Machine Learning Tools for Open Cluster Characterization with Gaia DR2 Data

Carlos David Álvaro Yunta

Advisor: Dr. César Augusto Guzmán Álvarez

unir LA UNIVERSIDAD EN INTERNET

Universidad Internacional de La Rioja

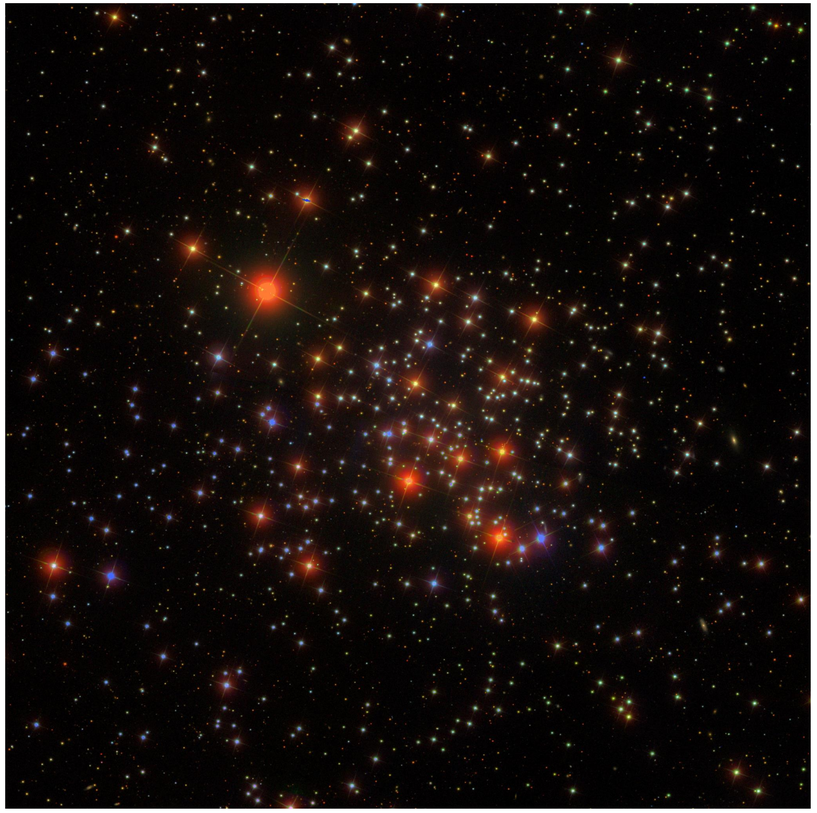
Index

Open Clusters

What is an open cluster?

NGC 2516 | right asc.: $7^h\ 58.3^m$ - dec.: *-60° 52'*

NGC 2632 | right asc.: $8^h\ 40.4^m$ - dec.: *19° 59'*



NGC 2682 | right asc.: $8^h\ 51.3^m$ - dec.: *+11° 49'*

Open Clusters

Why study them?

Young OCs allow us to *research star formation regions* and *improve our understanding about the mechanisms that create other 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.

Characterization

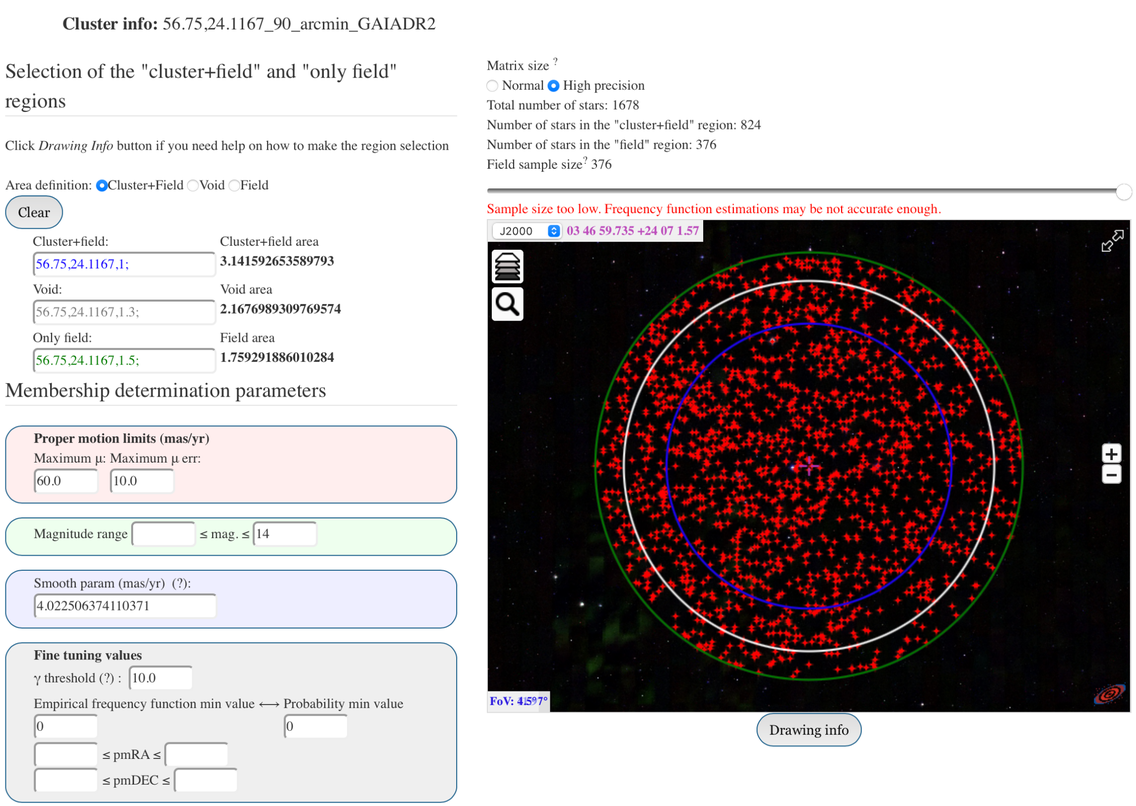
State of the art

Clusterix is an interactive web-based tool. It *takes the proper motion diagram* without making any prior assumption about the membership of the candidate star *and determines the probability value of each object to belong to the open cluster*.

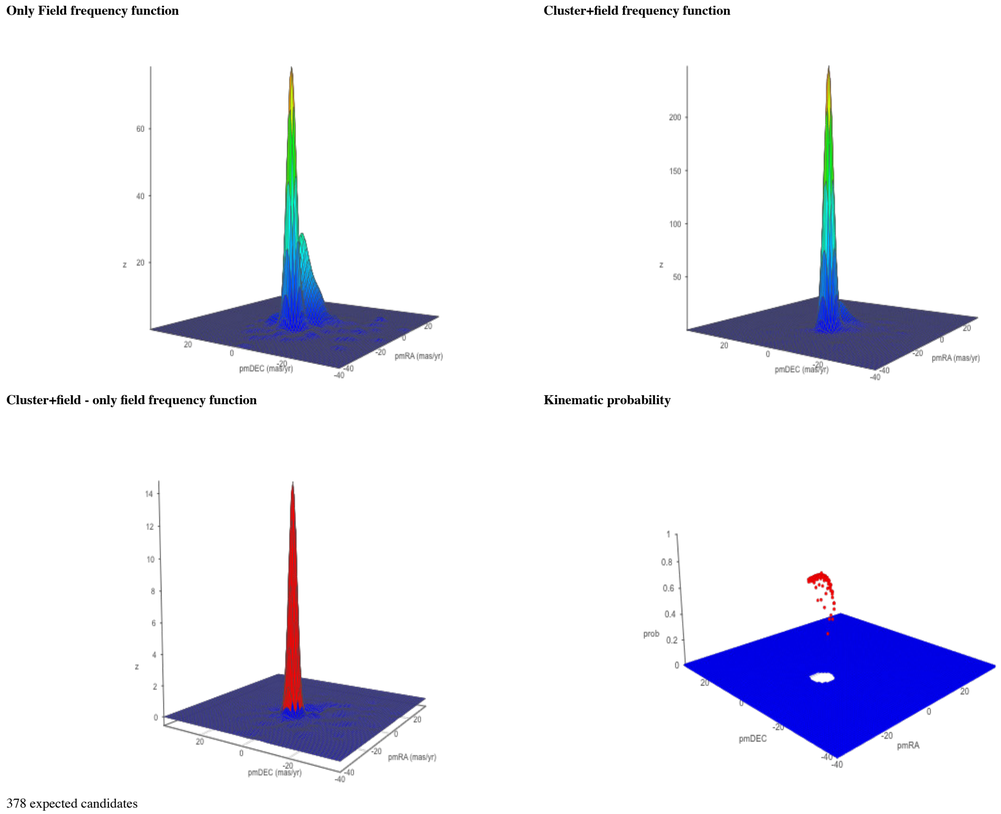
However, *the selection must be refined using tools such as TOPCAT*.

Characterization

State of the art



Clusterix 2.0 | Parameters panel



Clusterix 2.0 | Frequencies and Probability results

Characterization

State of the art

Castro-Ginard et al., 2020 work 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:

1. An *unsupervised clustering algorithm* is used, DBSCAN, *to search for overdensities*.
2. A *deep learning Artificial Neural Network* (ANN) is applied *to identify isochrone patterns* within the detected overdensities and thus proceed to confirm them as OC.

Characterization

Drawbacks

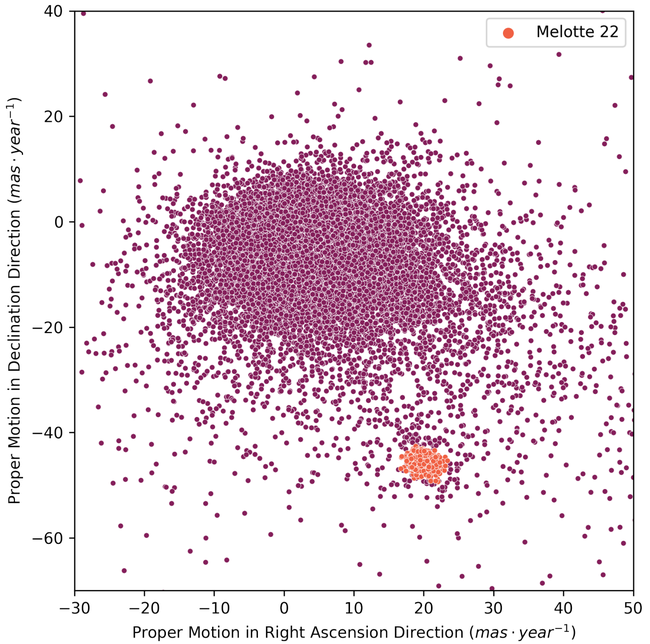
The near field around the open cluster is filled with ***two types of distinct star populations***:

- ***those who belong to the OC*** (tens or hundreds to a few thousand)
- ***a background*** made up of thousands or millions ***of stars that are not***.

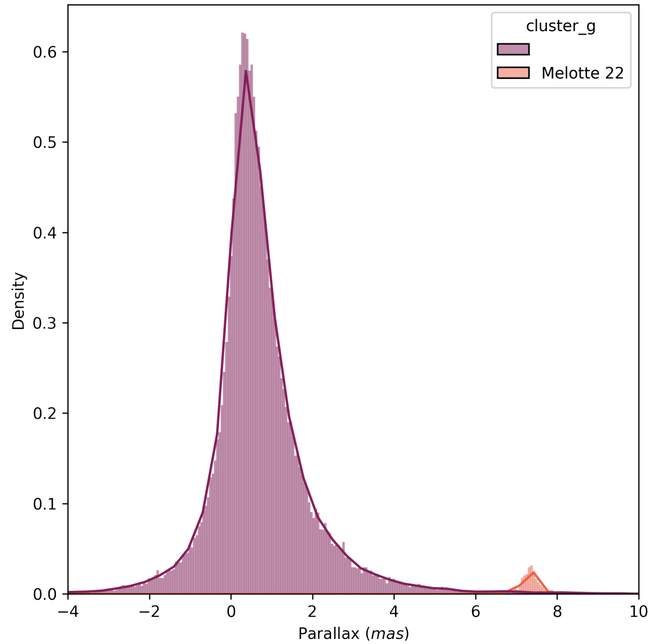
Characterization

Drawbacks

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



Melotte 22 | Proper motion distribution in right ascension and declination.
OC located at [20, -45].



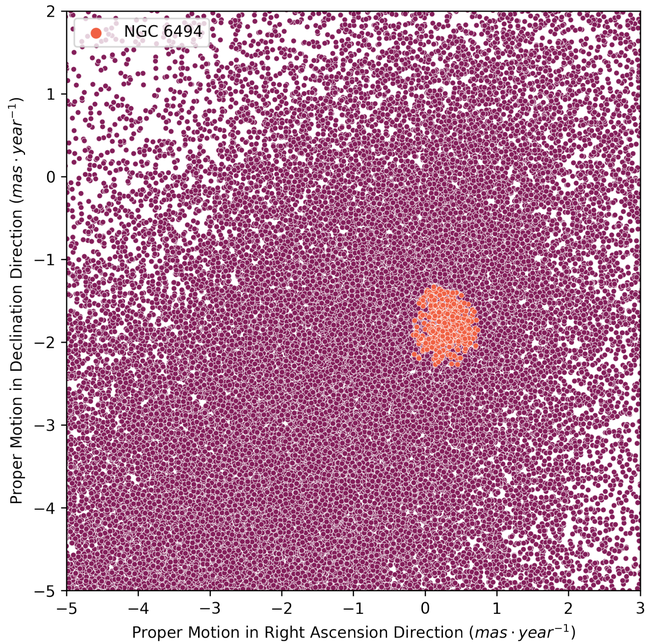
Melotte 22 | Parallax distribution
Confirms an overdensity at ~7.3 mas.

Characterization

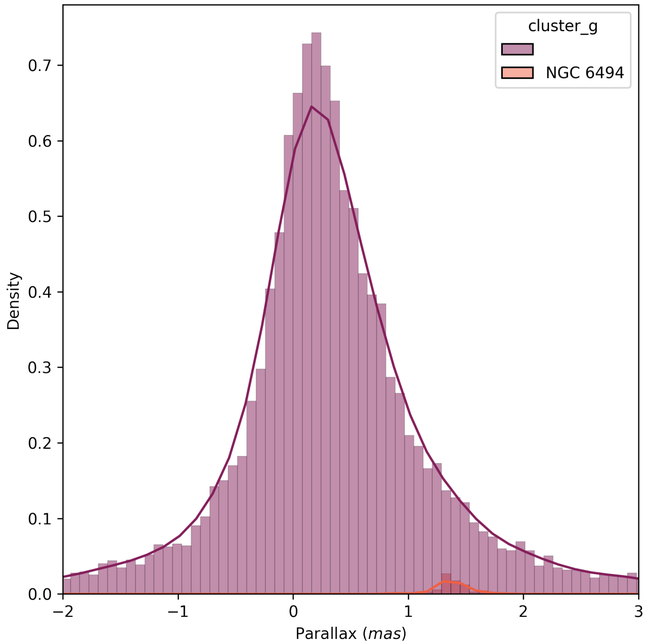
Drawbacks

However, *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.



NGC 6494 | Proper motion distribution in right ascension and declination.



NGC 6494 | Parallax distribution.

Characterization

Drawbacks

Virtual Observatory's 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*.

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

There is no rule of thumb that defines relative proportions of these areas.

Clusterix's results must be filtered with TOPCAT as a refinement process.

Characterization

Drawbacks

It should be noted that *for the execution of the method proposed in Castro-Ginard et al., 2020 work, MareNostrum 4 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.

These drawbacks make current methods *difficult to automate*.

Aims

The primary aim of this work is to ***develop and implement an unsupervised and non-parameterized clustering model for open cluster characterization***.

The model must be ***able to run under medium specs machines*** avoiding the necessity for supercomputers.

We want our model to be ***suitable for automated processes***.

The ***results*** of our model ***have to be comparable with other characterization techniques***, e.g.: Clusterix+TOPCAT.

Development

Clustering tools

We are looking for a clustering model that ideally separates our OC's components from the rest of the stars in the studied region.

Since ***we do not have data for training and validate our model***, we need it to be ***unsupervised self-trained***.

There are several clustering models: ***K-Means***, Mean-Shift, DBSCAN, among other.

Development

K-Means

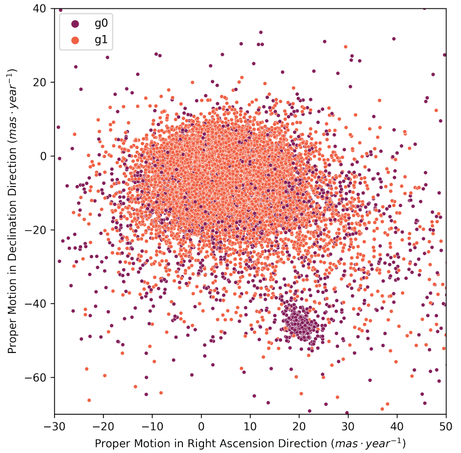
We have chosen K-Means by its ***simplicity and good results***.

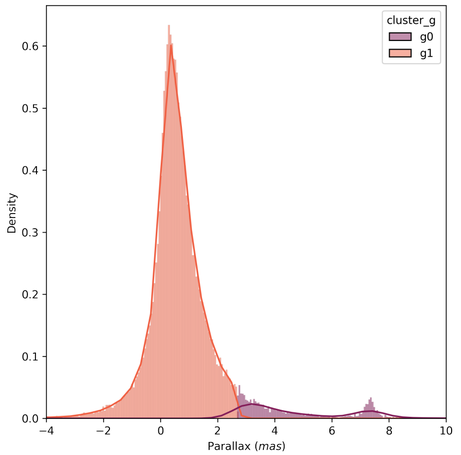
It ***requires a single parameter***, the number of clusters to build, ***N***.

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***.

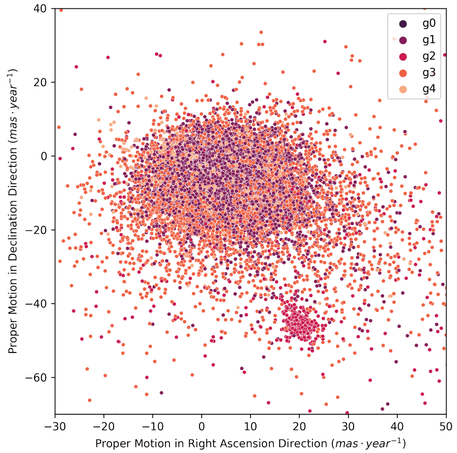
Development

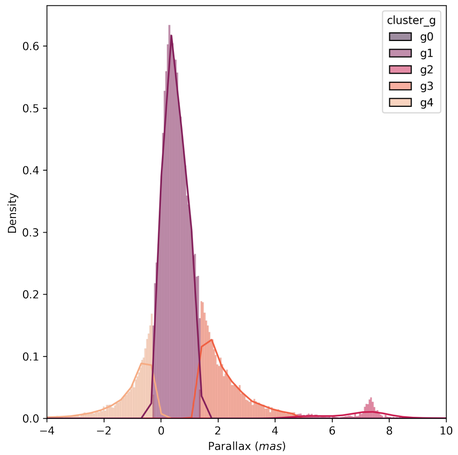
K-Means



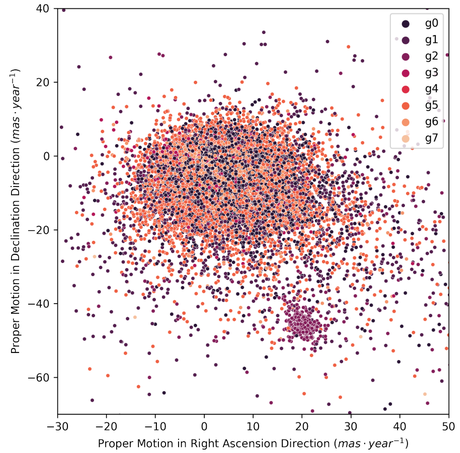


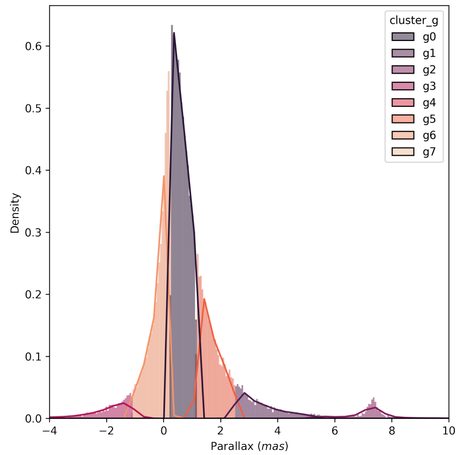
Melotte 22 - *g0* | K-Means *N=2*





Melotte 22 - *g2* | K-Means *N=5*





Melotte 22 - *g2* | K-Means *N=8*

Development

Deep Embedded Clustering (DEC)

The *Unsupervised Deep Embedding for Clustering Analysis* model or DEC (Xie, Girshick, & Farhadi, 2016) is a clustering model composed by a deep autoencoder and a clustering layer.

Unsupervised Deep Embedding for Clustering Analysis

Junyuan Xie JXIE@CS.WASHINGTON.EDU
University of Washington

Ross Girshick RBG@FB.COM
Facebook AI Research (FAIR)

Ali Farhadi ALI@CS.WASHINGTON.EDU
University of Washington

Development

Deep Embedded Clustering (DEC)

The *deep autoencoder is a neural network composed of encoder layers followed by decoder layers* that learns to copy its input to its output.

It *transforms the input data into a latent space* using a nonlinear mapping function $f_\theta : X \rightarrow Z$.

Although the number of features we deal is not too large, *this latent space helps us to start with a reduced number of features and avoids the "curse of dimensionality"*.

Development

Deep Embedded Clustering (DEC)

Once the autoencoder is trained and the data has been transformed, *a K-Means clusterer is used to make an initial clustering*. *K-Means 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 to it*.

This unsupervised algorithm allows us to improve the clustering.

Development

Features selection

Proper motion in right ascension and declination seems a natural choice since stars that belong to the same OC share a common motion vector.

Parallax is another important feature, since all stars within an OC were born from the same dust cloud, they must all have similar parallax.

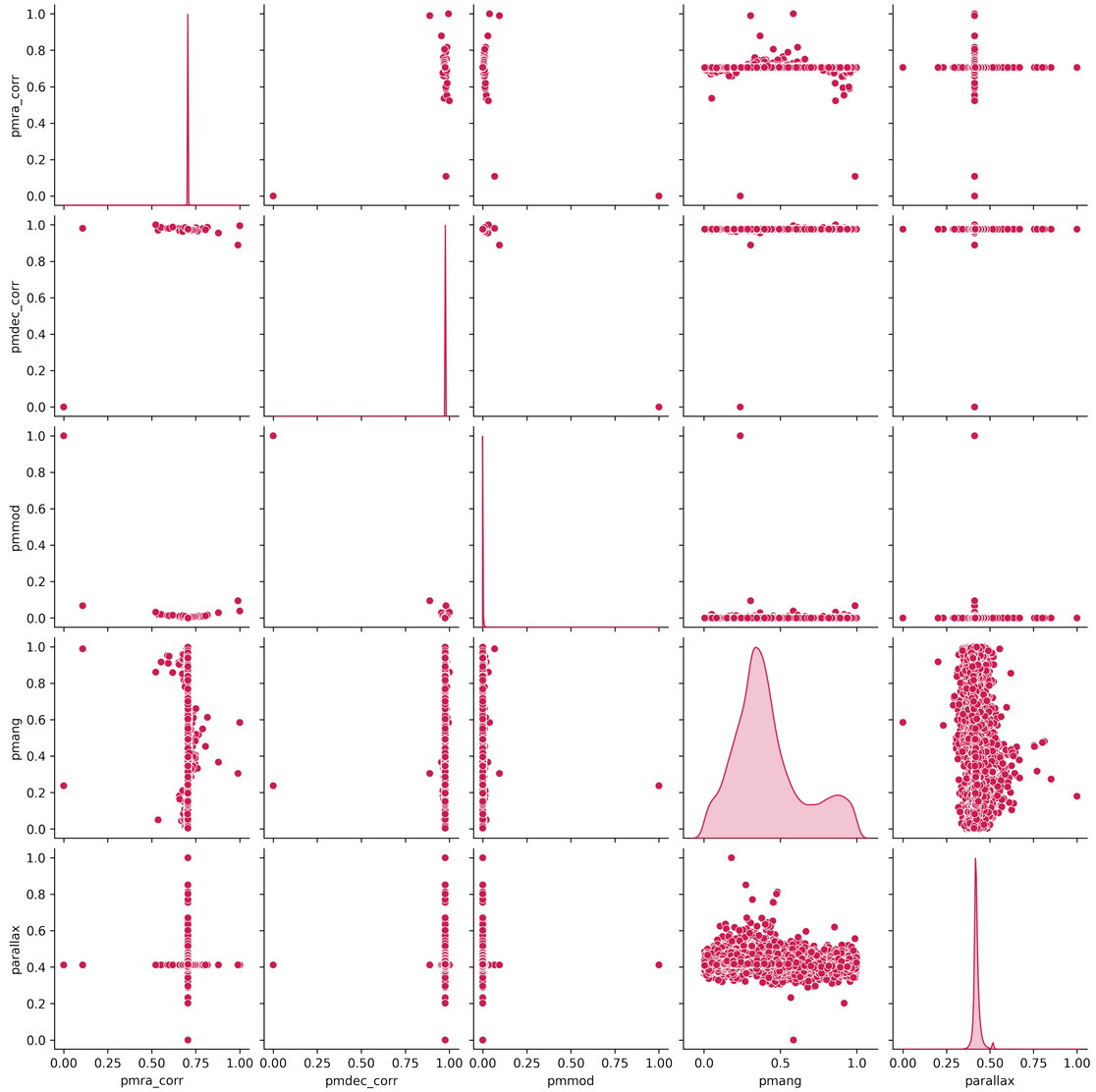
Auxiliary, *G-band mean magnitude* and *BP-RP color* are used to generate the HR-diagram to identify the OC among the obtained groups. So, we do not provide these variables to our model.

Development

Features selection

Features	Units	Description
pmra	Angular Velocity [mas/year]	Proper motion in right ascension
pmdec	Angular Velocity [mas/year]	Proper motion in declination
parallax	Angle [mas]	Parallax
pmra_corr	[year^{-1}]	pmra / parallax
pmdec_corr	[year^{-1}]	pmdec / parallax
pmmod	Angular Velocity [mas/year]	sqrt[pmra_corr^2 + pmdec_corr^2]
pmang	radians	arctan[pmdec_corr / pmra_corr]
phot_g_mean_mag	Magnitude [mag]	G-band mean magnitude
bp_rp	Magnitude [mag]	BP-RP color

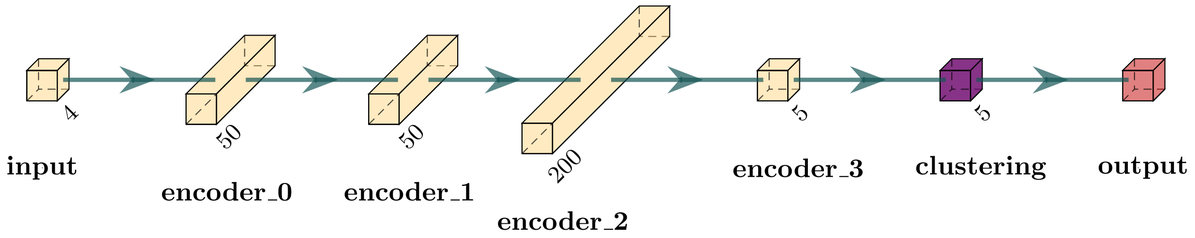
RAW features for our model | *Computed features* | *Auxiliary features*



Pairwise relationships among variables using Melotte 22 data

Development

DEC for Open Cluster Characterization (DECOCC)



DECOCC is an adaptation of DEC model for open cluster characterization.

Since the number of features in Unsupervised Deep Embedding for Clustering Analysis work is much greater than in ours, ***we have reduced the size of the deep autoencoder and the clustering layer***.

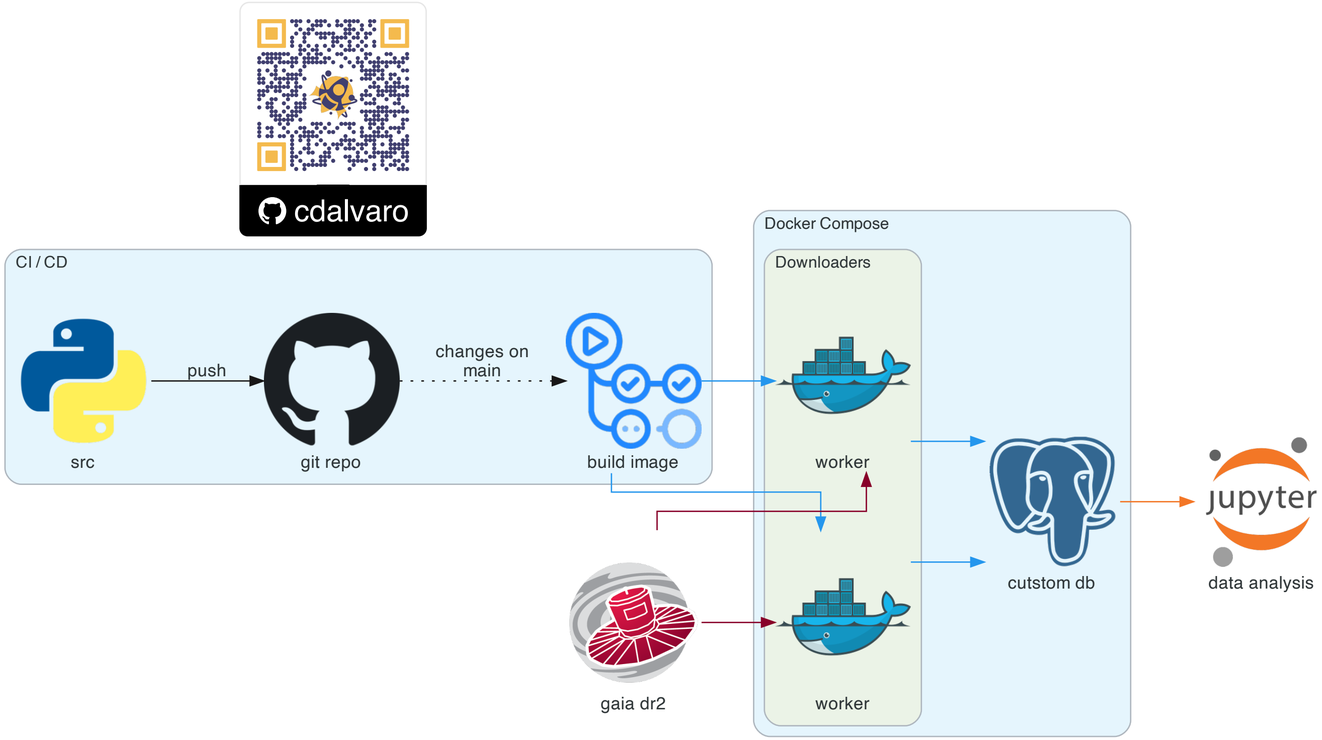
Also, ***we added a final step where those stars which fall outside the tails of the normal distribution of the selected group*** (below and above the 0.10 and 0.90 quartiles respectively) ***are filtered***.

Development

Procedure

Our *source code is hosted at GitHub*.

When a pull request is accepted, a GitHub workflow is triggered to build a new *Docker image with our custom PostgreSQL database and our Python model*.

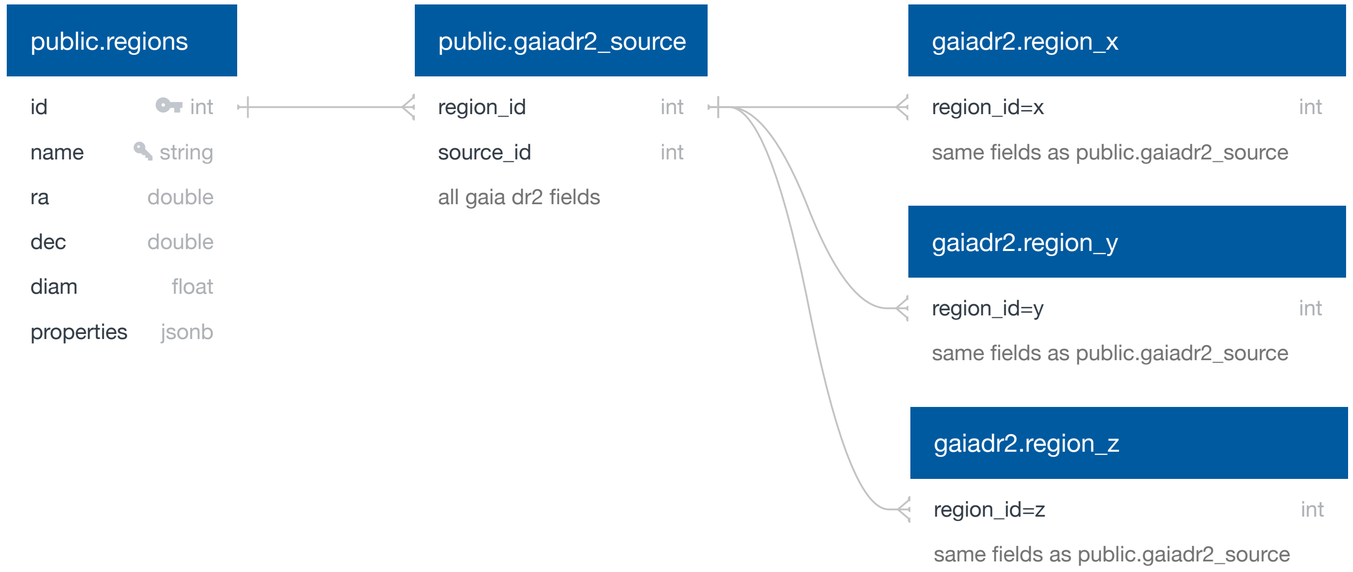


GitHub: [cdalvaro/machine-learning-master-thesis]

Development

Procedure

We use this ***custom db*** to store downloaded data from the Gaia DR2 dataset in order ***to have quick access to the most important data for our work***.

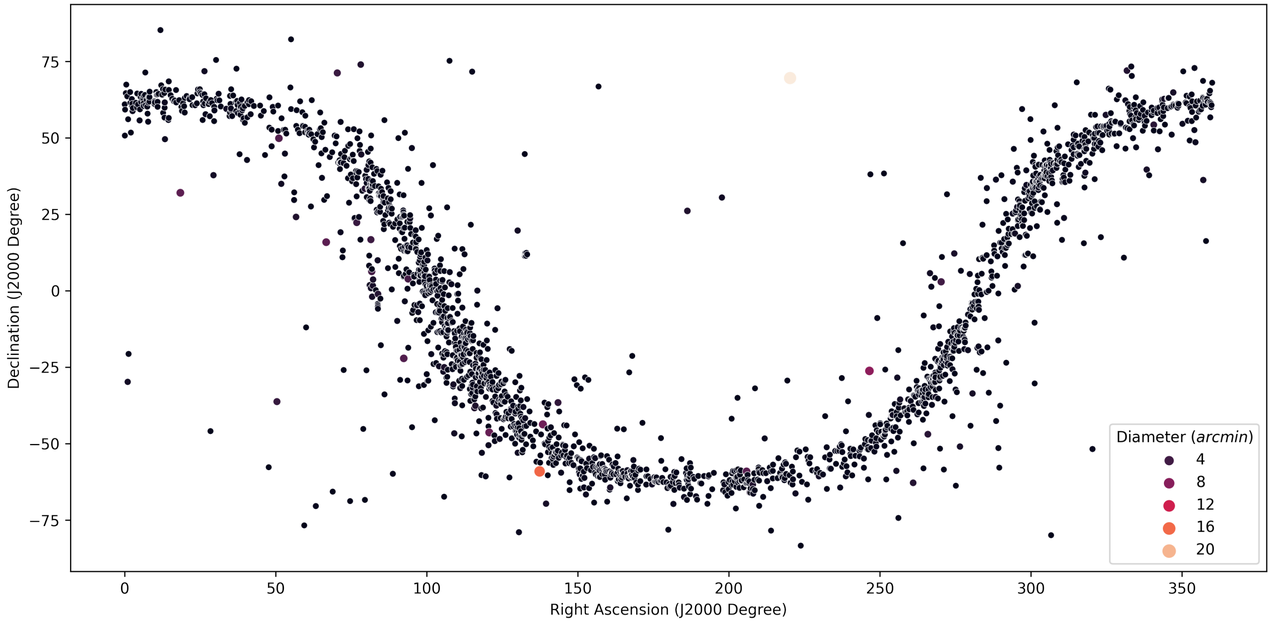


Development

Procedure

Gaia DR2 dataset has more than 1.7 billion stellar sources.

We have used the OpenClust catalogue as a region selector to download stars data from Gaia. However, the reduced selection is still too large. ***Around 114 million stars*** (~42GB of compressed data).



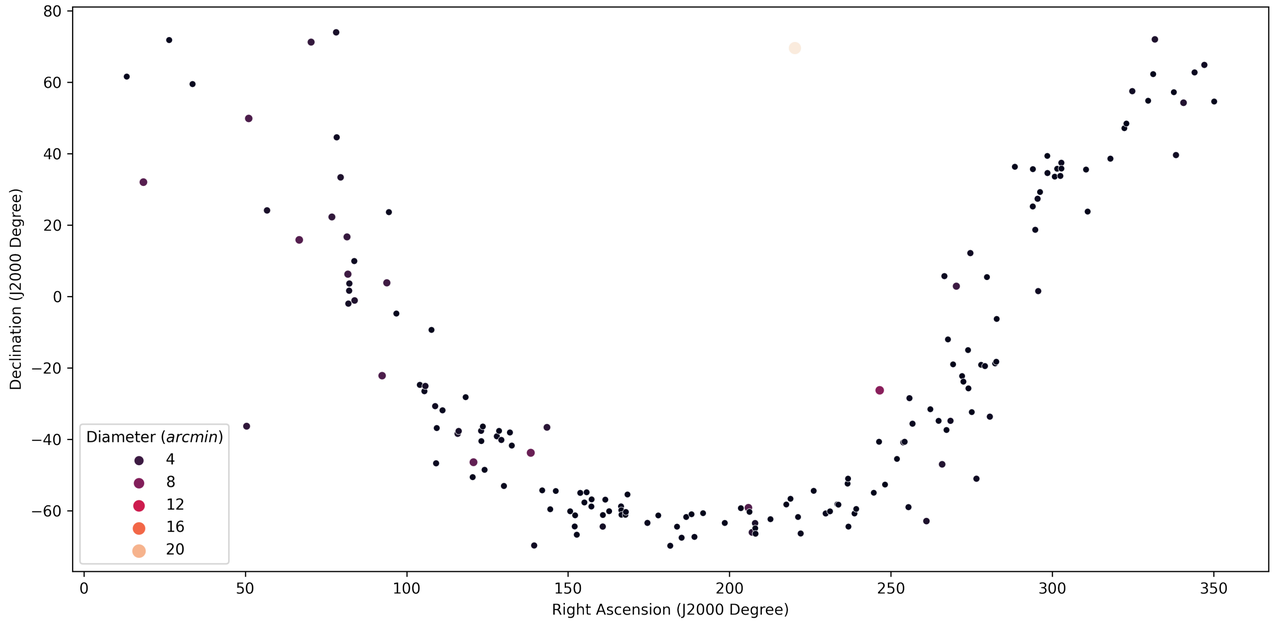
Known open cluster centers available at ***OpenClust*** catalogue
https://heasarc.gsfc.nasa.gov/W3Browse/star-catalog/openclust.html

Development

Procedure

We applied a series of filters in order to ***select only those cluster regions with good quality data***:

- Cluster diameter above 25.0 arcmin
- Parallax absolute value greater than 0.0
- All stars in the studied region fully identified (i.e.: without null values)



Open cluster centers after applying our filters
169 open clusters | ~75 million stars

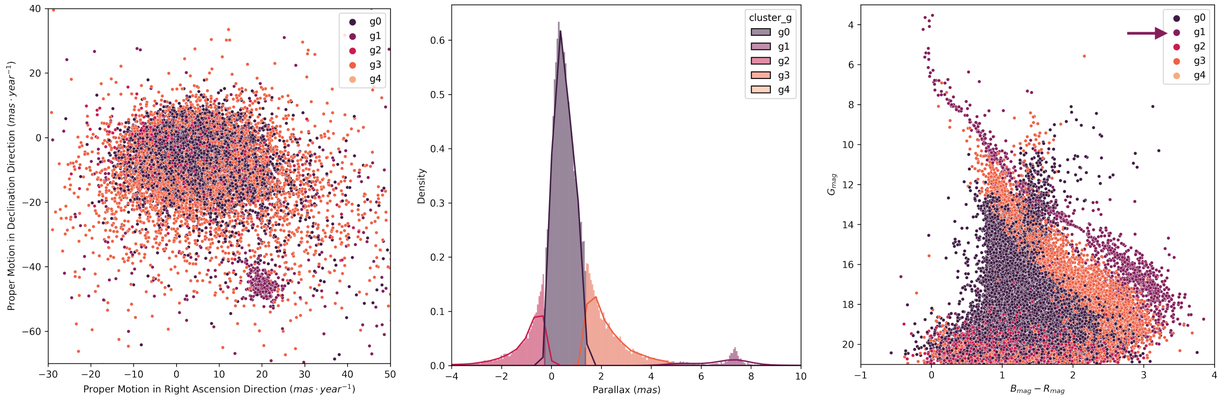
Results

Melotte 22

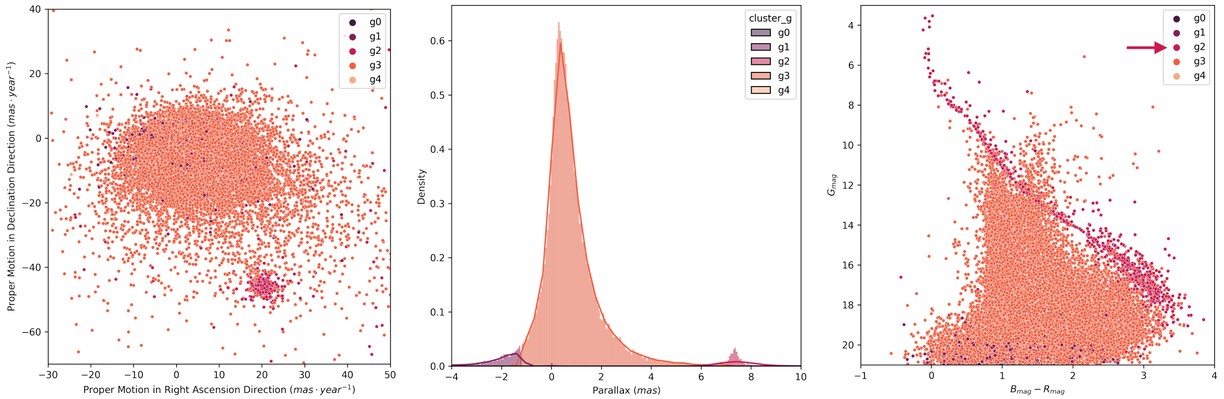
α J2000 (degrees)	δ J2000 (degrees)	Radius (arcmin)	# stars in region
56.75	24.117	60	61,552

Method	μ_α (mas·yr^{-1})	μ_δ (mas·yr^{-1})	ω (mas)	# stars
Simbad	20.0 ± 0.1	-45.6 ± 0.1	7.36 ± 0.01	1326
Clusterix+ TOPCAT	20.0 ± 1.3	-45.8 ± 1.5	7.33 ± 0.21	634
K-Means	20.3 ± 1.0	-38.0 ± 1.1	7.23 ± 0.06	1378
DECOCC	23.7 ± 1.3	-46.2 ± 1.5	8.04 ± 0.09	878
DECOCC (filtered)	19.5 ± 0.4	-44.2 ± 0.4	7.42 ± 0.01	438

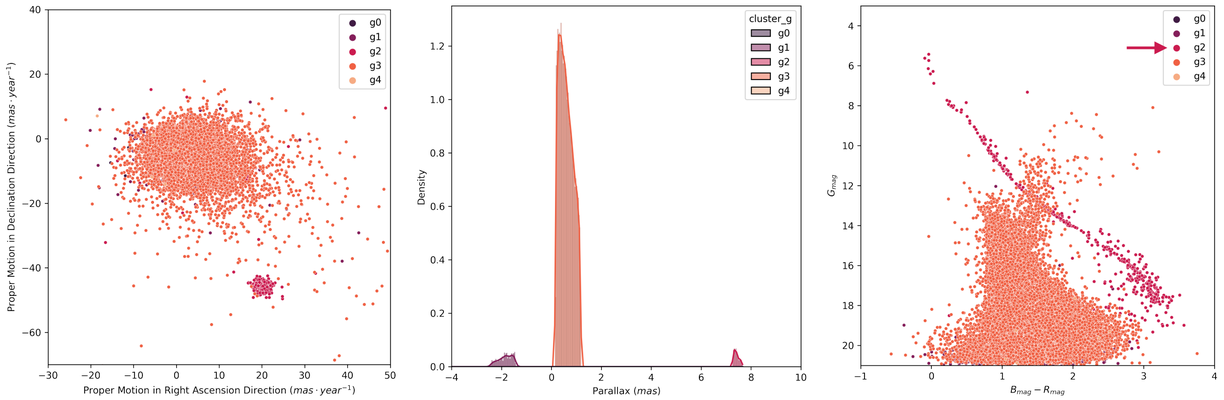
K-Means



DECOCC



DECOCC (filtered)



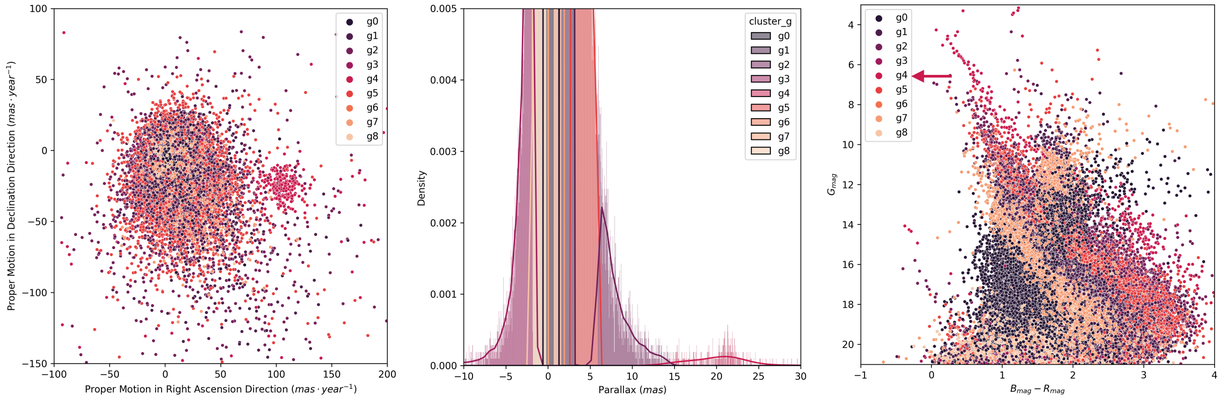
Results

Melotte 25

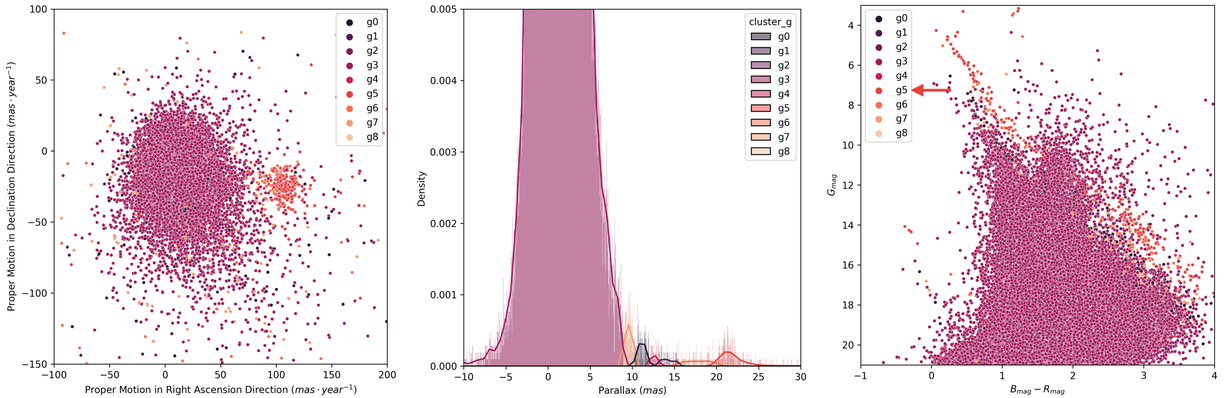
α J2000 (degrees)	δ J2000 (degrees)	Radius (arcmin)	# stars in region
66.725	15.867	165	433,996

Method	μ_α (mas·yr^{-1})	μ_δ (mas·yr^{-1})	ω (mas)	# stars
Simbad	104.9 ± 0.1	-28.00 ± 0.09	21.05 ± 0.07	-
Clusterix+ TOPCAT	106.8 ± 6.2	-24.9 ± 5.4	21.2 ± 1.1	109
K-Means	79.9 ± 3.7	-45.4 ± 3.8	20.9 ± 0.3	374
DECOCC	104.1 ± 3.2	-33.4 ± 2.5	22.1 ± 0.1	219
DECOCC (filtered)	106.0 ± 3.5	-30.0 ± 2.4	21.74 ± 0.07	175

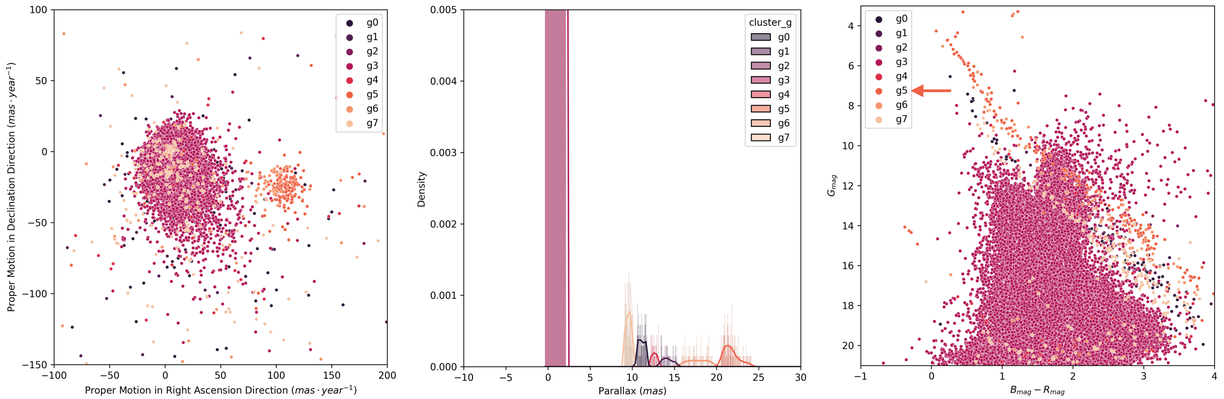
K-Means



DECOCC



DECOCC (filtered)



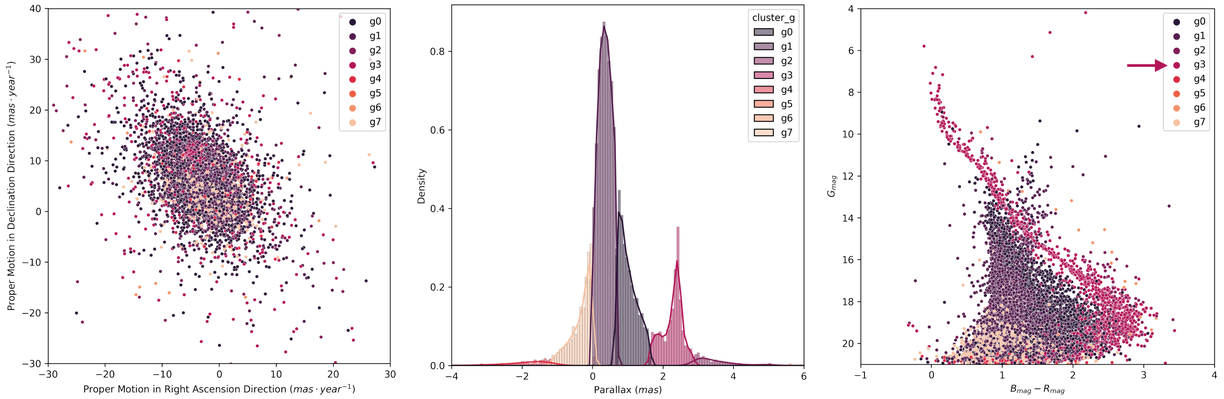
Results

NGC 2516

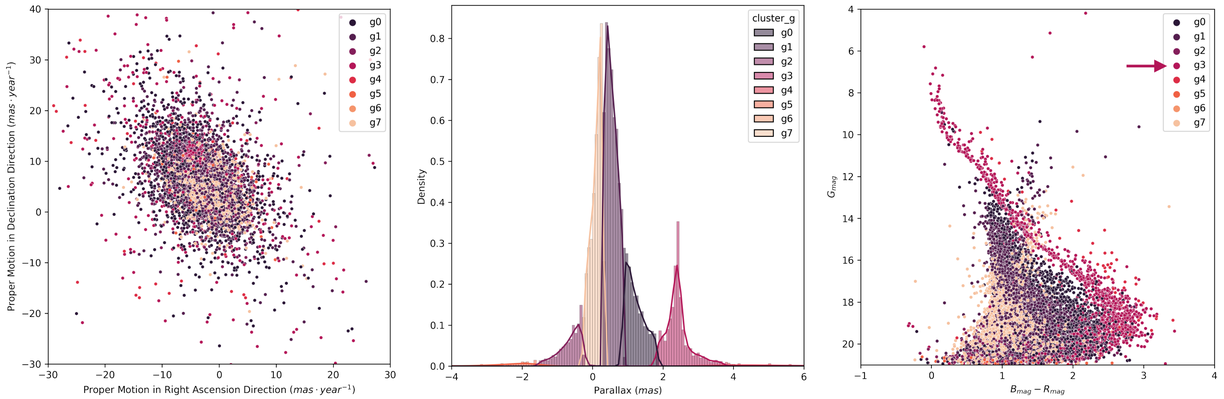
α J2000 (degrees)	δ J2000 (degrees)	Radius (arcmin)	# stars in region
119.517	-60.753	15	12,869

Method	μ_α (mas·yr^{-1})	μ_δ (mas·yr^{-1})	ω (mas)	# stars
Simbad	-4.66 ± 0.01	11.15 ± 0.01	2.4118 ± 0.0006	1727
Clusterix+ TOPCAT	-4.7 ± 0.5	11.2 ± 0.5	2.4 ± 0.1	638
K-Means	-4.3 ± 0.1	9.5 ± 0.2	2.23 ± 0.01	1542
DECOCC	-4.4 ± 0.2	10.0 ± 0.2	2.44 ± 0.01	1532
DECOCC (filtered)	-4.5 ± 0.1	10.1 ± 0.2	2.392 ± 0.004	1072

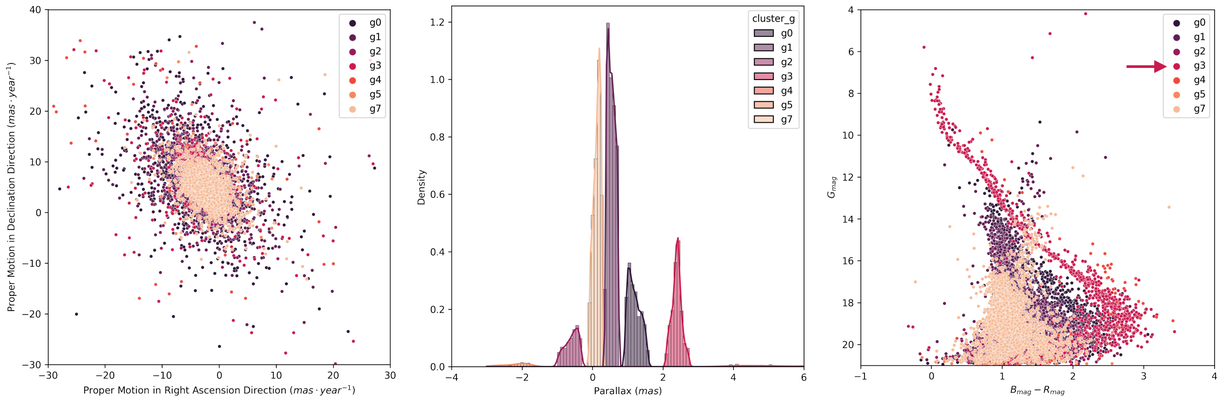
K-Means



DECOCC



DECOCC (filtered)



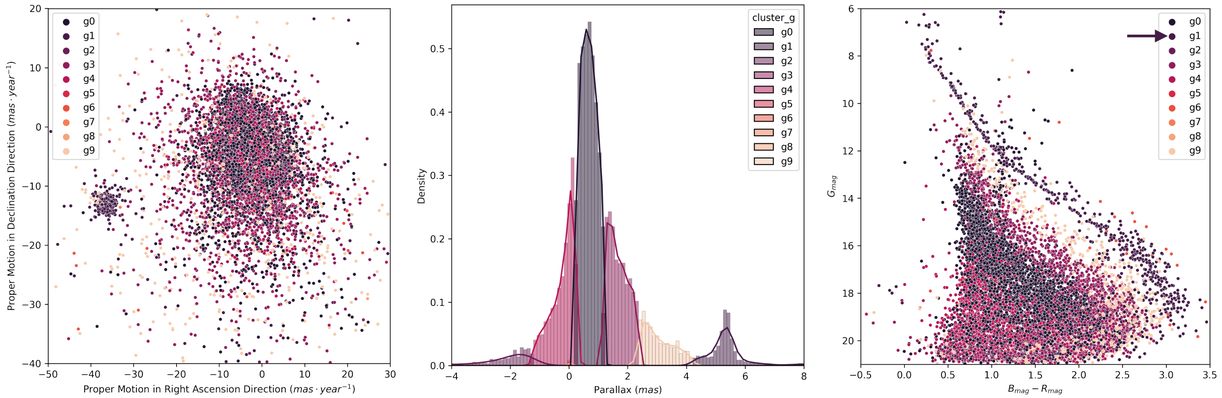
Results

NGC 2632

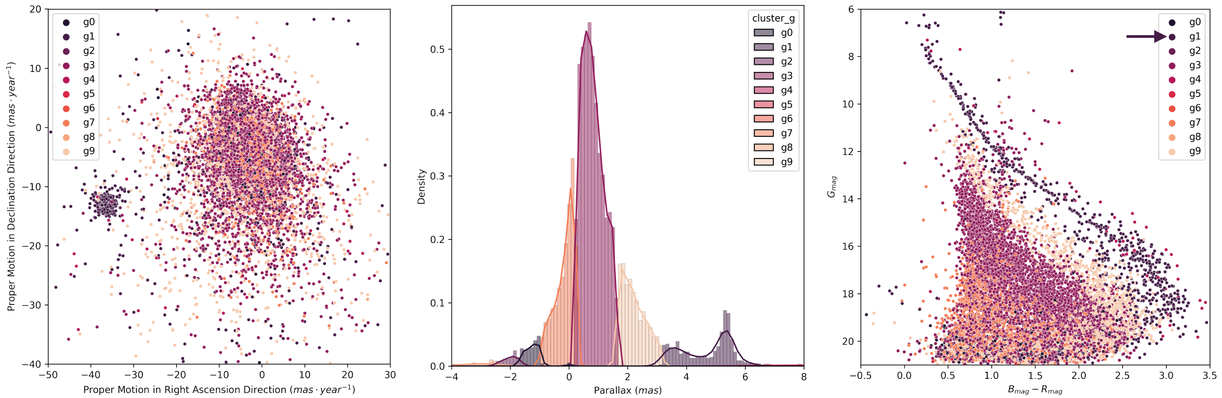
α J2000 (degrees)	δ J2000 (degrees)	Radius (arcmin)	# stars in region
130.1	19.667	35	13,167

Method	μ_α (mas·yr^{-1})	μ_δ (mas·yr^{-1})	ω (mas)	# stars
Simbad	-36.05 ± 0.1	-12.92 ± 0.07	5.371 ± 0.003	-
Clusterix+ TOPCAT	-36.2 ± 1.0	-12.9 ± 0.8	5.3 ± 0.2	371
K-Means	-26.4 ± 0.8	-15.8 ± 0.8	5.39 ± 0.03	629
DECOCC	-20.0 ± 0.7	-14.7 ± 0.6	4.69 ± 0.03	894
DECOCC (filtered)	-21.6 ± 0.7	-14.2 ± 0.6	4.72 ± 0.03	714

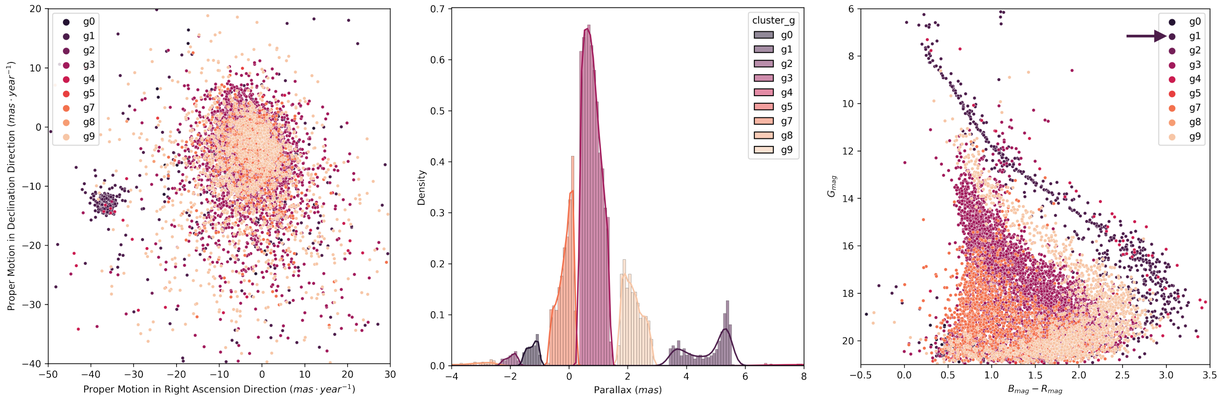
K-Means



DECOCC



DECOCC (filtered)



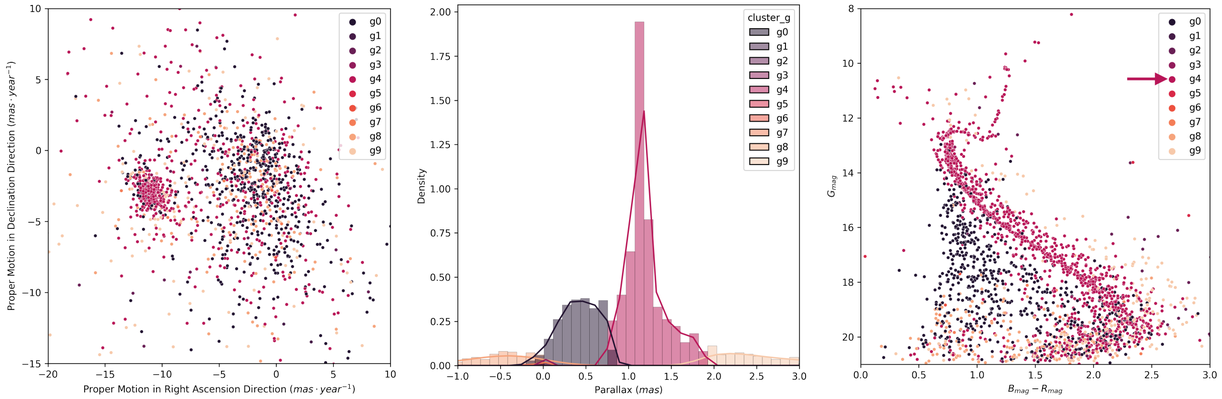
Results

NGC 2682

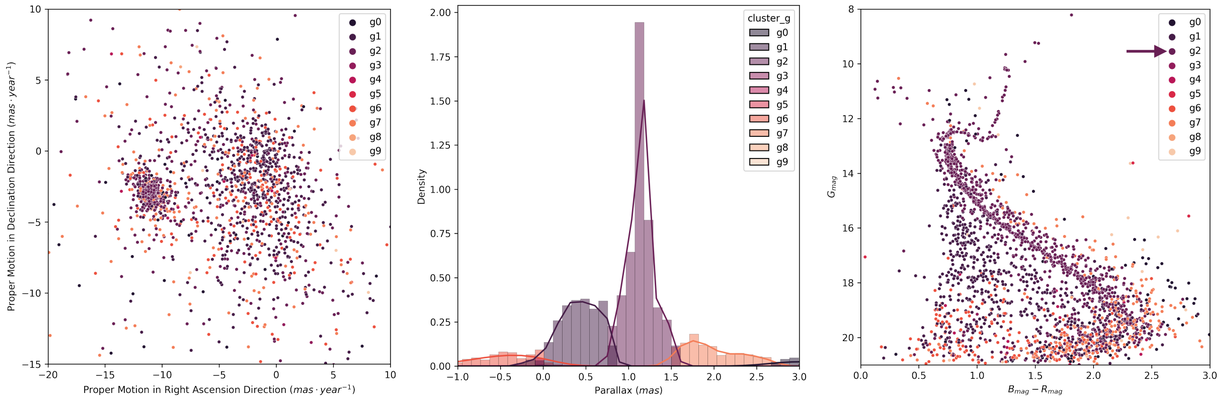
α J2000 (degrees)	δ J2000 (degrees)	Radius (arcmin)	# stars in region
132.825	11.8	12.5	2,839

Method	μ_α (mas·yr^{-1})	μ_δ (mas·yr^{-1})	ω (mas)	# stars
Simbad	-10.974 ± 0.006	-2.940 ± 0.006	1.133 ± 0.001	1194
Clusterix+ TOPCAT	-11.0 ± 0.3	-3.0 ± 0.3	1.14 ± 0.08	649
K-Means	-8.6 ± 0.2	-3.7 ± 0.2	1.20 ± 0.01	1374
DECOCC	-8.9 ± 0.2	-3.6 ± 0.2	1.144 ± 0.005	1238
DECOCC (filtered)	-9.6 ± 0.1	-3.3 ± 0.1	1.140 ± 0.003	990

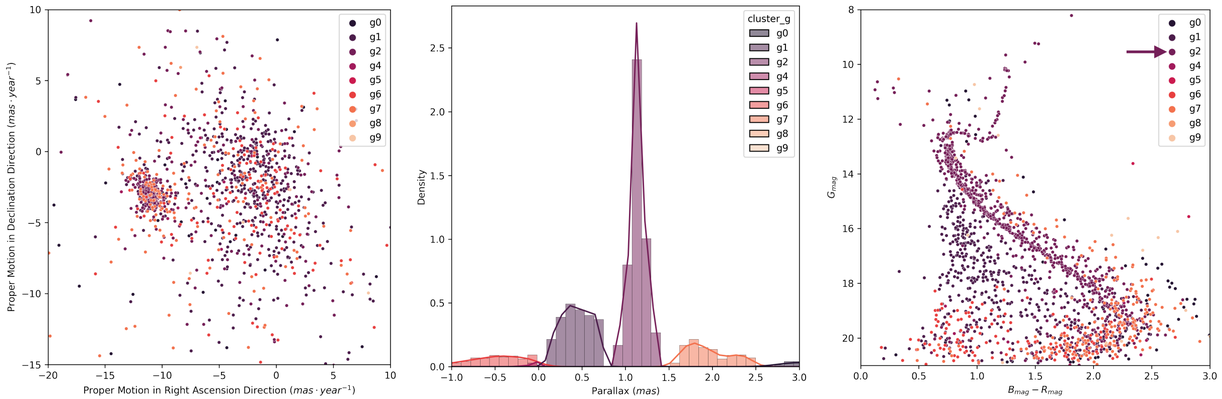
K-Means



DECOCC



DECOCC (filtered)



Results

Summary

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

- *NGC 2516* has it *proper motion center deviated from the origin* but it is *embedded inside a big cloud of stars* with similar proper motions.
- *NGC 2632* and *Melotte 22* is a cluster whose *proper motion center is not located at the origin* and *has a well separated parallax center*.
- *NGC 2682* 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*, are *clusters closer to us than the other*. That makes its *membership stars to be more scattered* than previous clusters which are more compact.

Results

Summary

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***.

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

Conclusions

We have succeeded developing a model that works fine identifying and characterizing open clusters.

The model ***is non-parameterized and unsupervised***.

It does not require complex hardware. Data analysis was performed on an *Apple Mac Pro Late 2013* with a *2.7GHz 12-Core Intel Xeon*, *64GB RAM* and *two graphic cards AMD FirePro D700* with 6GB.

Our model can deal with large and open regions. As an example, it has succeeded in characterizing Melotte 25 (Hiades) OC.

However, ***some improvements should be done in order to improve its accuracy and precision***.

Future Work

The initializer kernel prepares data before passing them to the ANN. This is one of the reasons because the *results may vary significantly for the same dataset among different runs*. In future works, we want to *avoid it to have a reliable model which does not depend on the dataset order*.

We also want to *test our model with a wider range of clusters*. In general, we could think of applying it to the entire OpenClust catalogue. So, we could *get a better idea about the current limitations of our model and improve it*.

Future Work

Another possible improvements are either to *use DBSCAN instead of K-Means as our initial clustering algorithm*, and *employ DR3 dataset instead of DR2 dataset of Gaia*.

An important aspect we noticed is that *with our model new uncatalogued clusters arise from the data apart of the main ones*. Making it of special interest to *study these individual new clusters*.

Future Work

European Astronomical Society Annual meeting 2021

Mar 1, 2021

ID 1193
Dear Cesar Augusto Guzman Alvarez,

Thank you for your abstract submission for the EAS 2021 virtual meeting.
Your submitted abstract details are as follows:

Abstract Submission

Title	Deep Embedded Clustering for Open Cluster Characterization with Gaia DR2 Data
Paper Status	Submitted
Presentation Type	Contributed talk
Theme	Data science & Archives
Session	SS32 : Machine learning and visualisation in data intensive era
Presenting Author	Carlos David Alvaro Yunta Affiliations: Meteologica
Co-Author	Cesar Augusto Guzman Alvarez Affiliations: Iac, UNAD
Co-Author	Joaquín Álvaro Contreras Affiliations: Nodo Ibérico de Europlanet, SEA, Federación de Asociaciones Astronómicas de España - FAAE

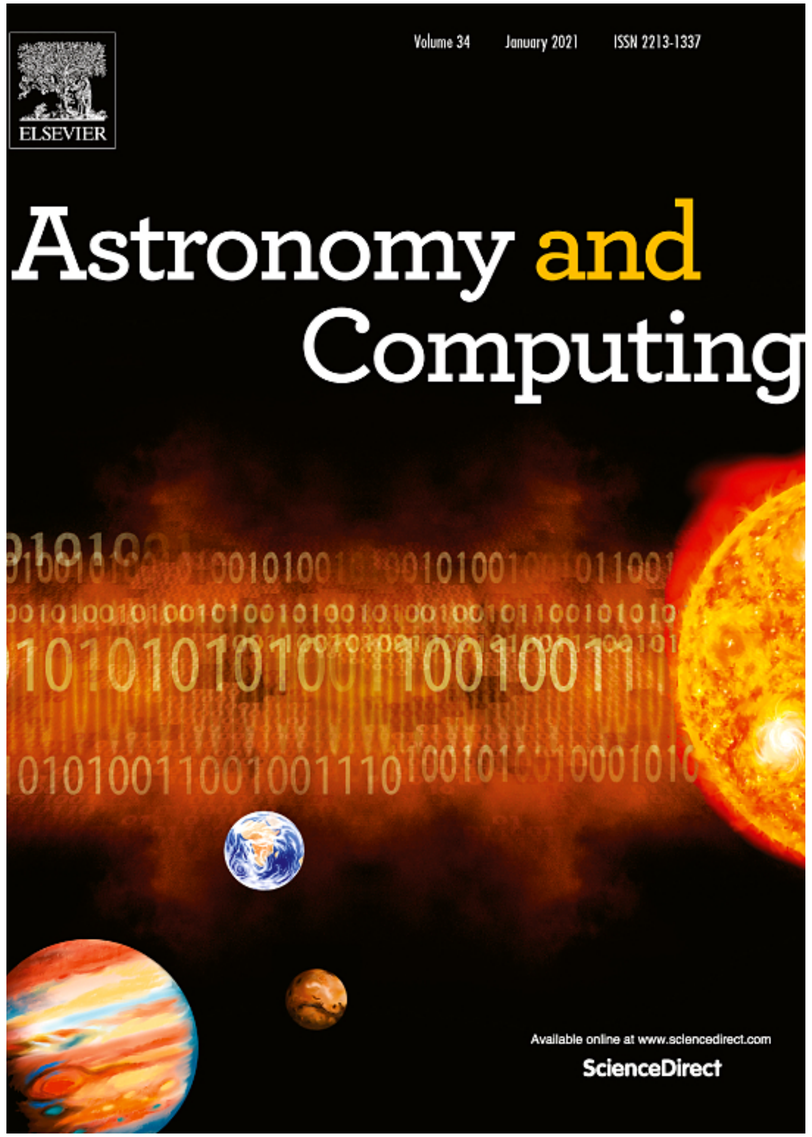
Future Work

Astronomy and Computing

Deep Embedded Clustering for Open Cluster Characterization with Gaia DR2 Data

--Manuscript Draft--

Manuscript Number:	
Article Type:	Full Length Article
Section/Category:	Astrostatistics and machine learning
Keywords:	characterization; data analysis; deep embedded clustering; gaia; machine learning; open cluster
Corresponding Author:	Cesar Augusto Guzman Alvarez, Ph.D. IAC: Instituto de Astrofisica de Canarias La Laguna, La Laguna SPAIN
First Author:	Carlos Alvaro Yunta
Order of Authors:	Carlos Alvaro Yunta
	Cesar Augusto Guzman Alvarez, Ph.D.
	Joaquín Alvaro Contreras
Abstract:	Characterize and understand 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 \emph{Artificial Neural Networks} (ANNs). More specifically, we adapted a state of the art model, the \emph{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 \emph{Gaia DR2 database} as the data source, and compared our model with the clustering technique \emph{K-Means}. Our method achieves good results, becoming even better (in some of the cases) than current techniques.
Suggested Reviewers:	Pablo Moreno Ger pablo.moreno@unir.net
	Juan Fdez-Olivares faro@decsai.ugr.es
	Laura Sebastia lsebastia@dsic.upv.es
	Adrian Colomer Granero adrian.colomer.g@campusviu.es



unir

LA UNIVERSIDAD
EN INTERNET

gaia

www.unir.net