly access to one of those tables. We only need to know the `id` for the desired region and query for its stars to the main table as follows:

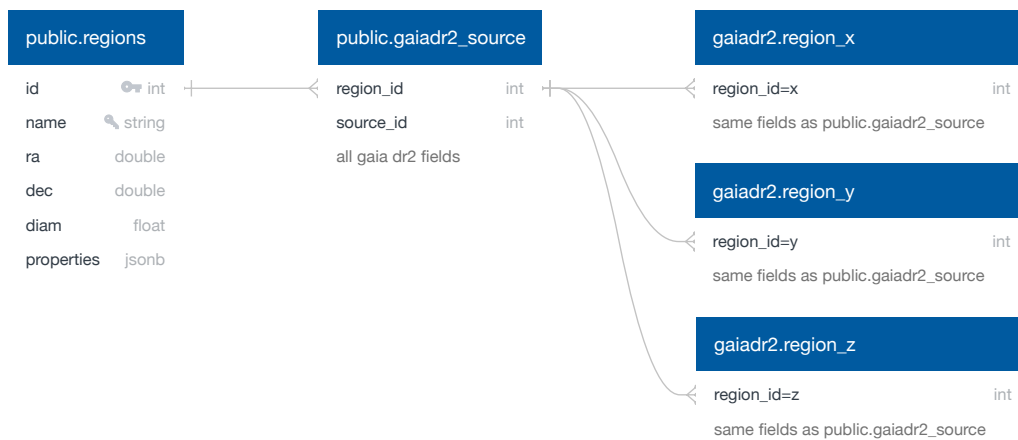


Figure 4.6: Custom DB diagram

```
-- Get region id
SELECT id FROM public.regions WHERE name = 'Melotte 22';
-- output: id -> 302

-- Get stars for the given id
-- NOTE: PostgreSQL cannot optimize this query when region_id is
--       provided as a subquery. This is why we get first the region id.
SELECT * FROM public.gaiadr2_source WHERE region_id = 302;
```

PostgreSQL knows that all stars we want are located inside `gaiadr2.region_302` so it does not need to query any other table. This allows us to download data for as many regions as we want without losing performance when accessing for that data.

`gaiadr2_source` and its partitions contain all fields available at the Gaia DR2 database. A complete list of these fields and their description can be found at the following link: https://gea.esac.esa.int/archive/documentation/GDR2/Gaia_archive/chap_datamodel/sec_dm_main_tables/ssec_dm_gaia_source.html

Table 4.1 shows a brief description of the fields used in this work.

The downloader takes data from the Gaia DR2 database and saves the selection into our custom database without modifying any value, so data preserves all its properties and units.

Python methods for retrieving information from the database are available through an instance of `cdalvaro.DB`. This class contains methods for retrieving star's information for a given region as a Pandas `DataFrame`, ready for data analysis.

Name	Units	Description
ra	Angle [deg]	Right ascension
dec	Angle [deg]	Declination
parallax	Angle [mas]	Parallax
pmra	Angular Velocity [mas/year]	Proper motion in right ascension
pmdec	Angular Velocity [mas/year]	Proper motion in declination
phot_g_mean_mag	Magnitude [mag]	G-band mean magnitude
bp_rp	Magnitude [mag]	BP - RP colour

Table 4.1: Properties and descriptions of the fields used in this work.

The developed code for this work can be found inside the `src` directory at GitHub: [cdalvaro/machine-learning-master-thesis](https://github.com/cdalvaro/machine-learning-master-thesis) repository. (For the printed version of this work, you can scan the QR code on cover for accessing it.)

4.2 Feature Selection

As mentioned before, we want our model to characterize open clusters by looking at their dynamic properties. Also, we want to maintain as simple as possible our clustering model in order to save computing resources.

Proper motion in right ascension and declination seems like a natural choice since, as we know, stars belonging to the same OC share a common motion vector.

Parallax is another important feature. It lets us know how far stars are from the Earth. In addition, since all stars within an open cluster were born from the same dust cloud, they must all have similar parallax.

However, we are not going to use these raw features. Instead, we are taking a combination of them.

First, we correct proper motion in right ascension and declination by dividing them by the parallax. That way, we normalize these quantities and help our clustering models to improve their performance.

The modulus of the proper motion is another computed property that we are considering. We use it to relate both features and therefore, to force our model to keep them tight.

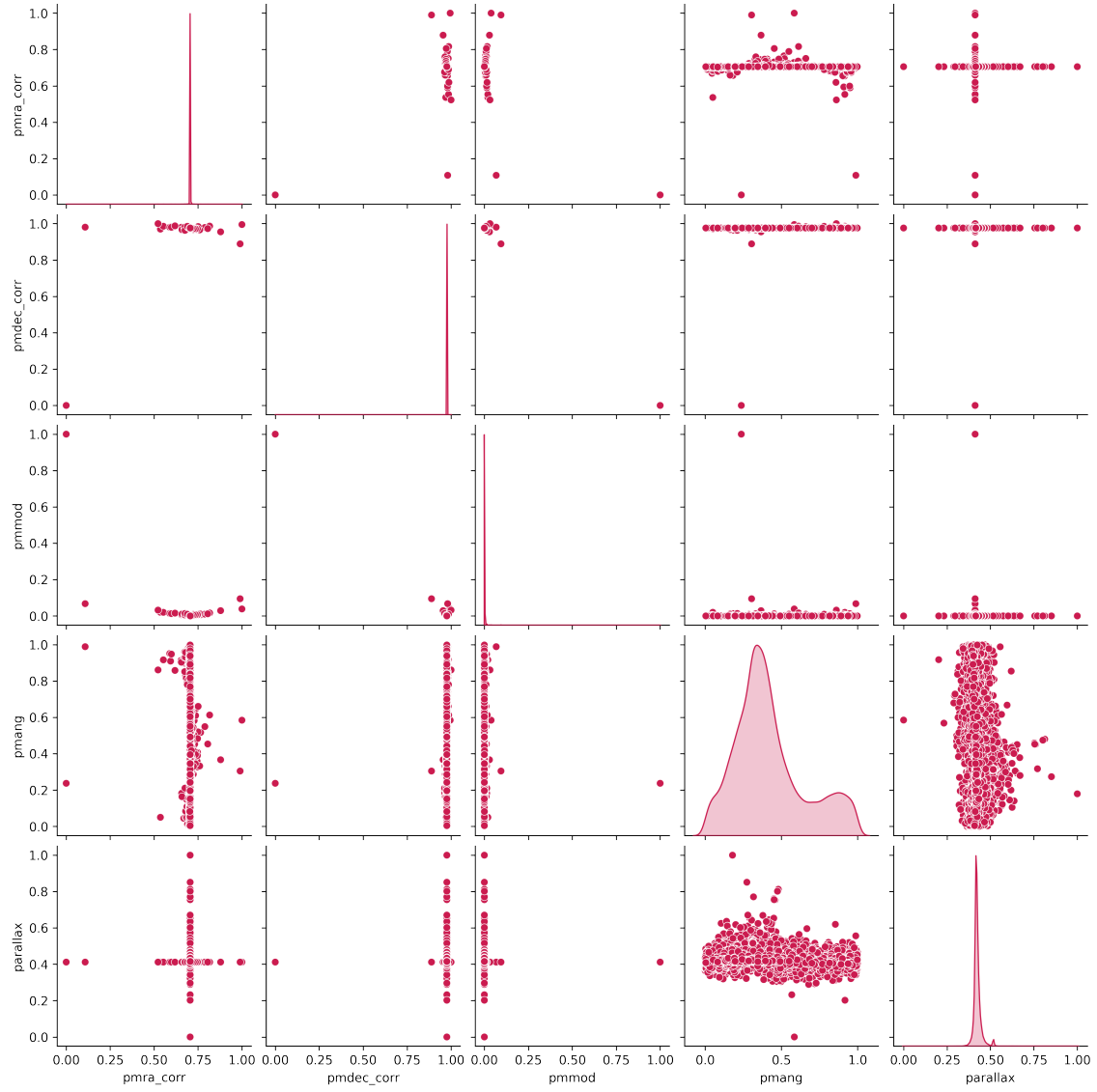


Figure 4.7: Pairwise relationships among variables using Melotte 22 data

Figure 4.7 shows a pairwise relationship among some features available in the dataset for Melotte 22 data. The diagonal of the grid shows the marginal distribution of the data in each column. This plot can be useful to find correlations between pairs of variables or to identify resonances in certain variables such as those in the parallax feature.

In summary, these are the features we are going to use as sources for our clustering models:

- *Proper motion in right ascension* (corrected by parallax): μ_α
- *Proper motion in declination* (corrected by parallax): μ_δ
- *Parallax*: ϖ
- *Proper motion modulus*: $\|\vec{\mu}\|$

This feature selection has been refined by iterating over the K-Means clustering process with a fixed number of clusters and varying the feature selection in each case. The final selection is the one with best *silhouette score* (Rousseeuw, 1987).

The silhouette score is a metric used to determine how good a cluster is based on intra-cluster distances and nearest-cluster distances. The best possible value is 1, and the worst is -1. Negative values mean that some samples have been assigned to the wrong cluster.

4.3 Soft Clustering with K-Means

Once we have found the set of features that best describes our problem, we can begin searching for the OC within the selected region.

Our first approach to find the open cluster is using the K-Means algorithm.

Since we are looking for a single cluster, it seems reasonable to use a clustering algorithm and set it to find two clusters. One for the desired OC and another which contains stars that do not belong to the open cluster. However, this idea is not completely right.

This is due to the fact that OC's stars are surrounded by other stars with possibly similar properties. So, setting the number of clusters to two is too low to separate them properly.

As shown in Figure 4.8, larger values for the number of clusters allow us to isolate more accurately the resonance in parallax at $\approx 7.3mas$. However, we have the disadvantage that more groups are formed.

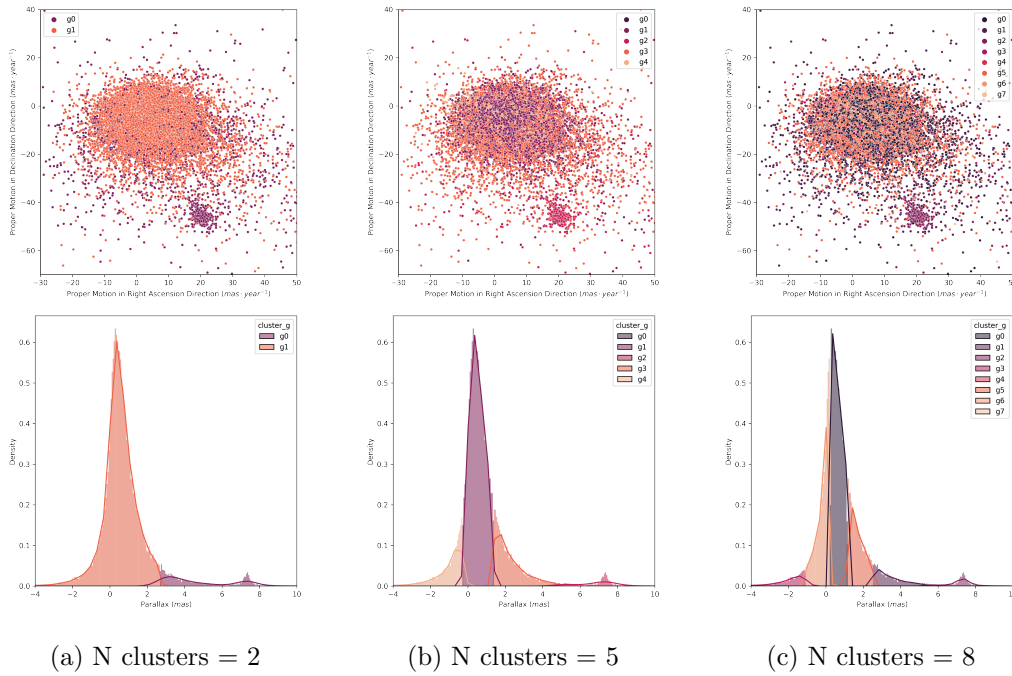


Figure 4.8: K-Means comparisons with Melotte 22

This effect complicates our task of finding the desired open cluster, since we would like to get just two groups, one for the OC and another with the remaining stars. Therefore, we have to find a way to set the right value for the number of clusters to isolate the searched cluster without creating too many groups.

To solve this issue, we will try to estimate the best number of clusters by using the *silhouette score* one more time.

The following example is a snippet copied from `cluster_characterization.ipynb` Jupyter notebook (available at [src/notebooks](#)). It shows how to estimate the best number of clusters for Melotte 22 by using `cdalvaro.ml.estimate_n_clusters` method.

```
n_clusters, kmeans = estimate_n_clusters(melotte22_df,
                                         min_clusters=3, max_clusters=7)

# Silhouette score for 3 clusters: 0.5420
# Silhouette score for 4 clusters: 0.5393
# Silhouette score for 5 clusters: 0.5608
# Silhouette score for 6 clusters: 0.5336
# Silhouette score for 7 clusters: 0.5306
# Best silhouette score is 0.5608 for 5 clusters
```


K-Means does a good job making an initial clustering, as shown in Figure 4.9. However, too many clusters arise from this characterization and the OC is still polluted with stars that do not belong to it. Moreover, we would like to reduce the amount of clusters too.

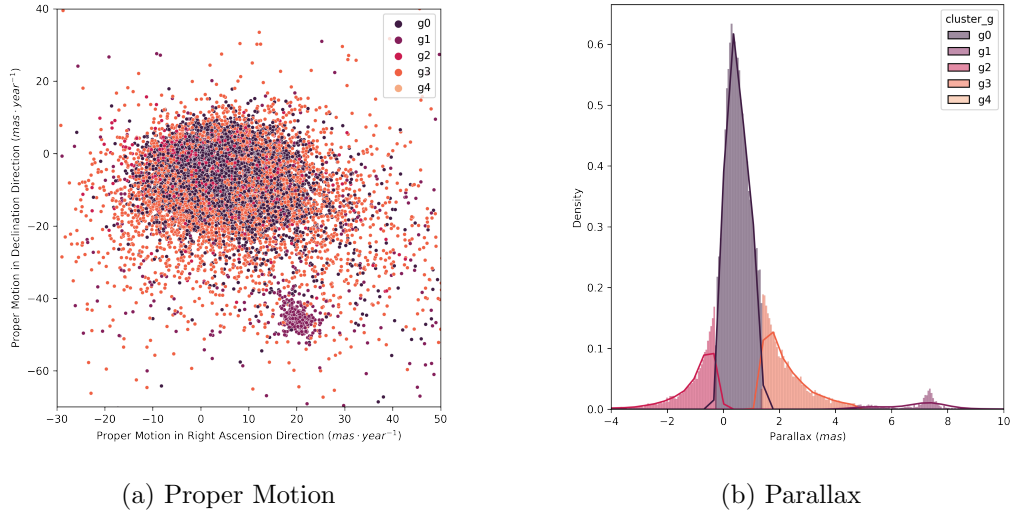


Figure 4.9: K-Means model applied to Melotte 22

If we take a look into the H-R diagram (Figure 4.10), we can identify the isochrone curve for Melotte 22 cluster. This curve has a good shape, but again, it contains outsider stars. Therefore, we would like to find a better model that improves this initial characterization by reducing the amount of clusters and also that removes outsiders from the OC.

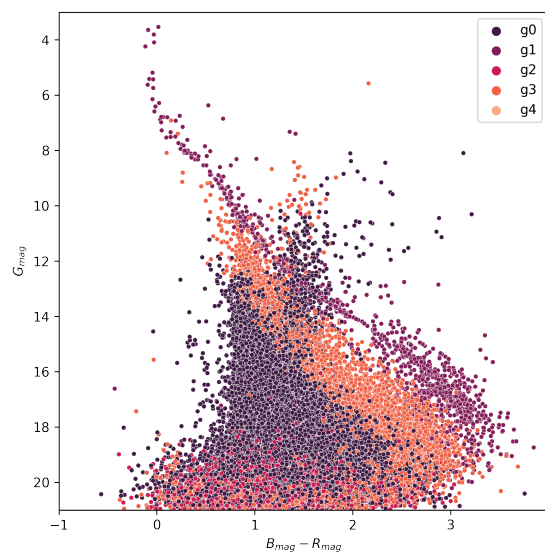


Figure 4.10: Melotte 22 H-R diagram with K-Means characterization

4.4 Deep Embedded Clustering (DEC)

Our initial approach allows us to find some groups that potentially contain the desired OC. However, as we have seen before, the result is not accurate enough. Therefore, we need a method for merging clusters by migrating stars from one group to another, so we can end up with the minimum number of clusters while preserving the open cluster.

Since we do not have a labeled dataset to train a supervised model, we have no choice but to use an unsupervised self-trained model.

For that reason we have adapted the *Unsupervised Deep Embedding for Clustering Analysis* to our work.

The implementation of this model is available in `cdalvaro.ml.DEC` and it is developed with the Keras framework.

The model is composed by a *deep autoencoder* and a *clustering layer*.

The autoencoder is used to transform the input data into a latent space using a non-linear mapping function $f_\theta : X \rightarrow Z$.

Although, as explained in Section 4.2, the number of features we are managing is not too large, this latent space helps us reduce the number of features and avoids the “*curse of dimensionality*” (Bellman, 1961).

The autoencoder is pretrained before fitting the model to generate predictions. Then, the encoder layers of the autoencoder are used with the aim of transforming input data to the latent space Z . Once the data has been transformed, a K-Means clusterer is used in order to make an initial clustering. K-Means cluster centers are used as the initial weights for the clustering layer.

With that initial configuration, the model iterates alternating between computing an auxiliary target distribution (Soft Assignment) and minimizing the Kullback-Leiber (KL) divergence (Kullback & Leibler, 1951) to it. This unsupervised algorithm allows us to improve the clustering.

$$p_{ij} = \frac{q_{ij}^2 / f_j}{\sum_{j'} q_{ij'}^2 / f_{j'}} \quad (4.1)$$

In the soft assignment stage, the *Student’s t-distribution* is used as a kernel to measure the similarity between the embedded points and the cluster centroid. While in the KL divergence minimization the algorithm iteratively refines clusters by learning from their high confidence assignments with the help of an auxiliary target distribution. The model

is trained by matching the soft assignment to the target distribution. The choice of this target distribution is crucial for DEC's performance. In this work we have taken the target distribution from DEC's original paper (Xie et al., 2016), which is defined in Equation 4.1.

Figure 4.11 shows the layer setup of our DEC model. It is simpler than the one tested on the original paper (Xie et al., 2016), since the number of selected features in our work is smaller than in the original one. Therefore, using the same configuration would result in a model so powerful that would incur in overfitting issues unable to make right predictions.

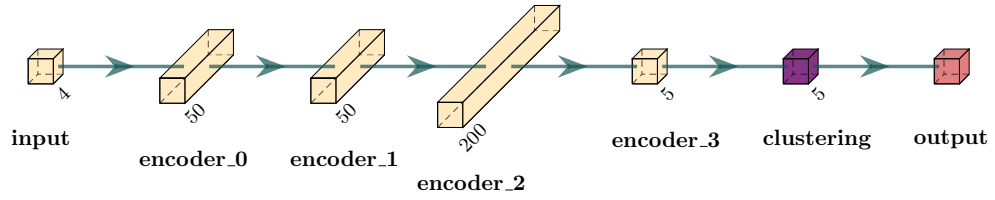


Figure 4.11: DEC model layer setup

Once the model is ready, we can test it by applying it to the Melotte 22 dataset. Notebook `cluster_characterization.ipynb` illustrates how DEC model processes a region of stars trying to characterize the open cluster hidden within that region.

As shown in Figure 4.12, the DEC model is able to move stars from one group to another, removing some clusters and improving the OC characterization. This is exactly what we are looking for.

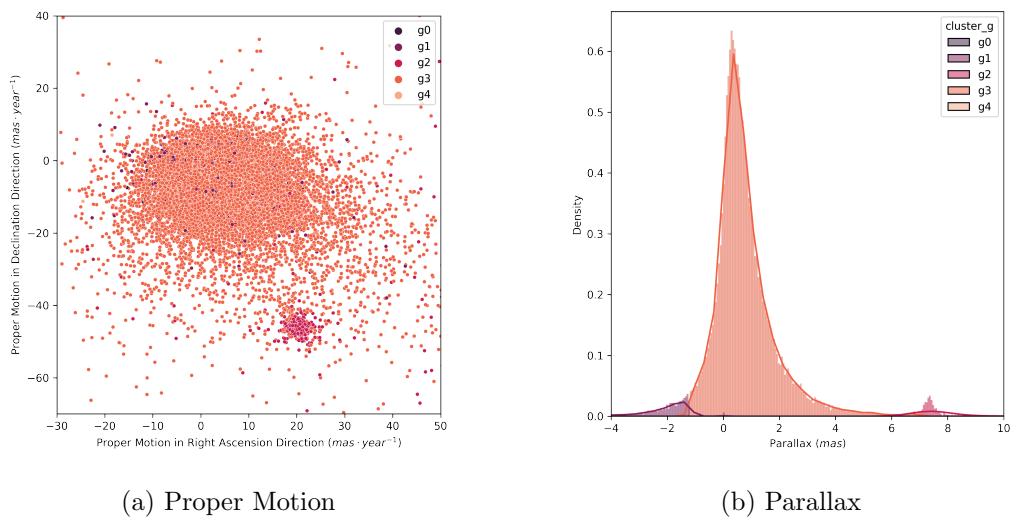


Figure 4.12: DEC model applied to Melotte 22

The H-R diagram has been improved as well, and the isochrone curve of the open

cluster is sharper than the one obtained with the K-Means algorithm (Figure 4.10).

Figure 4.13 shows a comparison between the H-R diagram obtained by DEC clustering and the one obtained using the Clusterix + TOPCAT method.

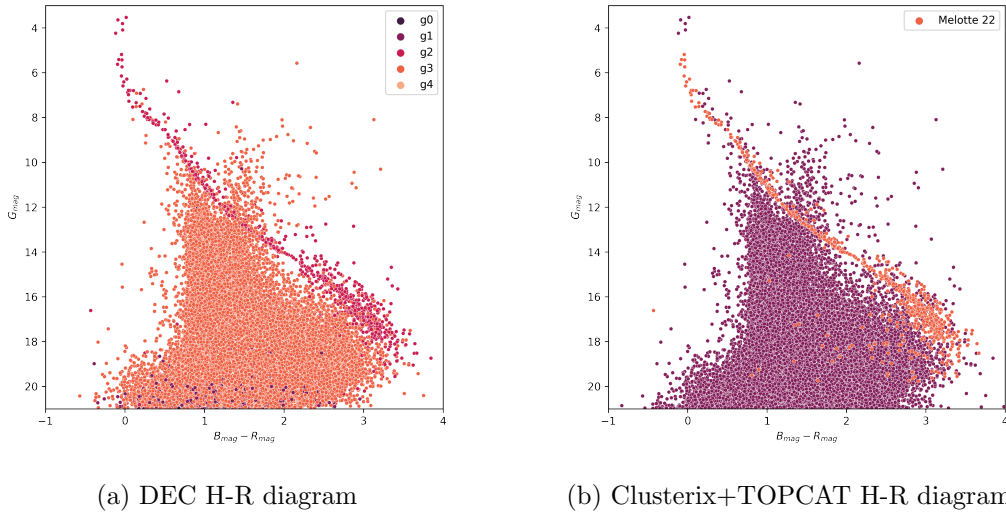


Figure 4.13: H-R diagram comparison between DEC clustering and Clusterix+TOPCAT method.

Finally, we can refine this selection by filtering those stars which are below and above the 0.10 and 0.90 quantiles for each group, respectively. That way we remove the most doubtful values from the selection. Figure 4.14 shows the result after filtering DEC clustering selection.

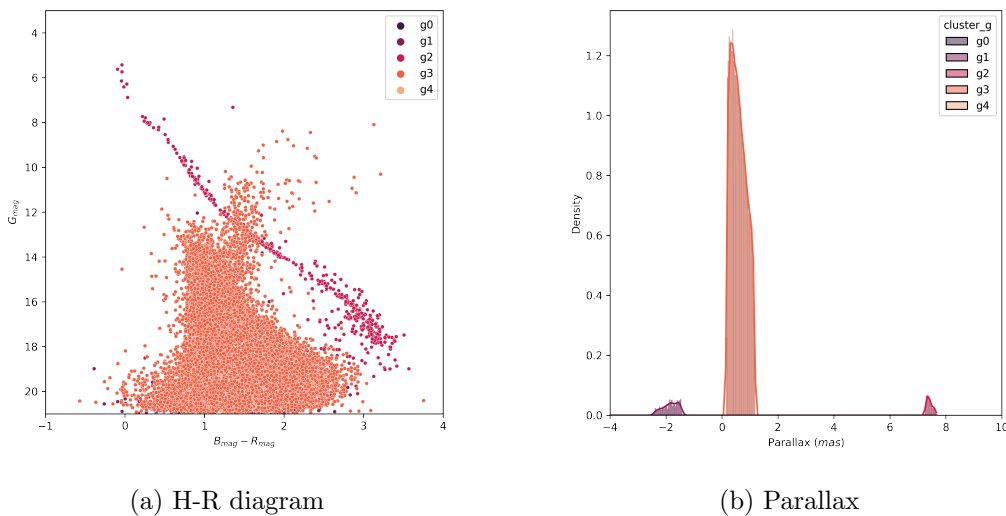


Figure 4.14: Melotte 22 H-R diagram and parallax histogram with stars outside 0.25 and 0.75 quantiles filtered.

Table 4.2 shows a summary of Melotte 22 with values taken from different sources and methods: SIMBAD Astronomical Database - CDS (Strasbourg), Clusterix+TOPCAT method, and the results obtained for each step of this method (K-Means, DEC and DEC filtered). Errors associated to K-Means and DEC predictions have been calculated as the *Standard Error*, $SE = \sigma/\sqrt{N}$, where N is the number of samples in the selected group.

Source / method	μ_α ($mas \cdot yr^{-1}$)	μ_δ ($mas \cdot yr^{-1}$)	ϖ (mas)	# stars
Simbad.u-strasbg.fr²	19.997 ± 0.127	-45.548 ± 0.101	7.364 ± 0.005	1326
Clusterix+TOPCAT	19.98 ± 1.25	-45.47 ± 1.48	7.33 ± 0.21	634
K-Means	20.25 ± 0.95	-38.01 ± 1.08	7.23 ± 0.06	1378
DEC	23.67 ± 1.29	-46.23 ± 1.50	8.04 ± 0.09	878
DEC (filtered)	19.50 ± 0.41	-44.23 ± 0.39	7.42 ± 0.005	438

Table 4.2: Parameters shown are proper motion in right ascension and declination, parallax with their respective deviations and number of stars corresponding to Melotte 22 data.

²Results have been taken from Babusiaux et al. (2018, p. 25. Table A.3. Pleiades)

Chapter 5

Results

We have seen in Section 4.4 that DEC model is capable of refining the clustering selection made by K-Means hence helping us have a more precise characterization of the cluster. Now, in order to validate our model and to be able to estimate how good or bad it is, we need a different technique to characterize the studied clusters and compare our results with those obtained with the validation method.

For that purpose, we have used an alternative method based on Virtual Observatory (VO) tools such as Clusterix and TOPCAT, to characterize the studied clusters.

5.1 Cluster Characterization with VO Tools

The aim of this section is to highlight the procedure differences between our method and the validation one, as well as using the results provided by the second technique to validate ours results. This validation method uses a combination of VO tools: *Clusterix 2.0* + *TOPCAT*. We take Melotte 22 again as our study case to make easier the comparison process between the method presented in Chapter 4 and the alternative one.

The initial point is taking advantage of Clusterix in order to avoid a subjective and manual selection for the cluster members. Figure 5.1 shows a screenshot of the Clusterix 2.0 selection panel.

This panel allows us to select a region by its coordinates or by searching a cluster by its name in the Gaia DR2 catalogue, as well as to specify the size of the region by introducing its radius. We can also set limit values in magnitude (for both minimum and maximum) to discard stars outside these limits. Another setting is the flag to enable or disable the *Q-Filter*. This filter discards those stars with *null* values in relevant fields.

Step 1/3: Information gathering (coordinates and physical parameters)

The image shows the Clusterix 2.0 selection panel with four search options:

- Search by Id** (pink panel): ID input field, Radius input field with a unit dropdown (set to arcmin), Catalogue dropdown (set to GAIA/DR2), a checked checkbox for "Q-Filter?", and Magnitude limits (min/max) input fields with a filter dropdown (set to G).
- Search by Coordinates** (green panel): RAJ2000(deg),DEJ2000(deg) input field with the value 56.75,24.1167, Radius input field with a unit dropdown (set to arcmin), Catalogue dropdown (set to GAIA/DR2), a checked checkbox for "Q-Filter?", and Magnitude limits (min/max) input fields with a filter dropdown (set to G). A "Search" button is located below this panel.
- Search in Webda ?** (purple panel): A single input field with a unit dropdown.
- Search by file ?** (yellow panel): A "Choose File" button and the text "no file selected".

Below the search panels is a checkbox labeled "Membership from proper motions".

Figure 5.1: Clusterix 2.0 selection panel with Melotte 22 settings. Coordinates correspond to the center of Melotte 22 cluster, while the radius is the value registered for this cluster in the OpenClust catalogue multiplied by 1.5 (as explained in 4.1)

Clusterix returns a position map for the downloaded stars from Gaia DR2 database using Aladin (Bonnarel et al., 2000) as an atlas of the sky (See Figure 5.2). At this point, we can choose sending the recovered data to TOPCAT or Aladin, or continue to the next step of the assistant.

We select the second choice and proceed with the proper motion frequency space analysis.

In this step, we define three regions that will be used for the frequency analysis:

- **Inner region** ($c+f$), with stars that belong to the OC and field stars.
- **External region** or field region (f), with stars that do not belong the open cluster.
- **Void region**. This region is excluded from the analysis since it is supposed to be a transition region with stars that could belong the OC.

It is necessary to remark the importance of this step, since the selection of these regions is critical to get a good result. However, despite how crucial this step is, there is no general rule to define these regions. We only have two references: the radius of the downloaded region and the estimated size for the OC that sets the *inner radius*. The other two radius

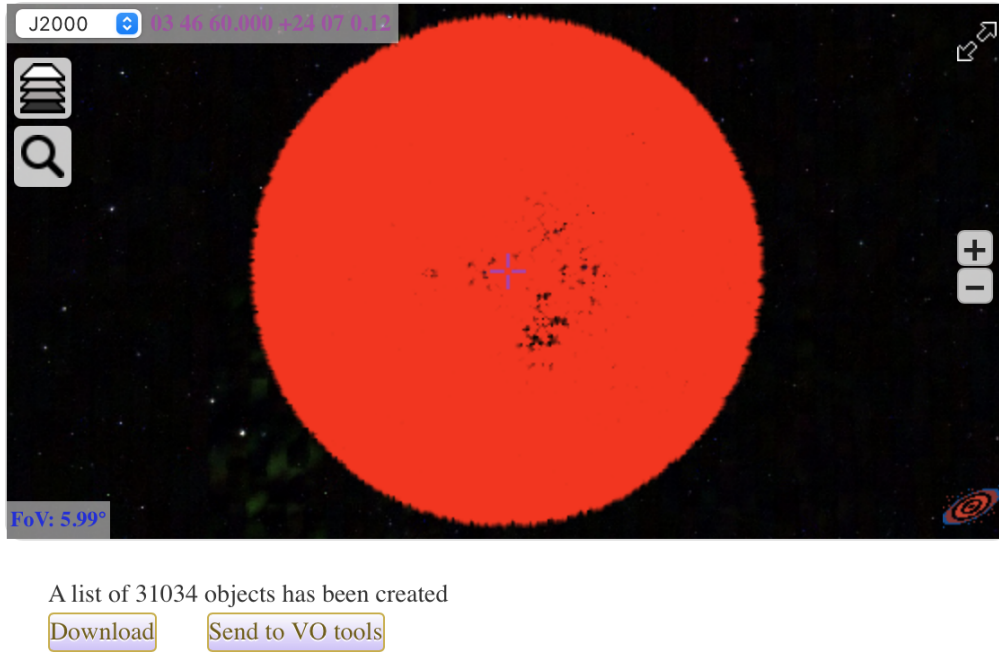


Figure 5.2: Aladin screenshot with equatorial coordinates for Melotte 22 stars referred to J2000.

must be set by trial and error. In general, they should be large enough to be sure that we are not taking polluted regions. This is a tedious process and requires several tries before achieving good enough probability values.

Furthermore, it is necessary to know in advance some additional profiles of the cluster, in addition to the estimated size. For example, in the case of Melotte 22, the proper motion in declination direction of the membership stars around $-45 \text{ mas} \cdot \text{year}^{-1}$, so it is necessary to open the μ range above the estimated values (in absolute value). If we do not do that we lose the possibility of detection, since Clusterix sets this value to $30 \text{ mas} \cdot \text{year}^{-1}$ by default. But even in this case, it is necessary to set a threshold for magnitude values, $G \leq 14$, to help getting the right result. Figure 5.3 shows the region selection panel on Clusterix 2.0.

This makes the selection process an art rather than a science, not to mention the required previous knowledge on certain profiles of the objects to characterize. This puts into question the *non-parameterized* character of the validation method.

After having set all these parameters, Clusterix performs the proper motion frequency analysis and returns the probability field associated to each star of the whole region (see Figure 5.4).

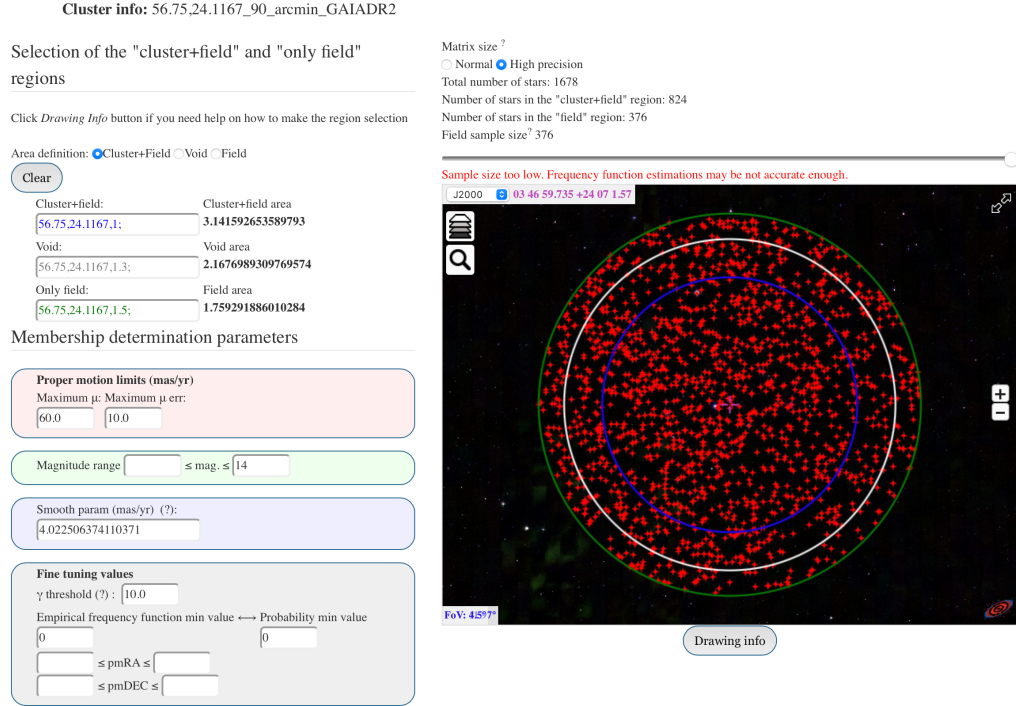


Figure 5.3: Clusterix 2.0 Step 2/3: Region selection.

This step can be repeated over and over again, varying radii and other parameters, until finally reach a valid probability that works. The last step presents a table with candidate objects and allows to export this dataset to TOPCAT, Aladin and VOSA to perform a deeper filtering and analysis based on the results provided by Clusterix.

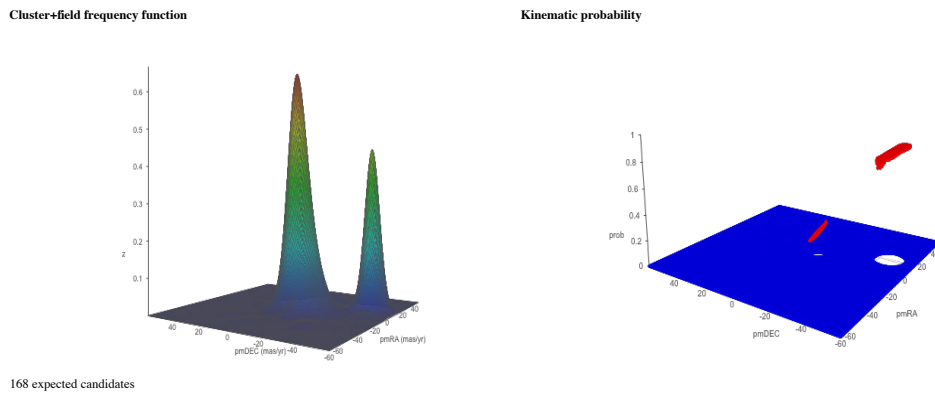


Figure 5.4: Kinematic probability associated to the membership of the open cluster. This probability is derived from the proper motion frequency analysis.

Once data is available in TOPCAT, we have to carefully analyze the imported data. At this point, we can perform a first star selection by excluding those stars with probability value lower than 0.8. This first result can be taken as a good starting point, but we still

need to make some fine tuning. As we can see in Figure 5.5, the selection is too restrictive and we have lost too many stars.

Looking at Figure 5.5b, we can assume that cluster members are around $7.3mas$ in the parallax histogram. Also, Figure 5.5a reveals that cluster stars have low scattering rate in the proper motion configuration space.

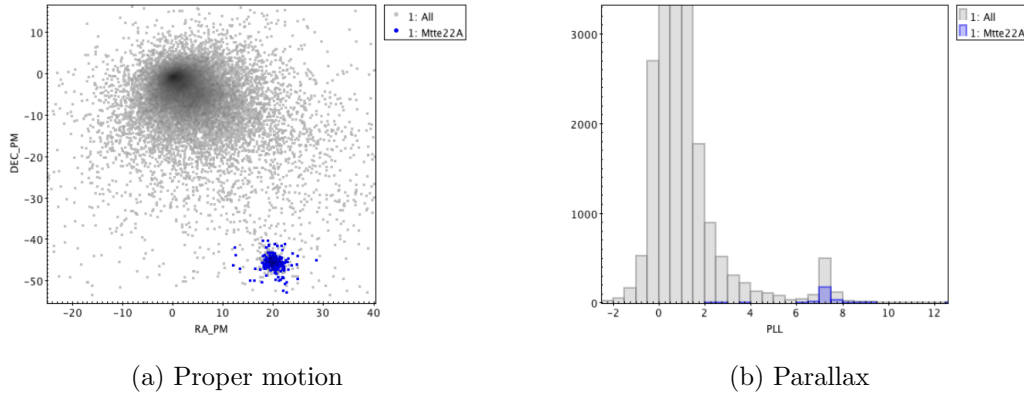


Figure 5.5: Melotte 22 first selection using validation method. Parallax is $7.304mas$ with 0.797 of standard deviation.

Taking these observations into account, we can proceed to make a new selection based on parallax properties. Our second selection makes use of the values obtained with the previous selection for parallax and its standard deviation and selects those stars from the whole set which fall inside the limits $7.304mas \pm 0.797$ (see Figure 5.6).

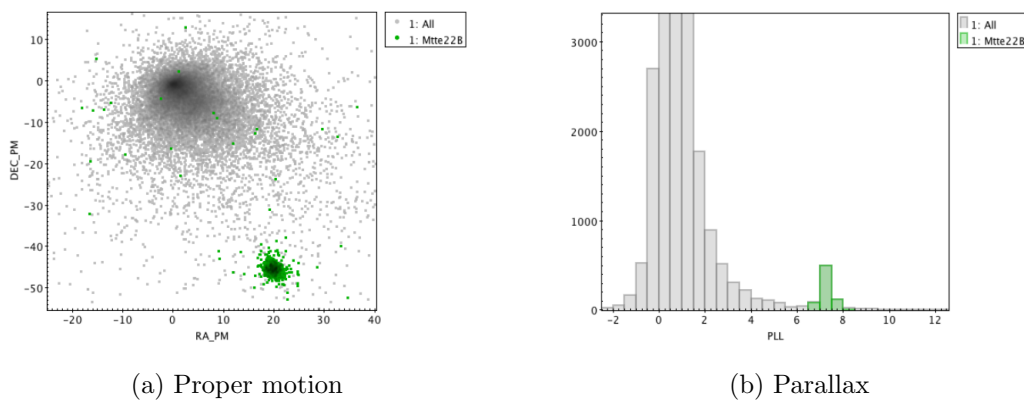
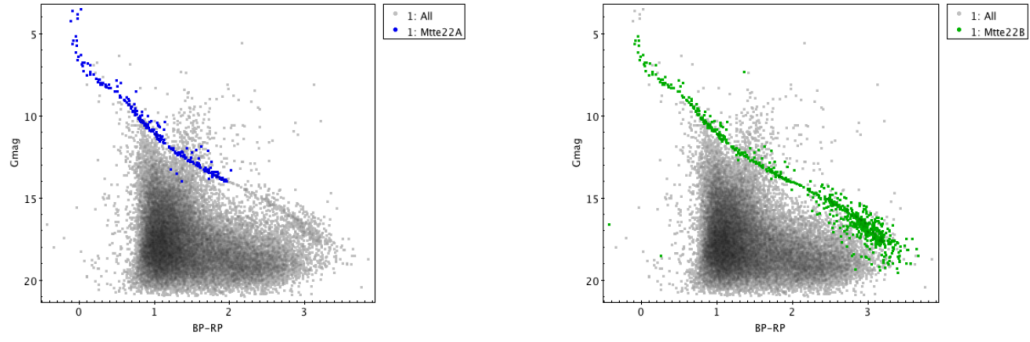


Figure 5.6: Melotte 22 second selection. The estimated number of cluster members is 709.

Looking at the H-R diagrams (Figure 5.7), we can compare both selections and see how we have extended the selection without getting too much noise around the isochrone curve.

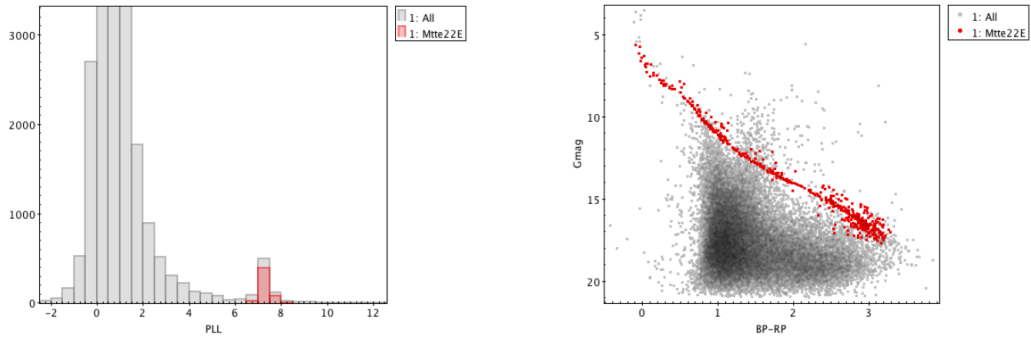


(a) H-R diagram from first selection.

(b) H-R diagram from second selection.

Figure 5.7: Comparison between first and second selection H-R diagrams.

Finally, we filter the second selection by discarding those stars whose parallax relative error in absolute value is behind 0.05 (see Figure 5.8).



(a) Parallax

(b) H-R diagram

Figure 5.8: Comparison between first and second selection H-R diagrams.

For the sake of completeness, Table 4.2 is shown again. This table summarizes the results obtained with the validation technique and our method.

Source / method	$\mu_{\alpha} \text{ (mas} \cdot \text{yr}^{-1}\text{)}$	$\mu_{\delta} \text{ (mas} \cdot \text{yr}^{-1}\text{)}$	$\varpi \text{ (mas)}$	# stars
Simbad.u-strasbg.fr	19.997 ± 0.127	-45.548 ± 0.101	7.364 ± 0.005	1326
Clusterix+TOPCAT	19.98 ± 1.25	-45.47 ± 1.48	7.33 ± 0.21	634
K-Means	20.25 ± 0.95	-38.01 ± 1.08	7.23 ± 0.06	1378
DEC	23.67 ± 1.29	-46.23 ± 1.50	8.04 ± 0.09	878
DEC (filtered)	19.50 ± 0.41	-44.23 ± 0.39	7.42 ± 0.005	438

5.2 Comparing and Validating

Until now, we have focused our study in the open cluster Melotte 22. Since our aim is to get a model that fits well for a wide range of clusters, we will now proceed to show some results obtained for a selection of clusters with different typologies.

All results have been computed with an Apple Mac Pro Late 2013, 2.7GHz 12-Core Intel Xeon E5-2697v2, 64GB RAM 1866MHz DDR3 and two AMD FirePro D700 6GB. The operative system is macOS BigSur 11.1, with Python 3.8, Keras 2.2 and PlaidML 0.7.

Most executions took a few minutes, while large clusters such as Melotte 25, took around an hour to perform the calculations. This is in contrast to [Castro-Ginard et al. \(2020\)](#) work, that used the power of the MareNostrum 42 supercomputer to accomplish its computations.

5.2.1 NGC 2516

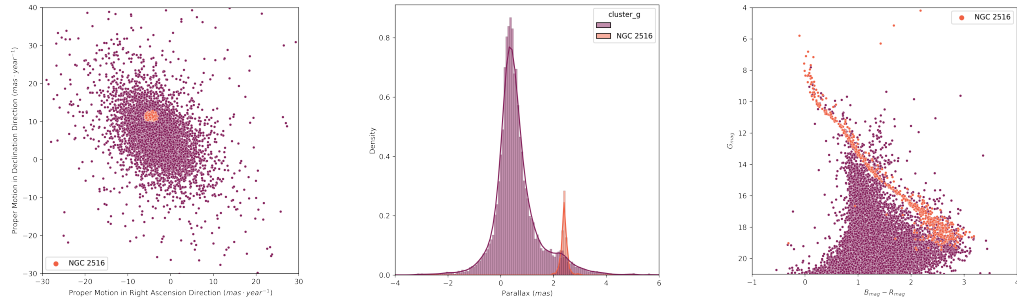


Figure 5.9: NGC 2516 Clusterix+TOPCAT characterization.

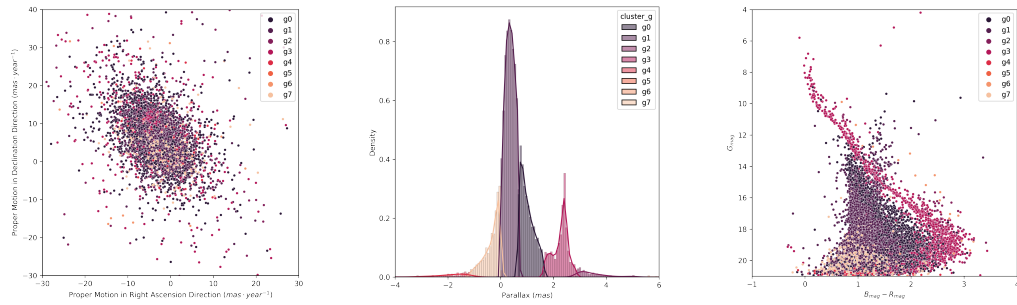


Figure 5.10: NGC 2516 K-Means characterization. Identified as cluster $g3$.

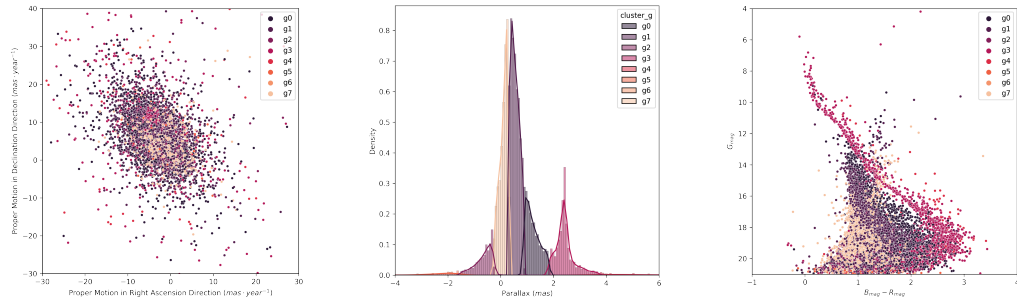


Figure 5.11: NGC 2516 DEC characterization. Identified as cluster $g3$.

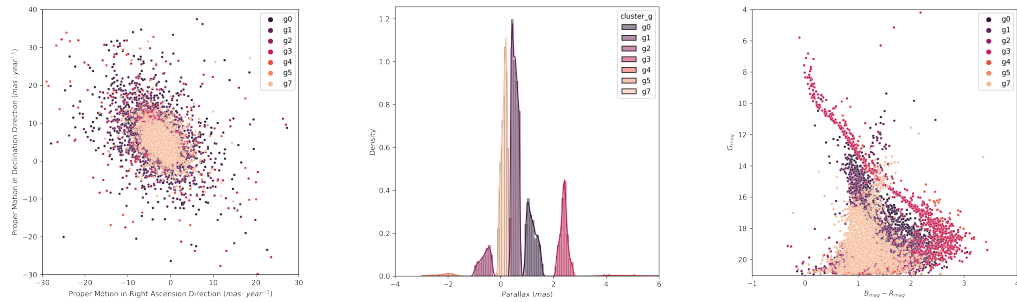


Figure 5.12: NGC 2516 DEC (filtered) characterization. Identified as cluster $g3$.

Figure 5.9 shows NGC 2516 characterized using Clusterix+TOPCAT tools. Figure 5.10 shows eight clusters identified by K-Means. In this case, cluster $g3$ is the one we are looking for. Figures 5.11 and 5.12 show the groups found using the DEC model and the DEC model filtered, respectively. Again, the cluster of interest is the group $g3$. Although in general, groups between K-Means model and DEC models may not match.

Table 5.1 shows the hyperparameters used for characterizing NGC 2516 with the DEC model. Table 5.2 shows a results summary for NGC 2516.

Hyperparameter	Value
Number of Clusters	8
Clustering Layer	[50, 50, 60]
Kernel Initializer Seed	2
Quantil Threshold	0.15

Table 5.1: NGC 2516 DEC model hyperparameters.

Source / method	μ_α ($mas \cdot yr^{-1}$)	μ_δ ($mas \cdot yr^{-1}$)	ϖ (mas)	# stars
Simbad.u-strasbg.fr	-4.6579 ± 0.0075	11.1517 ± 0.0075	2.4118 ± 0.0006	1727
Clusterix+TOPCAT	-4.652 ± 0.523	11.203 ± 0.454	2.409 ± 0.127	638
K-Means	-4.344 ± 0.14	9.507 ± 0.19	2.268 ± 0.01	1542
DEC	-4.426 ± 0.17	9.952 ± 0.20	2.436 ± 0.01	1532
DEC (filtered)	-4.502 ± 0.14	10.114 ± 0.17	2.392 ± 0.004	1072

Table 5.2: NGC 2516 results.

5.2.2 NGC 2632

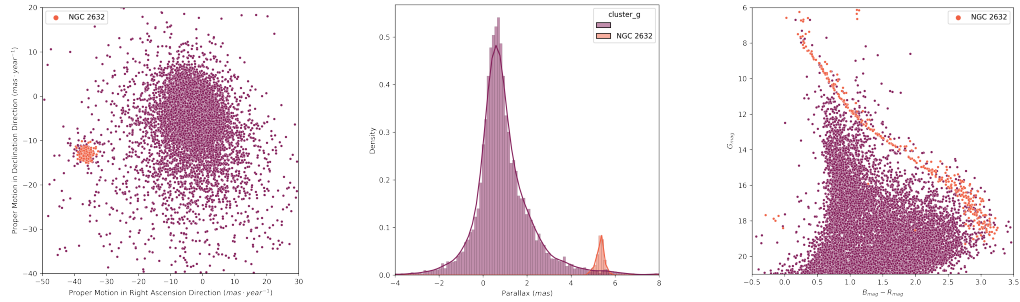


Figure 5.13: NGC 2632 Clusterix+TOPCAT characterization.

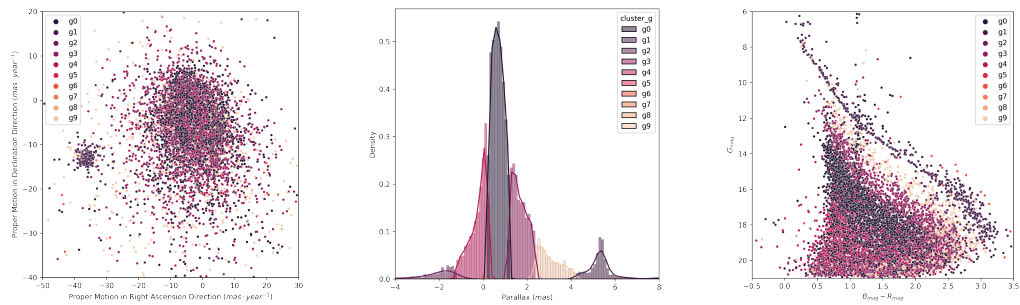


Figure 5.14: NGC 2632 K-Means characterization. Identified as cluster $g1$.

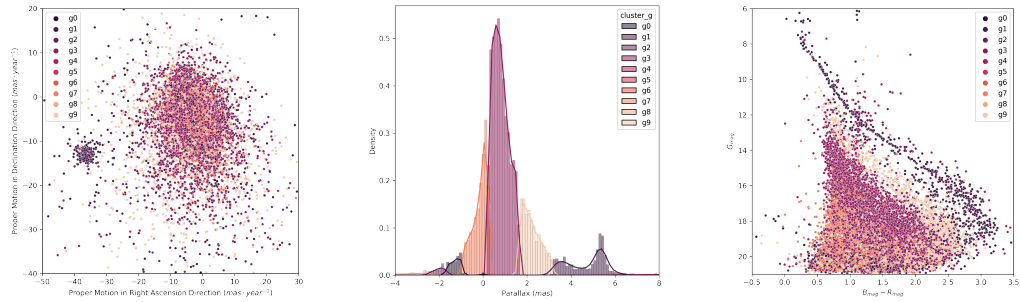


Figure 5.15: NGC 2632 DEC characterization. Identified as cluster $g1$.

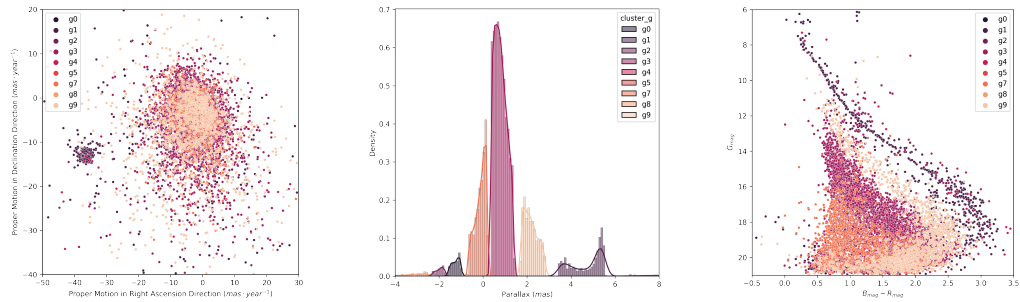


Figure 5.16: NGC 2632 DEC (filtered) characterization. Identified as cluster $g1$.

Figure 5.13 shows NGC 2632 characterized by the validation method. Figure 5.14 shows ten clusters identified by K-Means. Cluster *g1* is the open cluster we are looking for. Figures 5.15 and 5.16 show the groups found using the DEC model and the DEC model filtered, respectively. Open cluster NGC 2632 is labeled as group *g1*.

Table 5.3 shows the hyperparameters used for characterizing NGC 2632 with our model. Table 5.4 shows a results summary for NGC 2632 analysis.

Hyperparameter	Value
Number of Clusters	10
Clustering Layer	[50, 50, 40]
Kernel Initializer Seed	10
Quantil Threshold	0.1

Table 5.3: NGC 2632 DEC model hyperparameters.

Source / method	μ_α ($mas \cdot yr^{-1}$)	μ_δ ($mas \cdot yr^{-1}$)	ϖ (mas)	# stars
Simbad.u-strasbg.fr	-36.047 ± 0.110	-12.917 ± 0.066	5.371 ± 0.003	-
Clusterix+TOPCAT	-36.154 ± 1.001	-12.909 ± 0.806	5.327 ± 0.187	371
K-Means	-26.352 ± 0.82	-15.828 ± 0.76	5.394 ± 0.03	629
DEC	-20.012 ± 0.69	-14.742 ± 0.58	4.686 ± 0.03	894
DEC (filtered)	-21.571 ± 0.74	-14.234 ± 0.61	4.719 ± 0.03	714

Table 5.4: NGC 2632 results.

5.2.3 NGC 2682

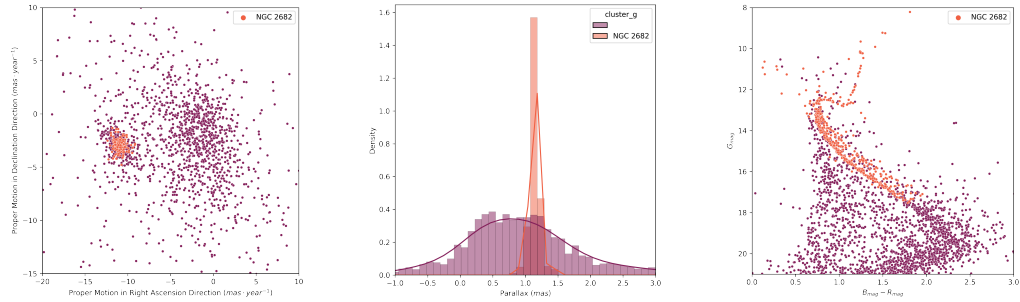


Figure 5.17: NGC 2682 Clusterix+TOPCAT characterization.

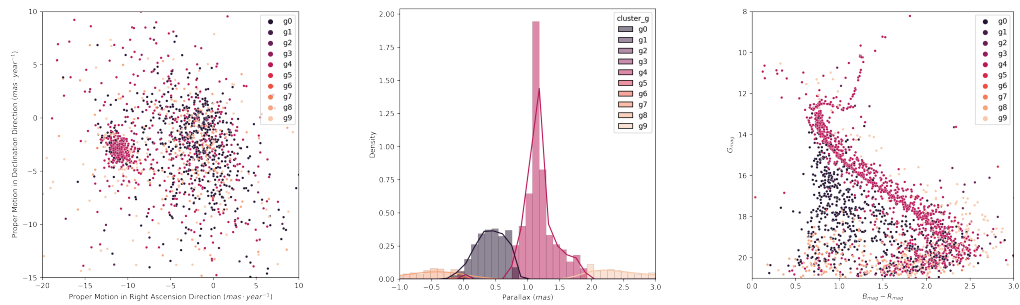


Figure 5.18: NGC 2682 K-Means characterization. Identified as cluster g_4 .

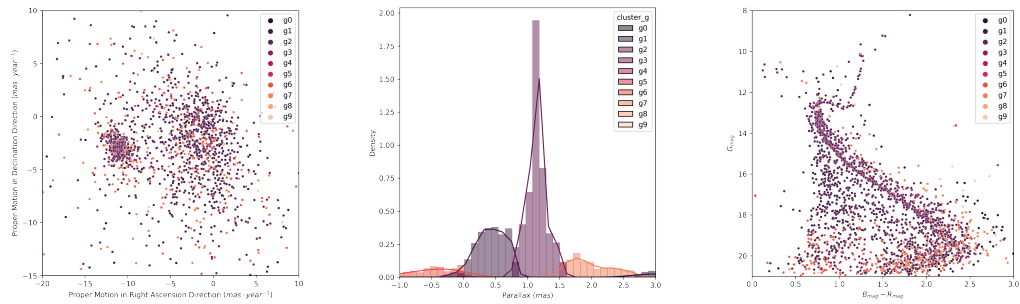


Figure 5.19: NGC 2682 DEC characterization. Identified as cluster g_2 .

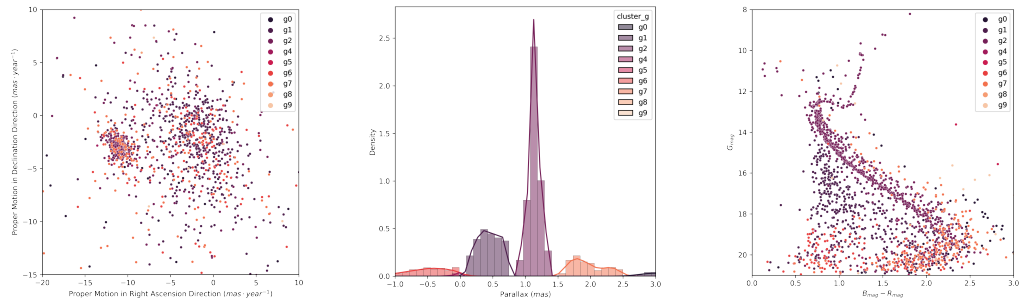


Figure 5.20: NGC 2682 DEC (filtered) characterization. Identified as cluster g_2 .

Figure 5.17 shows NGC 2682 characterized using Clusterix+TOPCAT tools. Figure 5.18 shows ten clusters identified by K-Means. K-Means labels NGC 2682 as group $g4$. Figures 5.19 and 5.20 show the groups found using the DEC model and the DEC model filtered, respectively. The cluster of interest is the group $g2$.

Table 5.5 shows the hyperparameters used for characterizing NGC 2682 with the DEC model. Table 5.6 shows a results summary for NGC 2682.

Hyperparameter	Value
Number of Clusters	10
Clustering Layer	[50, 50, 40]
Kernel Initializer Seed	0
Quantil Threshold	0.1

Table 5.5: NGC 2682 DEC model hyperparameters.

Source / method	μ_α ($mas \cdot yr^{-1}$)	μ_δ ($mas \cdot yr^{-1}$)	ϖ (mas)	# stars
Simbad.u-strasbg.fr	-10.9737 ± 0.0064	-2.9396 ± 0.0063	1.1325 ± 0.0011	1194
Clusterix+TOPCAT	-10.970 ± 0.322	-2.958 ± 0.327	1.142 ± 0.080	649
K-Means	-8.616 ± 0.15	-3.710 ± 0.16	1.196 ± 0.01	1374
DEC	-8.926 ± 0.15	-3.550 ± 0.15	1.144 ± 0.005	1238
DEC (filtered)	-9.619 ± 0.13	-3.317 ± 0.13	1.140 ± 0.003	990

Table 5.6: NGC 2682 results.

5.2.4 Melotte 25

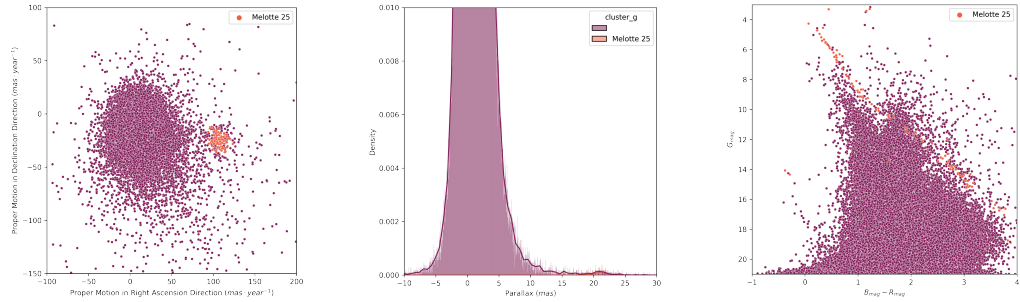


Figure 5.21: Melotte 25 Clusterix+TOPCAT characterization.

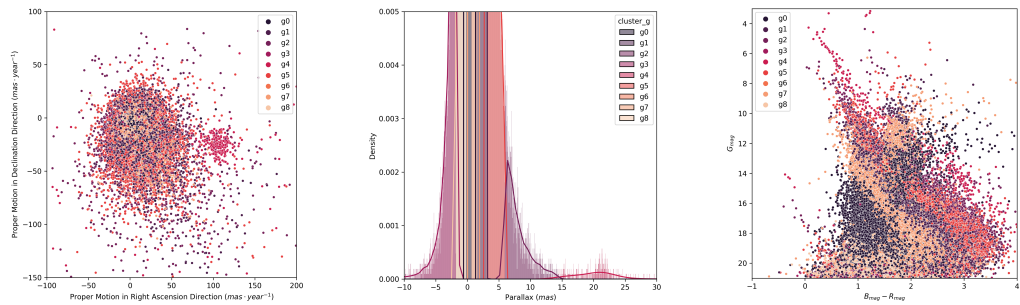


Figure 5.22: Melotte 25 K-Means characterization. Identified as cluster g_4 .

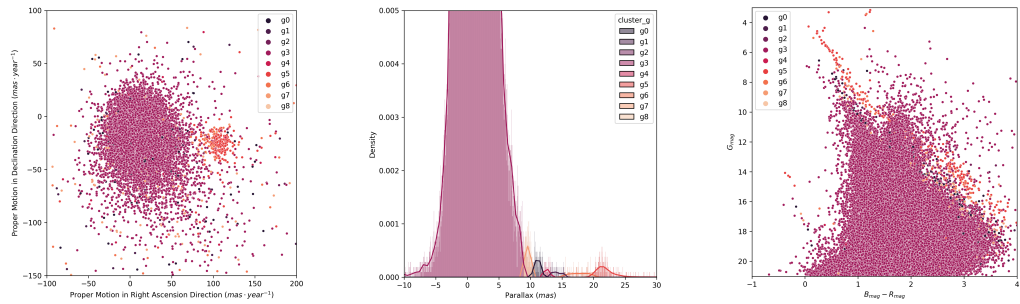


Figure 5.23: Melotte 25 DEC characterization. Identified as cluster g_5 .

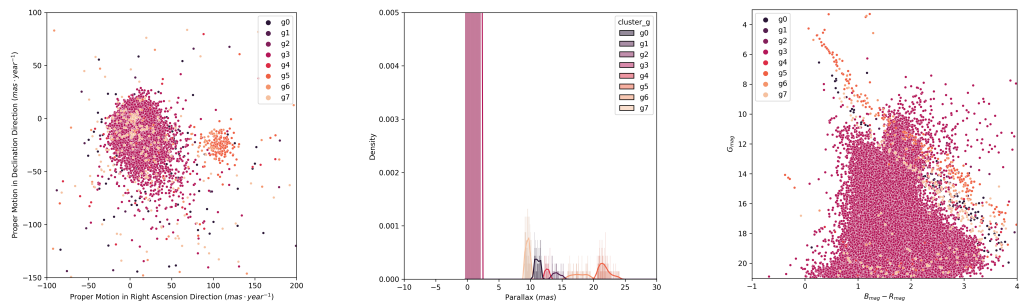


Figure 5.24: Melotte 25 DEC (filtered) characterization. Identified as cluster g_5 .

Figure 5.21 shows Melotte 25 characterized using Clusterix+TOPCAT tools. Figure 5.22 shows nine clusters identified by K-Means. K-Means labels Melotte 25 as group $g3$. Figures 5.23 and 5.24 show the groups found using the DEC model and the DEC model filtered, respectively. Melotte 25 is labeled as group $g5$ by DEC model.

Table 5.7 shows the hyperparameters used for characterizing Melotte 25 with the DEC model. Table 5.8 shows a results summary for Melotte 25.

Hyperparameter	Value
Number of Clusters	9
Clustering Layer	[50, 50, 40]
Kernel Initializer Seed	11
Quantil Threshold	0.1

Table 5.7: Melotte 25 DEC model hyperparameters.

Source / method	μ_α ($mas \cdot yr^{-1}$)	μ_δ ($mas \cdot yr^{-1}$)	ϖ (mas)	# stars
Simbad.u-strasbg.fr	104.92 ± 0.12	-28.00 ± 0.09	21.052 ± 0.065	-
Clusterix+TOPCAT	106.796 ± 6.229	-24.870 ± 5.417	21.210 ± 1.115	109
K-Means	79.936 ± 3.7	-45.362 ± 3.8	20.901 ± 0.31	374
DEC	104.051 ± 3.2	-33.424 ± 2.5	22.072 ± 0.13	219
DEC (filtered)	105.96 ± 3.5	-30.00 ± 2.4	21.74 ± 0.07	175

Table 5.8: Melotte 25 results.

5.3 Discussion

We have applied our proposed method to clusters with different typologies:

- NGC 2516 has its proper motion center deviated from the origin but it is embedded inside a big cloud of stars with similar proper motions.
- NGC 2632 (in addition to Melotte 22) is a cluster whose proper motion center is not located at the origin and has a well separated parallax center.
- NGC 2682, on the other hand, has its parallax centered within the region's gaussian, which complicates its detection although its proper motion center is deviated from the origin.
- Melotte 25 (as well as Melotte 22) is a cluster closer to us than the other. That makes its membership stars to be more scattered than previous clusters which are more compact.

All these clusters have a wide variety of diameters, from 25 to 330 arcmin, as well as the number of stars that belong to them. NGC 2682 has 3,000 stars while Melotte 25 is located inside a region with more than 400,000 stars. More information about the studied clusters is available in Table 5.9.

Open Cluster	α J2000 (degrees)	δ J2000 (degrees)	Radius (arcmin)	# stars
NGC 2516	119.517	-60.753	15	12,869
NGC 2632	130.1	19.667	35	13,167
NGC 2682	132.825	11.8	12.5	2,839
Melotte 22	56.75	24.117	60	61,552
Melotte 25	66.725	15.867	165	433,996

Table 5.9: Right ascension, declination, radius and number of stars of studied clusters. The number of stars corresponds to those stars contained within a cone of center (α, δ) and radius the cluster's radius multiplied by a factor of 1.5.

In all cases, the model has resolved properly the identity of the cluster and has characterized the membership stars, giving compatible results with the ones obtained with classic procedures and VO tools.

Compared with the mentioned tools, our model can be categorized as a non-parameterized model since it does not depend on parameters referred to the cluster itself but only hyperparameters such as the initial number of clusters to be found, or the structure of the Clustering Layer used by DEC model.

Our model is also non-supervised since we do not need to tell the model which stars belong to the cluster or assist it while its training stage.

Only at the end of the characterization, a fine tuning driven by the user, can be applied in order to improve the selection by discarding those stars which fall outside the given quantiles.

In any case, we do not make assumptions about the cluster profile, something that we have to take into account when using VO tools. This is another evidence that our model is non-parameterized.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

From the results shown in Chapter 5.2, we can say that, in general, our model works fine identifying and characterizing open clusters. So we have succeeded building a model non-parameterized and non-supervised for open cluster characterization. However, some improvements should be done in order to improve its accuracy and precision.

We have built a non-parameterized model since the user does not provide any information about the cluster profile or its properties. Nevertheless some tweaks can be done regarding the cluster, like changing the initial number of clusters or modifying the DEC's layer structure in order to improve the final characterization.

Another key point in our model is the initializer kernel. This kernel prepares data before passing it to the ANN and results may vary significantly depending on this kernel. For example, simply varying the kernel seed can make a huge difference on the result. This is something to avoid, so an improvement to solve this issue is necessary in order to have a reliable model which does not depend on the dataset order.

Despite that, comparing our method with Clusterix, we can still say that the number of parameters (or hyperparameters) in our model is smaller and that we do not need previous knowledge about the studied cluster. That makes easier to test different configurations and automate the process.

With all of these points in mind and once the mentioned issues have been solved, we can assert this model could be considered as a valid model for open cluster characterization. And thus, it could even be included as part of the Virtual Observatory toolset.

6.2 Further Research

The first aim after finishing this work is testing the model with a wider range of clusters. In general we could think of applying it to the whole VizieR catalogue. That way we could have a better idea about the current limits of our model and we could make adjustments to improve it. We could determine with higher accuracy the typologies that our model works well with, and which ones the model fails to resolve properly. Furthermore, we could establish different sets of hyperparameters regarding the typology of the cluster.

With our model, new secondary clusters arise from the data apart of the main ones. Hence, it is also interesting the individual study of these new clusters.

Another possible change in our model would be using DBSCAN instead of K-Means as our initial clustering algorithm. Maybe this algorithm gives us a better starting point for the DEC model which could improve the results.

Finally, we have used the Gaia DR2 database as our data source for this work, but recently the DR3 dataset has been released. Therefore it is evident the interest on testing our model with this new data.

References

- Astropy Collaboration, Price-Whelan, A. M., SipHocz, B. M., G”unther, H. M., Lim, P. L., Crawford, S. M., ... Astropy Contributors (2018, sep). The Astropy Project: Building an Open-science Project and Status of the v2.0 Core Package. *aj*, 156(3), 123. doi: 10.3847/1538-3881/aabc4f
- Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., Greenfield, P., Droettboom, M., Bray, E., ... Streicher, O. (2013, oct). Astropy: A community Python package for astronomy. *aap*, 558, A33. doi: 10.1051/0004-6361/201322068
- Babusiaux, C., van Leeuwen, F., Barstow, M., Jordi, C., Vallenari, A., Bossini, D., ... others (2018). Gaia data release 2-observational hertzsprung-russell diagrams. *Astronomy & Astrophysics*, 616, A10.
- Balaguer-Núñez, L., López del Fresno, M., Solano, E., Galadí-Enríquez, D., Jordi, C., Jimenez-Esteban, F., ... Paunzen, E. (2020). Clusterix 2.0: a virtual observatory tool to estimate cluster membership probability. *Monthly Notices of the Royal Astronomical Society*, 492(4), 5811–5843.
- Bayer, M. (2012). Sqlalchemy. In A. Brown & G. Wilson (Eds.), *The architecture of open source applications volume ii: Structure, scale, and a few more fearless hacks*. aosabook.org. Retrieved from <http://aosabook.org/en/sqlalchemy.html>
- Bayo, A., Rodrigo, C., y Navascués, D. B., Solano, E., Gutiérrez, R., Morales-Calderón, M., & Allard, F. (2008). Vosa: virtual observatory sed analyzer-an application to the collinder 69 open cluster. *Astronomy & Astrophysics*, 492(1), 277–287. Retrieved from <http://svo2.cab.inta-csic.es/theory/vosa/>
- Bellman, R. (1961). Curse of dimensionality. *Adaptive control processes: a guided tour*. Princeton, NJ, 3, 2.

Bonnarel, F., Fernique, P., Bienaymé, O., Egret, D., Genova, F., Louys, M., . . . Bartlett, J. G. (2000). The aladin interactive sky atlas-a reference tool for identification of astronomical sources. *Astronomy and Astrophysics Supplement Series*, 143(1), 33–40. Retrieved from <https://aladin.u-strasbg.fr/aladin.gml>

Bressan, A., Marigo, P., Girardi, L., Salasnich, B., Dal Cero, C., Rubele, S., & Nanni, A. (2012). Parsec: stellar tracks and isochrones with the padova and trieste stellar evolution code. *Monthly Notices of the Royal Astronomical Society*, 427(1), 127–145.

Brinkmann, N., Banerjee, S., Motwani, B., & Kroupa, P. (2017). The bound fraction of young star clusters. *Astronomy & Astrophysics*, 600, A49.

Cantat-Gaudin, T., Donati, P., Vallenari, A., Sordo, R., Bragaglia, A., & Magrini, L. (2016). Abundances and kinematics for ten anticentre open clusters. *Astronomy & Astrophysics*, 588, A120.

Cantat-Gaudin, T., Jordi, C., Vallenari, A., Bragaglia, A., Balaguer-Núñez, L., Soubiran, C., . . . others (2018). A gaia dr2 view of the open cluster population in the milky way. *Astronomy & Astrophysics*, 618, A93.

Castro-Ginard, A., Jordi, C., Luri, X., Cid-Fuentes, J. Á., Casamiquela, L., Anders, F., . . . others (2020). Hunting for open clusters in gaia dr2: 582 new open clusters in the galactic disc. *Astronomy & Astrophysics*, 635, A45.

Chollet, F., et al. (2015). *Keras*. Retrieved from <https://keras.io>

Clarke, C., Bonnell, I., & Hillenbrand, L. (2000). The formation of stellar clusters. In V. Mannings & A. Boss (Eds.), *Protostars and planets iv* (p. 151). University of Arizona Press.

Collaboration, G., et al. (2016). Description of the gaia mission (spacecraft, instruments, survey and measurement principles, and operations). *Gaia Collaboration et al.(2016a): Summary description of Gaia DR1*.

Dias, W., Alessi, B., Moitinho, A., & Lépine, J. (2002). New catalogue of optically visible open clusters and candidates. *Astronomy & Astrophysics*, 389(3), 871–873. Retrieved from <https://heasarc.gsfc.nasa.gov/W3Browse/star-catalog/openclust.html>

- Elsanhoury, W., & Nouh, M. (2019). Ppmxl and gaia morphological analysis of melotte 22 (pleiades) and melotte 25 (hyades). *New Astronomy*, 72, 19–27.
- Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* (Vol. 96, pp. 226–231).
- Gaia, C., Brown, A., Vallenari, A., Prusti, T., de Bruijne, J., Babusiaux, C., ... others (2018). Gaia data release 2 summary of the contents and survey properties. *Astronomy & Astrophysics*, 616(1).
- Janes, K., & Adler, D. (1982). Open clusters and galactic structure. *The Astrophysical Journal Supplement Series*, 49, 425–445.
- Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., ... Willing, C. (2016). Jupyter notebooks – a publishing format for reproducible computational workflows. In F. Loizides & B. Schmidt (Eds.), *Positioning and power in academic publishing: Players, agents and agendas* (p. 87 - 90).
- Krone-Martins, A., & Moitinho, A. (2014). Upmask: unsupervised photometric membership assignment in stellar clusters. *Astronomy & Astrophysics*, 561, A57.
- Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1), 79–86.
- Merkel, D. (2014). Docker: lightweight linux containers for consistent development and deployment. *Linux journal*, 2014(239), 2.
- Ochsenbein, F., Bauer, P., & Marcout, J. (2000). The vizier database of astronomical catalogues. *Astronomy and Astrophysics Supplement Series*, 143(1), 23–32. Retrieved from <https://vizier.unistra.fr> doi: 10.26093/cds/vizier
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Portegies Zwart, S. F., McMillan, S. L., & Gieles, M. (2010). Young massive star clusters. *Annual review of astronomy and astrophysics*, 48, 431–493.
- PostgreSQL. (2020). *Postgresql: The world's most advanced open source relational database*. Retrieved from <https://www.postgresql.org>

- Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20, 53–65.
- Taylor, M. B. (2005). Topcat & stil: Starlink table/votable processing software. In *Astronomical data analysis software and systems xiv* (Vol. 347, p. 29). Retrieved from <http://www.starlink.ac.uk/topcat/>
- Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. Scotts Valley, CA: CreateSpace.
- Waskom, M., Botvinnik, O., O’Kane, D., Hobson, P., Lukauskas, S., Gemperline, D. C., ... Qalieh, A. (2017, September). *mwaskom/seaborn: v0.8.1 (september 2017)*. Zenodo. Retrieved from <https://doi.org/10.5281/zenodo.883859> doi: 10.5281/zenodo.883859
- Wenger, M., Ochsenbein, F., Egret, D., Dubois, P., Bonnarel, F., Borde, S., ... others (2000). The simbad astronomical database-the cds reference database for astronomical objects. *Astronomy and Astrophysics Supplement Series*, 143(1), 9–22.
- Xie, J., Girshick, R., & Farhadi, A. (2016). Unsupervised deep embedding for clustering analysis. In *International conference on machine learning* (pp. 478–487).

Machine Learning Tools for Open Cluster Characterization with Gaia DR2 Data

CD. Álvaro, C. Guzman, J. Álvaro

Universidad Internacional de la Rioja, Logroño (España)

24th of January, 2021

Abstract

The characterization and understanding of *Open Clusters* (OCs) allow us to understand better properties and mechanisms about the Universe such as stellar formation and the regions where these events occur. They also provide information about stellar processes and the evolution of the galactic disk.

In this paper, we present a novel method to characterize OCs. Our method employs a model built on *Artificial Neural Networks* (ANNs). More specifically, we adapted a state of the art model, the *Deep Embedded Clustering* (DEC) model for our purpose. The developed method aims to improve classical state of the arts techniques. We improved not only in terms of computational efficiency (with lower computational requirements), but in usability (reducing the number of hyperparameters to get a good characterization of the analyzed clusters). For our experiments, we used the *Gaia DR2 database* as the data source, and compared our model with the clustering technique *K-Means*. Our method achieves good results, becoming even better (in some of the cases) than current techniques.

I. Introduction

Stellar OCs [1] are groups of stars gravitationally bounded originated from a single molecular gas cloud. They share the same chemical composition and age, and that is the reason that their metallicity must be uniform since those stars were born from the same gas cloud and at the same time stage. Likewise, they have similar relative positions inherited from their original gas cloud, which means that their distances to the Earth are the same. Therefore, they have a narrow dispersion in their parallax value. Moreover, they share similar values of proper motion, both in right ascension and declination. For all these properties, we can see why the stellar OCs are relevant to

understand the spiral structure, dynamics, and chemical evolution of our galaxy.

The study of OCs advanced thanks to the huge and precise dataset from the Gaia DR2 mission, available since 2018 from [2] and [3]. Although, the Gaia DR2 database misreported the chemical composition of some OCs, the database has helped to review already known OCs and to find new ones.

The characterization of an OC refers to the statistical determination of the stars that form it. Therefore, the characterization is based on the probability of a star to be a member of the cluster [4].

Normally, the characterization process considers astrometric features such as proper motion in right ascension and declination, or par-

unir
LA UNIVERSIDAD
EN INTERNET

Key Words

characterization,
data analysis,
deep embedded
clustering, gaia,
machine learning,
open cluster

allax. In some cases [5], it also considers photometric features, which helps to generate the H-R diagram [6] of the OC candidate stars. Those OC candidates should present a sharp profile corresponding to an isochrone curve derived from a theoretical model. The model considers metallicity, mass and brightness of the stars involved. For all this purpose, we used a set of tools that require supervised and parameterized models. It means that we need a previous knowledge of the cluster, otherwise we must repeat the process iteratively to achieve valid results.

In the present paper, we propose a machine learning model capable of characterizing OCs inside a stellar region with no previous knowledge about the region. Our model takes advantage of some features (such as proper motion in right ascension and declination, and parallax) to train an Artificial Neural Network. The ANN clusters all the stars within a given region in groups. One of those groups will be the open cluster that interests us.

Our model contributes with several novelties: i) it is *non-supervised* and *non-parameterized*, making easier the automation process of analyzing a wide range of regions with different typologies, and ii) it is computationally efficient to run in common workstations because it has been developed with Python using the Keras framework which takes advantage of modern GPUs to perform its computations. That increases significantly the computational capacity of our model doing it possible to run on regular workstations.

This paper is organized as follows. Section II presents some related works and Section III describes in detail our model. The following Section IV, shows the results and compares them with a method that involves tools from the Virtual Observatory (VO) such as Clusterix [7]. Finally, the last section concludes and outlines future research lines.

II. Related work

In [8], they present one of the initial approaches to detect OCs. They search for overdensities

in the astrometric space of the galactic disk. Then, they identify with photometric information possible OCs. Although it seems easy, it is quite hard to face the problem that the near field around the OC has two kinds of populations. The first population belongs to the members of the OC stars (from tens or hundreds to a few thousand). The second population has a background of stars that do not belong to the OC (from tens to hundreds of thousands). Thus, we summarize our problem in finding out which stars belong to the OC.

Other works (TOPCAT [9], Clusterix 2.0 [7], Aladin [10], or VOSA [11]), first analyze the proper motion configuration space of the region of interest. TOPCAT cannot find an open cluster. Instead, it requires parameterizing some of the properties of the cluster. This last process requires previous knowledge of the cluster. In other words, they perform a supervised and manual selection of groups based on overdensities in the proper motion configuration space. Clusterix 2.0 is an interactive web-based tool that can help the choice of groups. Take the proper motion diagram without making any previous assumptions about the membership of the candidate star. And empirically determine the frequency functions. It employs normal Gaussian kernel functions, defined as:

$$K(a, b) = \frac{1}{2\pi h^2} \exp \left[-\frac{1}{2} \frac{(a - a_i)^2 + (b - b_j)^2}{h^2} \right]$$

where (a, b) are the proper motion configuration space, (i, j) is a point located in the center of that provides the maximum contribution for calculating the local density, and h is the *smoothing* parameter, which it is measured in the same units as the proper motion.

Clusterix critically depends on the selection of three radii in the studied region. The inner radius contains stars belonging or not to the cluster. While the outermost radii defines a ring which only contains stars that do not belong to the cluster. In addition, it is also sensitive to other parameters, such as the soften parameter h , or other restrictions related with the searched proper motions.

If the analysis performed by Clusterix is successful, it returns the probability of each star to belong to the open cluster. The results can be imported then into TOPCAT to continue the refining process to get a valid characterization.

Another recent approach [8] solves the problem using machine learning techniques. It searches systematically for overdensities in the astrometric space of the galactic disk and a subsequent identification of open clusters using photometric information. It includes two phases:

1. it employs DBSCAN, an unsupervised clustering algorithm, to search for overdensities,
2. and it applies a deep learning ANN to identify isochrone patterns within the detected overdensities and thus proceed to confirm them as OCs. They trained previously the ANN with magnitude diagrams.

They execute their experiments in the Barcelona Supercomputing Center, MareNostrum 42. Thereby, the neural network handles the image recognition process with isochrone patterns without applying theoretical models derived from values such as metallicity or masses, among other. Their work concludes with the identification of 582 new open clusters distributed along the galactic disk for a galactic declination below 20 degrees, increasing the number of known open clusters by 45%.

In contrast to [8], we lack of a super computer like MareNostrum for our experiments. For this reason, our aim is to obtain a new novel method that allows the characterization of open clusters without doing a blind search for clusters characterization. We advocate of the idea of building an unsupervised and non-parameterized method. So, it can be suitable for automated processes.

III. Open Cluster Characterization method

In the literature, we have several clustering algorithms, such as K-Means, Mean-Shift Clustering, DBSCAN [12] among other, that can

help us to achieve our aim. Each one behaves better according to the distribution of the objects to clusterize. However, we need to set a large number of clusters to find the open cluster we are looking for. It complicates the identification of the OC because most of the time falls in many outliers.

The unsupervised *Deep Embedding for Clustering Analysis* (DEC) model [13] is a refinement K-Means based on ANN. It starts with K-Means. And then, it trains an autoencoder to reduce the feature space and pass them through a Clustering Layer which refines the previous selection.

With all this in mind, our method is based on DEC model. We adapted it to manage data groups based on the dynamic properties of the stars. Thus, we have an unsupervised clustering model for open cluster characterization. The model fits a wide range of clusters without the need for fine-tuning a high number of hyperparameters.

A Selection of the catalogue

In this work we make use of Gaia DR2 since DR3 has not been released in time for us to include it. Gaia DR2 is a multidimensional dataset obtained by ESA's Gaia mission (located at L2, 1.5 million kilometers from Earth) and operational since 2014. The catalogue has high precision and accuracy astrometric data for more than 1.7 billion stellar sources, and magnitudes in three photometric filters (G, BP and RP) for more than 1,300 million sources.

We start by selecting a region from the OpenClust [14] catalogue and we downloaded it from Gaia DR2 database. The radius of the downloaded region from Gaia is 1.5 times larger than the one registered in OpenClust, ensuring to include several stars that do not belong to the open cluster.

B Selection of features

For the selection of features, we studied the Melotte 22 dataset. Table 1 describes an example of the region Melotte 22. Each row represents the features of a star.

$pmra \text{ (mas} \cdot \text{yr}^{-1})$	$pmdec \text{ (mas} \cdot \text{yr}^{-1})$	$parallax \text{ (mas)}$
2.848682	-3.291204	0.399680
0.894901	-3.445501	0.416639
7.924372	-0.241281	0.397743
-4.433802	-2.584965	0.410695
0.055990	-1.760018	0.413813

Table 1: Features of Melotte 22 region.

Proper motion in right ascension (column $pmra$) and declination (column $pmdec$) seems like a natural choice since stars belonging to the same OC share a common motion vector. Parallax (column $parallax$) is another relevant feature. It lets us know how far stars are from us. Besides, since all stars within an open cluster were born from the same dust cloud, they must all have similar parallax.

However, we are not going to use these raw features. Instead, we are taking a combination of them. Figure 1 shows a pairwise relationship of our combination of features from Melotte 22.

First, we correct proper motion in right ascension and declination by dividing them by the parallax (variables $pmra_corr$ and $pmdec_corr$, respectively). That way, we normalize these quantities and help our clustering models to improve their performance. The modulus of the proper motion (variable pm_mod in Figure 1) is another computed property that we considered. We use it to relate both features and therefore force our model to keep them tight.

C Deep Open Clustering of stars

In this section we explain in more detail our model.

In order to define our model, one of the main problems is that we do not have a labeled dataset to train a supervised model. Thus, we have to deal with an unsupervised self-trained model. We adapted the Unsupervised *DEC* model [13] to our requirements.

Same as the DEC model, we have two main components:

- **deep autoencoder:** It is trained before passing the data through the *clustering layer*. It is composed by encoder layers fol-

lowed by decoder layers. Recent research has shown that this autoencoder provides meaningful and well-separated representations on real-world datasets [15] and [16] to the DEC model.

- **clustering layer:** This layer receives data transformed by the autoencoder and it is iteratively trained until a convergence criterion is met.

The deep autoencoder is used to transform the input data into a latent space using a non-linear mapping function $f_\theta : X \rightarrow Z$.

As it was described in Section B, the number of features we deal is not too large. This latent space helps us to start in a reduced number of features and avoids the “*curse of dimensionality*” [17].

The autoencoder is pretrained before fitting the model to generate predictions. Then, the encoder layers of the autoencoder are used with the aim of transforming input data to the latent space Z . Once the data has been transformed, a K-Means clusterer is used in order to make an initial clustering. K-Means cluster centers are used as the initial weights for the clustering layer.

With that initial configuration, the model iterates alternating between computing an auxiliary target distribution (Soft Assignment) and minimizing the Kullback-Leiber (KL) divergence [18] to it. This unsupervised algorithm allows us to improve the clustering.

$$p_{ij} = \frac{q_{ij}^2 / f_j}{\sum_{j'} q_{ij'}^2 / f_{j'}} \quad (1)$$

In the soft assignment stage, the *Student’s t-distribution* is used as a kernel to measure the similarity between the embedded points and the cluster centroid. While in the KL divergence minimization the algorithm iteratively refines clusters by learning from their high confidence assignments with the help of an auxiliary target distribution. The model is trained by matching the soft assignment to the target distribution. The choice of this target distribution is crucial for DEC’s performance. In

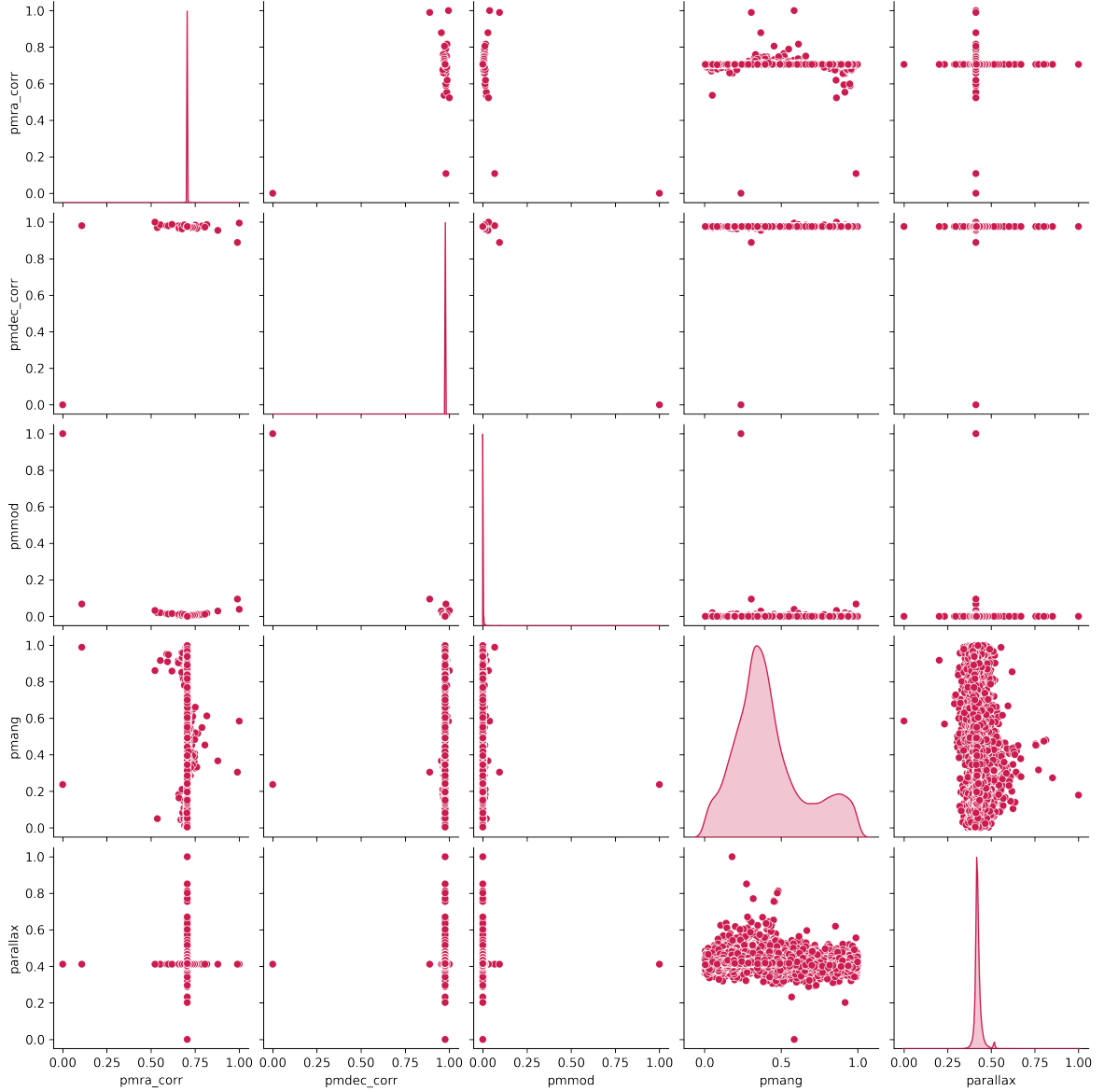


Figure 1: Pairwise relationships among variables using Melotte 22 data

this work we have taken the target distribution from DEC’s original paper [13], which is defined in Equation 1.

Figure 2 shows the layer setup of our DEC model. It is simpler than the one tested on the original paper [13], since the number of selected features in our work is smaller than in the original one. Therefore, using the same configuration would result in a model so powerful that would incur in overfitting issues unable to make right predictions.

Finally, we can refine this selection by filtering those stars which are below and above the

0.10 and 0.90 quantiles for each group, respectively. That way we remove the most doubtful values from the selection.

Discussion. Since we are looking for a single cluster, it seems reasonable to use a clustering algorithm like K-Means to find two clusters. One for the desired OC and another for stars that do not belong to the OC. However, the idea is not completely right. This is due to the fact that OC’s stars are surrounded by other stars with possibly similar properties. Therefore, setting the number of clusters to two is

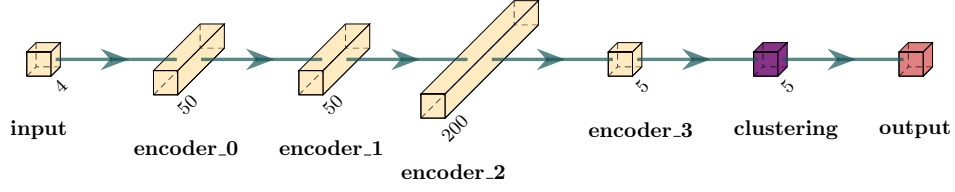


Figure 2: DEC model layer setup

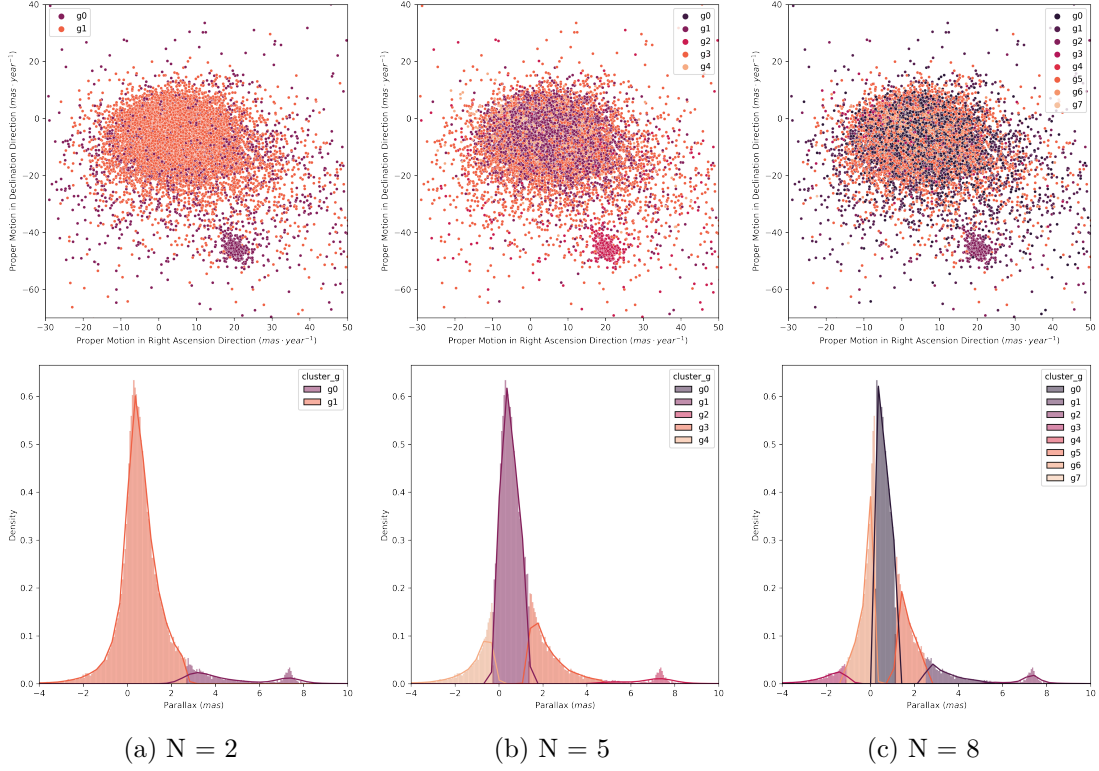


Figure 3: K-Means with Melotte 22

too low to separate them properly.

Figure 3 shows an example. It describes the results of applying K-Means with $N=2$, $N=5$ and $N=8$ in Melotte 22. As we can see, large values of the number of clusters allow us to isolate more accurately the resonance in parallax at $\approx 7.3\text{mas}$.

The disadvantage is that it forms more groups of stars, which complicates the task of finding the desired OC, since we would like to get just two groups. Therefore, with a K-Means we have to find a way to set the right value for the number of clusters to isolate the searched cluster without creating too many groups. To solve this issue, we can try to

estimate the best number of clusters by using the *silhouette score* [19].

K-Means does a good job making an initial clustering. However, too many clusters arise from this characterization and the OC is still polluted with stars that do not belong to it. Moreover, we would like to reduce the amount of clusters too.

IV. Results

We have tested our model with the Melotte 25 dataset, which is one of the most difficult dataset to characterize OC by its distribution

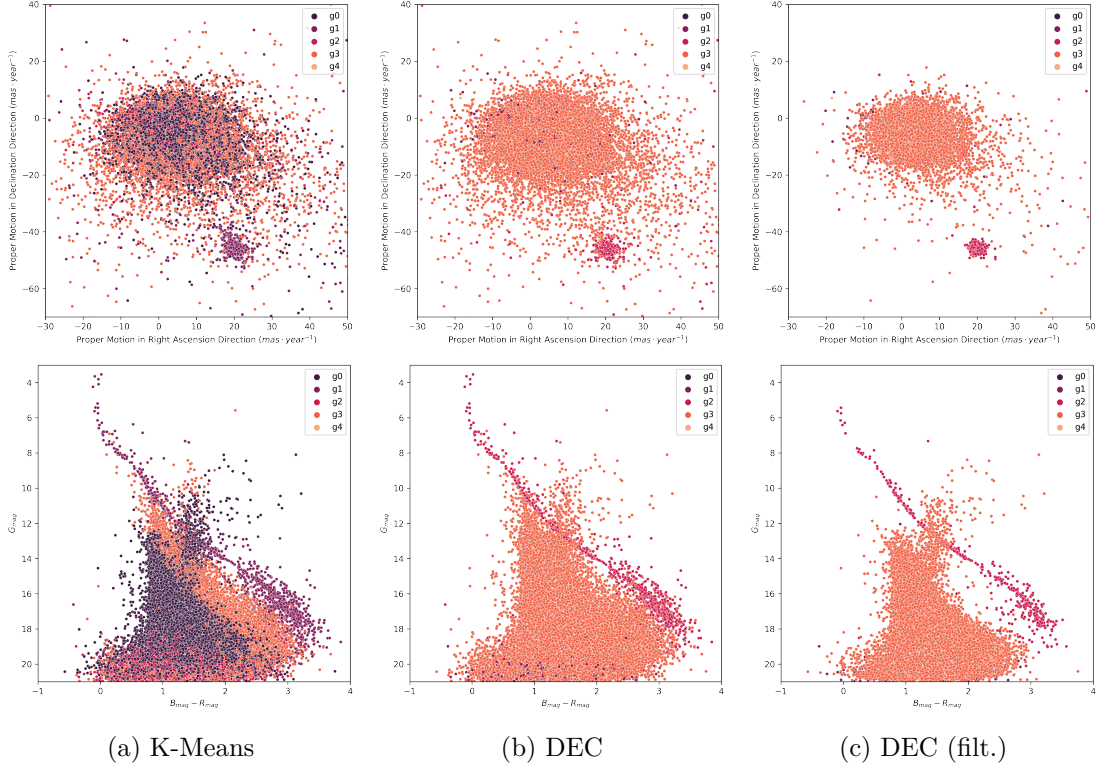


Figure 4: Evolution of Melotte 22 characterization

and extension. In order to support the discussion presented in Section C, we compared our model against a K-Means classification algorithm. In addition, we perform a second execution only in our model where we filtered the quantiles in the Melotte 22 dataset.

Our model was implemented in Python 3.8 using the Keras 2.2 framework and Jupyter Notebooks [20]. All tests were run on an Apple Mac Pro Late 2013 with a 2.7GHz 12-Core Intel Xeon E5-2697v2, and 64GB RAM 1866MHz DDR3. For the GPU, we use a graphic card AMD FirePro D700 with 6GB ¹.

Figure 4 shows the evolution of the characterization of Melotte 22. The first column corresponds to the characterization obtained by applying K-Means alone to the dataset. The middle column is the characterization made by the DEC model. And the column on the right shows the characterization made by the DEC and having filtered the quantiles lower

than 0.10 and higher than 0.90. The top row shows the proper motion configuration space while the bottom one shows the Hertzsprung–Russell diagrams.

Melotte 25 is located at 66.725 degrees in right ascension and 15.867 degrees in declination with a radius of 165 arcmin. This region contains more than 400,000 stars.

Table 2 shows a results summary for Melotte 25.

Method	μ_α (mas · yr ⁻¹)	μ_δ (mas · yr ⁻¹)	ϖ (mas)	# stars
Simbad ²	104.92 ± 0.12	-28.00 ± 0.09	21.052 ± 0.065	-
Clusterix	106.796 ± 6.229	-24.870 ± 5.417	21.210 ± 1.115	109
K-Means	79.936 ± 3.7	-45.362 ± 3.8	20.901 ± 0.31	374
DEC	104.051 ± 3.2	-33.424 ± 2.5	22.072 ± 0.13	219
DEC (filt.)	105.96 ± 3.5	-30.00 ± 2.4	21.74 ± 0.07	175

Table 2: Melotte 25 results.

Parameters shown are proper motion in right ascension and declination, parallax with their respective deviations and number of stars corresponding to the OC.

¹All resources developed for this project are available at <https://github.com/cdalvaro/machine-learning-master-thesis>

²Results have been taken from the SIMBAD astronomical database[21]

Figure 5 shows the groups found using the DEC model. This model labels Melotte 25 as group *g5*.

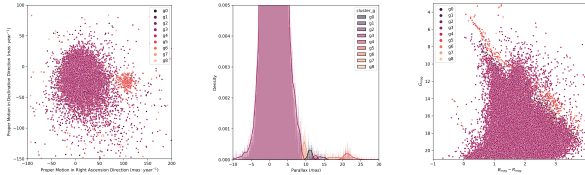


Figure 5: Melotte 25 characterization using the DEC model. From the left to the right: proper motion configuration space, parallax histogram and H-R diagram.

Table 3 shows the hyperparameters used for the characterization.

Hyperparameter	Value
Number of Clusters	9
Clustering Layer	[50, 50, 40]
Kernel Initializer Seed	11
Quantil Threshold	0.1

Table 3: Melotte 25 DEC hyperparameters.

V. Result discussions

We have applied our proposed method to clusters with different typologies:

- NGC 2516 (Ap. A) has its proper motion center deviated from the origin but it is embedded inside a big cloud of stars with similar proper motions.
- NGC 2632 (Ap. B), in addition to Melotte 22, is a cluster whose proper motion center is not located at the origin and has a well separated parallax center.
- NGC 2682 (Ap. C), on the other hand, has its parallax centered within the region’s gaussian, which complicates its detection although its proper motion center is deviated from the origin.
- Melotte 22 (Ap. D), as well as Melotte 25, is a cluster closer to us than the

other. That makes its membership stars to be more scattered than previous clusters which are more compact.

All these clusters have a wide variety of diameters, from 25 to 330 arcmin, as well as the number of stars that belong to them. NGC 2682 has 3,000 stars while Melotte 25 is located inside a region with more than 400,000 stars. More information about the studied clusters is available in Table 4.

OC	α J2000 (degrees)	δ J2000 (degrees)	Radius (arcmin)	# stars
NGC 2516	119.517	-60.753	15	12,869
NGC 2632	130.1	19.667	35	13,167
NGC 2682	132.825	11.8	12.5	2,839
Melotte 22	56.75	24.117	60	61,552
Melotte 25	66.725	15.867	165	433,996

Table 4: Right ascension, declination, radius and number of stars of studied clusters. The number of stars corresponds to those stars contained within a cone of center (α, δ) and radius the cluster’s radius multiplied by a factor of 1.5.

In all cases, the model has resolved properly the identity of the cluster and has characterized the membership stars, giving compatible results with the ones obtained with classic procedures and VO tools.

Compared with the mentioned tools, our model can be categorized as a non-parameterized model since it does not depend on parameters referred to the cluster itself but only hyperparameters such as the initial number of clusters to be found, or the structure of the Clustering Layer used by DEC model.

The proposed method is also non-supervised since we do not need to tell the model which stars belong to the cluster or assist it while its training stage.

Only at the end of the characterization, a fine tuning driven by the user, can be applied in order to improve the selection by discarding those stars which fall outside the given quantiles.

In any case, we do not make assumptions about the cluster profile, something that we have to take into account when using VO tools. This is another evidence that our model is non-parameterized.

VI. Conclusions

From the results shown in Section IV, we can say that, in general, our model works fine identifying and characterizing open clusters. So we have succeeded building a model non-parameterized and non-supervised for open cluster characterization. However, some improvements should be done in order to improve its accuracy and precision.

Another achievement is that we have developed a model that does not require complex hardware. Instead, it can be run on workstations with a common GPU, making it accessible to be implemented in many research centers.

We have built a non-parameterized model since the user is not required to provide any information about the cluster profile or its properties. Nevertheless some tweaks can be done regarding the cluster, like changing the initial number of clusters or modifying the clustering layer structure in order to improve the final characterization.

Another key point in our model is the initializer kernel. This kernel prepares data before passing it to the ANN and the results may vary significantly depending on this kernel. This is something to avoid, so an improvement to solve this issue is necessary in order to have a reliable model which does not depend on the dataset order.

Despite that, comparing our method with Clusterix, we can still say that the number of parameters (or hyperparameters) in our model is smaller and that we do not need previous knowledge about the studied cluster. That makes easier to test different configurations and automate the process.

Furthermore, our model is able to deal with large and open regions. As an example, it has succeeded characterizing Melotte 25 (Hades) OC. Also, the relative proximity of this cluster, 45 parsec, makes it difficult to characterize. Clusterix fails characterizing this cluster. In this case, our method proposes at least two more non-catalogued clusters. The analysis of these new clusters brings out an interesting stellar dynamics in the studied region. It

could mean that there was a recent interaction among different molecular clouds.

For all this, we consider that the presented method is good enough to be included in the Virtual Observatory toolset. Only a few tweaks must be done in order to make our model compatible with other tools of the VO.

A Future work

One open point to do from now is testing the model with a wider range of clusters. In general we could think of applying it to the whole VizieR catalogue. That way we could have a better idea about the current limitations of our model and we could make some adjustments to improve it. We could determine with higher accuracy the typologies that our model works well with, and which ones the model does not resolve properly. Furthermore, we could establish different sets of hyperparameters regarding the typology of the cluster.

With our model, new uncatalogued clusters arise from the data apart of the main ones. Hence, it is also interesting the individual study of these new clusters.

Another possible change in our model would be using DBSCAN instead of K-Means as our initial clustering algorithm. Maybe this algorithm gives us a better starting point for the DEC model which could improve the results.

Finally, we have used the Gaia DR2 database as our data source for this work, but recently the DR3 dataset has been released. Therefore, it is evident the interest on testing our model with this new data.

A. Appendix

A NGC 2516

Method	μ_α (mas · yr ⁻¹)	μ_δ (mas · yr ⁻¹)	ϖ (mas)	# stars
Simbad	-4.6579 ± 0.0075	11.1517 ± 0.0075	2.4118 ± 0.0006	1727
Clusterix	-4.652 ± 0.523	11.203 ± 0.454	2.409 ± 0.127	638
K-Means	-4.344 ± 0.14	9.507 ± 0.19	2.268 ± 0.01	1542
DEC	-4.426 ± 0.17	9.952 ± 0.20	2.436 ± 0.01	1532
DEC (filt.)	-4.502 ± 0.14	10.114 ± 0.17	2.392 ± 0.004	1072

Table 5: NGC 2516 results.

Table 5 shows a results summary for NGC 2516. The top row of Figure 6 shows the characterization using Clusterix+TOPCAT tools while the bottom row shows eight clusters found by K-Means.

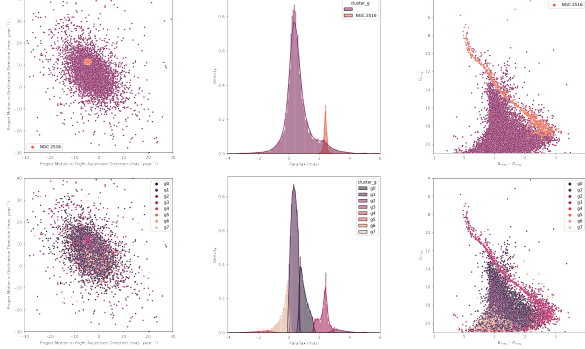


Figure 6: NGC 2516 characterization. Top: Clusterix+TOPCAT. Bottom: K-Means ($g3$).

Figure 7 shows the groups found using the DEC model (top row) and the DEC model filtered (bottom row). K-Means and DEC have labeled NGC 2516 as $g3$. Although in general, groups between K-Means and DEC models may not match. Hyperparameters used in this characterization are listed in Table 6.

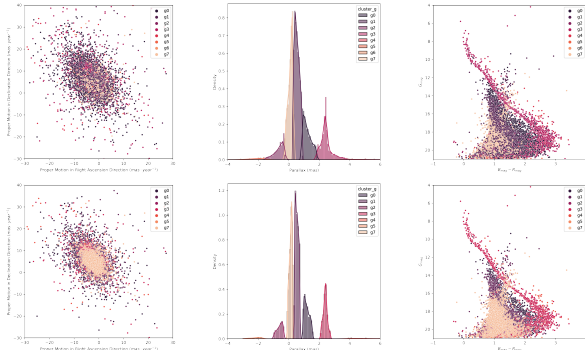


Figure 7: NGC 2516 characterization (labeled as $g3$). DEC (top). DEC filtered (bottom).

Hyperparameter	Value
Number of Clusters	8
Clustering Layer	[50, 50, 60]
Kernel Initializer Seed	2
Quantil Threshold	0.15

Table 6: NGC 2516 DEC hyperparameters.

B NGC 2632

Figure 8 shows NGC 2632 characterized by the validation method (first row) and the characterization made by K-Means showing ten clusters (second row). K-Means labels NGC 2632 as $g1$.

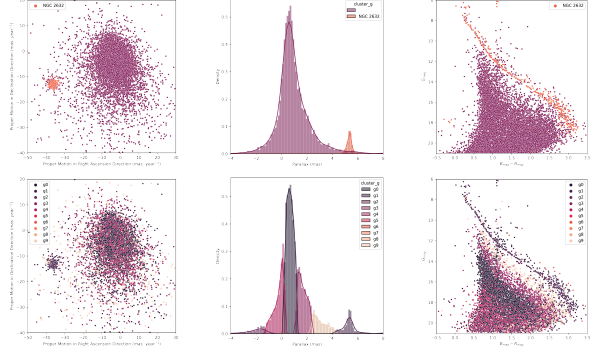


Figure 8: NGC 2632 characterization using Clusterix+TOPCAT (top) and K-Means (bottom). K-Means identifies NGC 2632 as $g1$.

Figure 9 shows the groups found using the DEC model (first row) and the DEC model filtered (second row). Open cluster NGC 2632 is labeled as group $g1$.

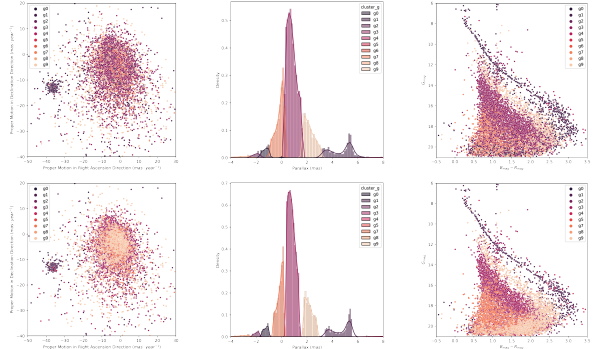


Figure 9: NGC 2632 characterization using DEC model (top) and DEC filtered (bottom). NGC 2632 is identified as $g1$.

Method	μ_α (mas \cdot yr $^{-1}$)	μ_δ (mas \cdot yr $^{-1}$)	ϖ (mas)	# stars
Simbad	-36.047 ± 0.110	-12.917 ± 0.066	5.371 ± 0.003	-
Clusterix	-36.154 ± 1.001	-12.909 ± 0.806	5.327 ± 0.187	371
K-Means	-26.352 ± 0.82	-15.828 ± 0.76	5.394 ± 0.03	629
DEC	-20.012 ± 0.69	-14.742 ± 0.58	4.686 ± 0.03	894
DEC (filt.)	-21.571 ± 0.74	-14.234 ± 0.61	4.719 ± 0.03	714

Table 7: NGC 2632 results.

Table 7 shows a results summary for NGC

2632 analysis. While Table 8 shows the hyperparameters used for characterizing NGC 2632 with our model.

Hyperparameter	Value
Number of Clusters	10
Clustering Layer	[50, 50, 40]
Kernel Initializer Seed	10
Quantil Threshold	0.1

Table 8: NGC 2632 DEC hyperparameters.

C NGC 2682

Table 9 shows a results summary for NGC 2682.

Method	μ_α (mas \cdot yr $^{-1}$)	μ_δ (mas \cdot yr $^{-1}$)	ϖ (mas)	# stars
Simbad	-10.9737 ± 0.0064	-2.9396 ± 0.0063	1.1325 ± 0.0011	1194
Clusterix	-10.970 ± 0.322	-2.958 ± 0.327	1.142 ± 0.080	649
K-Means	-8.616 ± 0.15	-3.710 ± 0.16	1.196 ± 0.01	1374
DEC	-8.926 ± 0.15	-3.550 ± 0.15	1.144 ± 0.005	1238
DEC (filt.)	-9.619 ± 0.13	-3.317 ± 0.13	1.140 ± 0.003	990

Table 9: NGC 2682 results.

Figure 10 shows NGC 2682 characterized using Clusterix+TOPCAT tools (top row) and ten clusters identified by K-Means (bottom row). The characterization made by K-Means labels NGC 2682 as group $g4$.

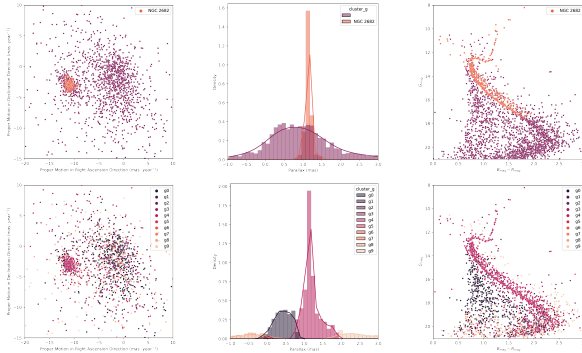


Figure 10: NGC 2682 K-Means characterization. Top row corresponds to Clusterix+TOPCAT characterization, while bottom row shows K-Means results identifying NGC 2682 as cluster $g4$.

Figure 11 shows the groups found using the DEC model (first row) and the DEC model fil-

tered (second row). The cluster of interest is the group $g2$.

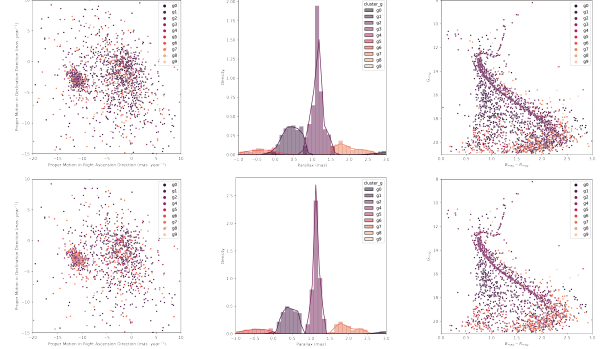


Figure 11: NGC 2682 is identified as cluster $g2$. Top row: DEC — Bottom row: DEC (filtered)

Table 10 shows the hyperparameters used for characterizing NGC 2682 with DEC method.

Hyperparameter	Value
Number of Clusters	10
Clustering Layer	[50, 50, 40]
Kernel Initializer Seed	0
Quantil Threshold	0.1

Table 10: NGC 2682 DEC hyperparameters.

D Melotte 22

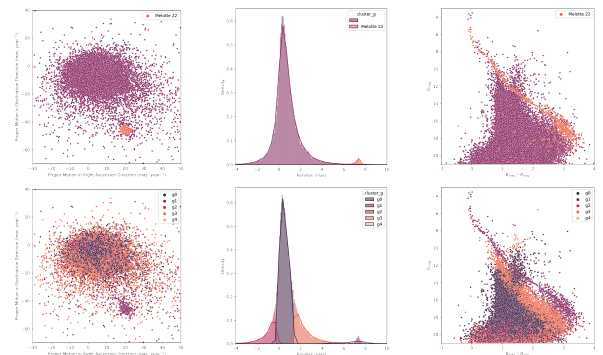


Figure 12: Melotte 22 characterization. Top row: Clusterix+TOPCAT. Bottom row: K-Means. Melotte 22 is labeled as cluster $g1$.

Figure 12 shows Melotte 22 characterized using Clusterix+TOPCAT tools (top row) and the characterization made by K-Means (bot-

tom row) showing nine clusters. Melotte 22 is labeled as group *g1*.

Figure 13 shows the groups found using DEC model (first row) and DEC model filtered (second row). Melotte 22 is labeled as group *g2* by DEC model.

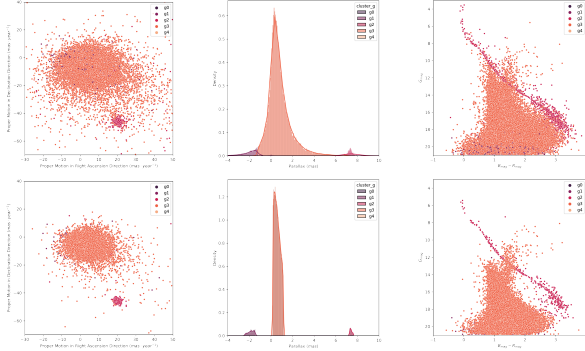


Figure 13: Melotte 22 is identified as cluster *g2*. Top row: DEC — Bottom row: DEC (filtered)

Table 11 shows a results summary for Melotte 22.

Method	μ_α ($mas \cdot yr^{-1}$)	μ_δ ($mas \cdot yr^{-1}$)	ϖ (mas)	# stars
Simbad	19.997 ± 0.127	-45.548 ± 0.101	7.364 ± 0.005	1326
Clusterix	19.98 ± 1.25	-45.47 ± 1.48	7.33 ± 0.21	634
K-Means	20.25 ± 0.95	-38.01 ± 1.08	7.23 ± 0.06	1378
DEC	23.67 ± 1.29	-46.23 ± 1.50	8.04 ± 0.09	878
DEC (filt.)	19.50 ± 0.41	-44.23 ± 0.39	7.42 ± 0.005	438

Table 11: Melotte 22 results.

While Table 12 shows the hyperparameters used for characterizing Melotte 22 with the DEC model.

Hyperparameter	Value
Number of Clusters	5
Clustering Layer	[50, 50, 200]
Kernel Initializer Seed	11
Quantil Threshold	0.1

Table 12: Melotte 22 DEC hyperparameters.

References

[1] K Janes and D Adler. Open clusters and galactic structure. *The Astrophys-*

ical Journal Supplement Series, 49:425–445, 1982.

[2] Gaia Collaboration et al. Description of the gaia mission (spacecraft, instruments, survey and measurement principles, and operations). *Gaia Collaboration et al.(2016a): Summary description of Gaia DR1*, 2016.

[3] Collaboration Gaia, AGA Brown, A Valenari, T Prusti, JHJ de Bruijne, C Babusiaux, ÁL Juhász, G Marschalló, G Marton, L Molnár, et al. Gaia data release 2 summary of the contents and survey properties. *Astronomy & Astrophysics*, 616(1), 2018.

[4] Laura Sampedro. Caracterización de sistemas estelares en espacios de n-dimensiones: Simulaciones y aplicación al catálogo astrométrico ucac4. 2016.

[5] AF Oliveira, H Monteiro, WS Dias, and TC Caetano. Fitting isochrones to open cluster photometric data-iii. estimating metallicities from ubv photometry. *Astronomy & Astrophysics*, 557:A14, 2013.

[6] Arkadiusz Hypki. Gaia data release 2: Observational hertzsprung-russell diagrams. *Astronomy & Astrophysics*, 616, 2018.

[7] L Balaguer-Núñez, M López del Fresno, E Solano, D Galadí-Enríquez, C Jordi, F Jimenez-Esteban, E Masana, J Carbajo-Hijarrubia, and E Paunzen. Clusterix 2.0: a virtual observatory tool to estimate cluster membership probability. *Monthly Notices of the Royal Astronomical Society*, 492(4):5811–5843, 2020.

[8] Alfred Castro-Ginard, C Jordi, X Luri, J Álvarez Cid-Fuentes, Laia Casamiquela, Friedrich Anders, Tristan Cantat-Gaudin, M Monguió, Lola Balaguer-Núñez, S Solà, et al. Hunting for open clusters in gaia dr2: 582 new open clusters in the galactic disc. *Astronomy & Astrophysics*, 635:A45, 2020.

- [9] Mark B Taylor. Topcat & stil: Starlink table/votable processing software. In *Astronomical data analysis software and systems XIV*, volume 347, page 29, 2005.
- [10] François Bonnarel, Pierre Fernique, Olivier Bienaymé, Daniel Egret, Françoise Genova, Mireille Louys, François Ochsenbein, Marc Wenger, and James G Bartlett. The aladin interactive sky atlas-a reference tool for identification of astronomical sources. *Astronomy and Astrophysics Supplement Series*, 143(1):33–40, 2000.
- [11] A Bayo, C Rodrigo, D Barrado y Navascués, E Solano, R Gutiérrez, M Morales-Calderón, and F Allard. Vosa: virtual observatory sed analyzer-an application to the collinder 69 open cluster. *Astronomy & Astrophysics*, 492(1):277–287, 2008.
- [12] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.
- [13] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487, 2016.
- [14] WS Dias, BS Alessi, A Moitinho, and JRD Lépine. New catalogue of optically visible open clusters and candidates. *Astronomy & Astrophysics*, 389(3):871–873, 2002.
- [15] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.
- [16] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.
- [17] Robert Bellman. Curse of dimensionality. *Adaptive control processes: a guided tour*. Princeton, NJ, 3:2, 1961.
- [18] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.
- [19] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.
- [20] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, and Carol Willing. Jupyter notebooks – a publishing format for reproducible computational workflows. In F. Loizides and B. Schmidt, editors, *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pages 87 – 90. IOS Press, 2016.
- [21] Marc Wenger, François Ochsenbein, Daniel Egret, Pascal Dubois, François Bonnarel, Suzanne Borde, Françoise Genova, Gérard Jasiewicz, Suzanne Laloë, Soizick Lesteven, et al. The simbad astronomical database-the cds reference database for astronomical objects. *Astronomy and Astrophysics Supplement Series*, 143(1):9–22, 2000.