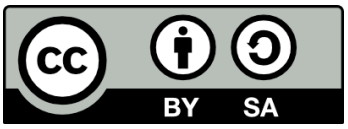
UNIVERSITAT DE BARCELONA

Detection, characterisation and use of open clusters in a Galactic context in a Big Data environment

Alfred Castro Ginard



Aquesta tesi doctoral està subjecta a la llicència ***Reconeixement- CompartIgual 4.0. Espanya de Creative Commons***.

Esta tesis doctoral está sujeta a la licencia ***Reconocimiento - CompartirIgual 4.0. España de Creative* Commons**.

This doctoral thesis is licensed under the ***Creative Commons Attribution-ShareAlike 4.0. Spain License***.

DETECTION, CHARACTERISATION AND USE OF OPEN CLUSTERS IN A GALACTIC CONTEXT IN A BIG DATA ENVIRONMENT

ALFRED CASTRO GINARD

DIRECTORS:
Dr Xavier Luri
Dr Carme Jordi

UNIVERSITAT DE BARCELONA

Departament de Física Quàntica i Astrofísica
Facultat de Física
Universitat de Barcelona

Alfred Castro Ginard, *Detection, characterisation and use of open clusters in a Galactic context in a Big Data environment* ,
PhD Thesis
Barcelona, 24 Febrer 2021

UNIVERSITAT DE BARCELONA
DEPARTAMENT DE FÍSICA QUÀNTICA I ASTROFÍSICA
Programa de doctorat en física
Línia de recerca en astronomia i astrofísica

Detection, characterisation and use of open clusters in a Galactic context in a Big Data environment

Memòria presentada per
Alfred Castro Ginard
per optar al grau de
Doctor en física per la Universitat de Barcelona

Directors de la tesi:
Dr Xavier Luri
Dr Carme Jordi
Tutor de la tesi:
Dr Alberto Manrique

Barcelona, 24 Febrer 2021

UNIVERSITAT DE BARCELONA

DECLARATION

This thesis is presented following the regulations of the University of Barcelona (Approved by the CdG in the session of the 16th of March of 2012 and modified by the CdG on the 9th of May and 19th of July of 2012, 29th of May and 3rd of October of 2013, 17th of July of 2014, 16th of July of 2015, 15th of June and 21st of November of 2016, 5th of December of 2017, 4th of May of 2018, 15th of May and 22nd of July of 2019 and 7th of October of 2020). The listed regulations allow for the presentation of a PhD thesis as a "compendia of published articles". According to the regulations, the thesis must contain a minimum of three published or accepted articles. This thesis contains the published version of five articles, which is sufficient to allow its presentation. It also contains one additional article, submitted but not yet accepted for publication at the moment of thesis submission.

Barcelona, 24 Febrer 2021

Alfred Castro Ginard

Als meus pares, en Jaume i na Catalina.
I a la meva germana, na Marta.

ACKNOWLEDGMENTS

Vull començar per agrair als meus directors, els Dr Xavier Luri i Dr Carme Jordi, per guiar-me durant aquests anys. Al Xavi, per donar-me l'oportunitat de començar un doctorat i confiar en mi des del principi. I a la Carme per haver-me acollit com a estudiant i tenir paciència amb mi. Sempre diré amb orgull qui van ser els meus directors. També vull agrair al Francesc Julbe, amb qui vaig aprendre molt al principi de la tesi i em va donar els fonaments que necessitava per continuar endavant.

També vull agrair a tot el grup *Gaia* de la UB, tant científics: Cesca, Lola, Teresa, Mercè, Eduard, Josep Manel, Maria, Roger, Tristan, Friedrich; com enginyers: Jordi Portell, Javi Castañeda, Sergio Soria. Vull mencionar apart al Dani Molina, que sempre ha tengut bones paraules i solucions pels meus problemes informàtics. També mereix un agraïment apart el JR, per evitar-me mals de cap a la hora de fer papers.

This thesis would not have been possible without the collaboration of many other people. I want to thank Dr Paul McMillan, for hosting me at Lund and introduce me to the world of the dynamics, where I have a lot to learn. Also to Paul and all the people in Lund, who worried for me and help me get back home when the pandemic started. I also want to thank Dr Alberto Krone-Martins and the people from COIN, who gave me the opportunity to learn a lot of the science done outside *Gaia* while visiting Chamonix. También quiero agradecer al Dr Octavio Valenzuela y Dr Luis Aguilar, por su contagioso entusiasmo a la hora de hablar sobre dinámica, y sobretodo por la oportunidad de disfrutar de México, donde conocí gente maravillosa.

Vull agrair especialment a tots els que m'han acompanyat en aquest camí i l'han fet més cómode. A en Dani, per oferir-me la seva sincera amistat des de que vaig arribar, i sempre tenir temps per mi. A en Pau, el meu primer (i darrer) company de pis, on hem passat moments molt bons. A n'Edgar, per separar el llit 5 cm. A en Nico, per sempre estar disposat a sortir a dinar, sopar i el que calgui. A na Núria Torres, que la trobem molt a faltar al departament. A n'Oscar, que ha aportat vida nova al departament. I a tota la resta d'estudiants i postdocs amb qui he compartit aquests anys: Lluís, José Luis, David, Sam, Ali, Katie, i tota la resta de gent que segur m'he deixat.

També vull agrair al Santi, Victor i Ignasi, un grup de *joves* amb els que he descobert parts de Catalunya i costums muntanyenques. I també a la Núria Miret i l'Aleix, per acompanyar-nos a veure món.

Vull agrair molt especialment als meus amics de sempre, en Toni, Tomeu, Ivan, Miquel Àngel i Álvaro per mantenir-me amb un peu a

les meves arrels, i ser els únics que em poden donar suport i consell sense haver d'escollir les paraules.

També agrair als meus pares i la meva germana, en Jaume, na Catalina i na Marta, a qui va dedicada aquesta tesi. Ells m'han donat tota la llibertat que he necessitat per aprendre tot per jo mateix, i sempre han estat allà quan les coses no han anat bé. Vos estim.

Finalment, a na Laia, que ha acabat acompanyant-me a la vida. En tu sempre he trobat les forces per arribar aquí. Gràcies.

ABSTRACT

Context. Open clusters are groups of gravitationally bound stars, that were born from the same gas molecular cloud and, thus, share similar position, kinematics, age and chemical composition. Traditional methods to detect open clusters rely in the human-assisted inspection of regions of the sky to look for positional overdensities of stars, which then are checked to follow an isochrone pattern in a color-magnitude diagram. The publication of the second *Gaia* data release, with more than 1.3 billion stars with parallax and proper motion measurements together with mean photometry in three broad bands, boosted the development of novel machine learning-based techniques to automatise the search for open clusters, using both the astrometric and photometric information.

Open clusters are popular tracers of properties of the Galactic disc such as the structure and evolution of the spiral arms, or testbed for stellar evolution studies, because their astrophysical parameters are estimated with a greater precision than for field stars. Therefore, a good understanding of the open cluster population in the Milky Way is key for Galactic archaeology studies.

Aims. Our aim for this thesis is to transform classical methodologies to detect different kinds of patterns from astronomical data, that mostly relies on human-assisted inspection, to an automatic data mining procedure to extract meaningful information from stellar catalogues. We also aim to use the result of the application of machine learning techniques to *Gaia* data, in a broad Galactic context.

Methods. We have developed a data mining methodology to blindly search for open clusters in the Galactic disc. First, we use a density-based clustering algorithm, `DBSCAN`, to search for overdensities in the five-dimensional astrometric parameter space in *Gaia* data $(l, b, \varpi, \mu_{\alpha^*}, \mu_\delta)$. The deployment of the clustering step in a Big Data environment, at the MareNostrum supercomputer located in the Barcelona Supercomputing Center, avoids computational constraints for the search. Second, the detected overdensities are classified into mere statistical or physical overdensities using an artificial neural network trained to recognise the isochrone pattern that open cluster member stars follow in a color-magnitude diagram.

We estimate astrophysical parameters such as ages, distances and line-of-sight extinctions for the whole open cluster population using an artificial neural network trained on well-known open clusters. We use this additional information, together with radial velocities gathered from different space-based and ground-based surveys, to trace the Galactic spiral present-day structure using `GaussianMixtureModels` to

associate the young ($\leq$ 30 Myr) open clusters to their mother spiral arms. We also analyse the spiral arms evolution during the last 80 Myr to provide new insights into the nature of the Milky Way spiral structure.

Finally, we use a machine learning pipeline to detect and characterise young stellar objects using a combination of data from different surveys. For this purpose, we first train a `RandomForest` algorithm to identify young stellar objects based on near-infrared photometry from combinations of measurements from the Spitzer Space Telescope with data from 2MASS, UKIDSS and VVV surveyss. The resulting young stellar objects, with astrometry available from the *Gaia* second data release, are grouped together using the `HDBSCAN` method, and these groups are associated to the Local, Sagittarius and Scutum arms as an example on the use of the catalogue.

Results. The automatization of the open cluster detection procedure, together with its deployment in a Big Data environment, has resulted in more than 650 new open clusters detected with this methodology. The new UBC clusters (named after the University of Barcelona) represent one third of the open cluster census (2017 objects with *Gaia* DR2 parameters), and it is the largest single contribution to the whole open cluster catalogue.

We are able to add 264 young open clusters ($\leq$ 30 Myr) to the 84 high-mass star forming regions traditionally used to trace spiral arms, to increase the Galactocentric azimuth range where the Milky Way spiral arms are defined, and better estimate their present-day parameters. By analysing the age distribution of the open clusters across the Galactic spiral arms, and computing the spiral arms pattern speeds following the open clusters orbits from their birth places, we are able to disfavour classical density waves as the main mechanism for the formation of the Milky Way spiral arms, favouring a transient behaviour.

The application of machine learning techniques to data in the near-infrared and infrared regimes, resulted in the detection of 117 446 young stellar objects (of which $\sim$ 90 000 are new identifications). This new young stellar object catalogue represents the largest homogeneous catalogue for the inner Galactic midplane.

Conclusions. The implementation of our methodology to search for unknown open clusters in the *Gaia* data, based on a Big Data environment, has shown to be efficient and prepared for future *Gaia* data releases as well as other large surveys. This thesis has shown that the use of machine learning, with a proper treatment of the computational resources, will play a key role in a data-dominated future for Astronomy.

RESUM EN CATALÀ

Aquesta tesi té com a objectiu la detecció i caracterització de cúmuls estel·lars oberts, i el seu ús en el context dels estudis de l'estructura de la nostra Galàxia. Pel gran volum de dades dels catàlegs astronòmics actuals, com ara el derivat de la missió espaial *Gaia*, la detecció d'aquests cúmuls es fa en un entorn de *Big Data*, cosa que facilita l'anàlisi de grans regions del cel minimitzant l'impacte de les limitacions computacionals.

Els cúmuls estel·lars oberts són conjunts d'estels, lligats gravitatòriament entre si, que van néixer del mateix núvol de gas molecular i, per tant, tenen posició, cinemàtica, edat i composició química similars. Aquests cúmuls són molt usats com a traçadors de les propietats del disc Galàctic, com per exemple l'estructura i l'evolució dels braços espirals, o actuen com a laboratoris per a estudis sobre l'evolució estel·lar, ja que les seves propietats astrofísiques poden estimar-se amb una major precisió que per a estels individuals. Així doncs, un bon coneixement de la població de cúmuls oberts a la Via Làctia és clau per als estudis d'arqueologia Galàctica.

Els mètodes tradicionals per a la detecció de cúmuls oberts es basen en la inspecció de regions particulars del cel cercant sobredensitats posicionals d'estels. Si aquests estels tenen relació física entre ells i no són meres sobredensitats estadístiques, segueixen un determinat patró en un diagrama de color magnitud (segueixen una isòcrona), i llavors poden ser considerats un cúmul obert. La publicació del segon arxiu de dades de *Gaia*, que conté més de 1 300 milions d'estels als quals s'han pogut mesurar paral·laxis i moviments propis a més de fotometria en tres filtres de banda ampla, impossibiliten els mètodes tradicionals degut al gran volum del catàleg. Per això, el desenvolupament de tècniques automàtiques, basades en *machine learning*, per detectar objectes com cúmuls oberts ha crescut juntament amb el volum dels catàlegs a analitzar.

L'objectiu d'aquesta tesi és transformar els mètodes tradicionals per detectar diferents tipus de patrons en dades astronòmiques, que majoritàriament es basen en una inspecció manual, en processos automàtics de mineria de dades per extreure informació rellevant dels catàlegs estel·lars. També tenim l'objectiu de fer servir el resultat de l'aplicació de les tècniques de *machine learning* sobre l'arxiu *Gaia*, en un context Galàctic ampli.

Durant aquesta tesis, hem desenvolupat una metodologia basada en mineria de dades per a la cerca a cegues de cúmuls oberts al disc Galàctic. Primer, hem utilitzat un algoritme de *clustering* basat en densitat, `DBSCAN`, per trobar sobredensitats en l'espai astromètric de cinc

dimensions en les dades de *Gaia* $(l, b, \varpi, \mu_{\alpha^*}, \mu_\delta)$. La implementació del mètode de *clustering* en un entorn de *Big Data*, en el nostre cas en el superordinador MareNostrum al Barcelona Supercomputing Center, ens permet cercar cúmuls oberts basant-nos en les seves propietats físiques i no estar restringits per limitacions computacionals. Després, les sobredensitats detectades es classifiquen en simples sobredensitats estadístiques o cúmuls oberts reals per mitjà d'una xarxa neuronal artificial entrenada per reconèixer isòcrones en un diagrama de color magnitud.

Per a tota la població de cúmuls oberts, hem pogut estimar les seves propietats físiques com distància, edat i extinció en la línia de visió, fent servir una xarxa neuronal artificial entrenada sobre cúmuls oberts ben caracteritzats. Hem fet servir aquesta informació addicional per a cada cúmul, juntament amb mesures de velocitat radial recollides de diferents catàlegs, per traçar l'estructura espiral actual de la nostra Galàxia mitjançant la tècnica dels `GaussianMixtureModels` per associar els cúmuls oberts més joves (menys de 30 milions d'anys) al braç espiral on s'han format. També hem analitzat l'evolució dels braços espirals de la Via Làctia durant els últims 80 milions d'anys, trobant nova informació sobre la seva naturalesa.

Finalment, hem desenvolupat un algoritme per detectar i caracteritzar objectes estel·lars joves fent servir dades combinades de diferents catàlegs. Per a això, hem entrenat un `RandomForest` per reconèixer aquests objectes basat en dades fotomètriques en l'infrarroig de combinacions de l'Spitzer Space Telescope amb els catàlegs de 2MASS, UKIDSS i VVV. Hem agrupat els objectes estel·lars joves resultants, amb astrometria de *Gaia* DR2, mitjançant l'algoritme `HDBSCAN` per associar-los als braços Local, Sagitari i Scutum, com a exemple d'aplicació del catàleg.

L'automatització del procediment de detecció de cúmuls oberts, juntament amb la seva implementació en un entorn de *Big Data*, ha resultat en més de 650 cúmuls nous detectats amb aquesta metodologia. Els nous cúmuls UBC (nomenats així per la Universitat de Barcelona) representen un terç de la població actualment coneguda de cúmuls oberts (2017 objectes amb paràmetres de *Gaia* DR2), i és la contribució individual més gran al catàleg global.

Hem pogut augmentar el nombre de traçadors dels braços espirals, afegint 264 cúmuls oberts joves (menys de 30 milions d'anys) a les 84 regions de formació estel·lar d'alta massa utilitzats tradicionalment. Això ens ha permès d'augmentar el rang en azimut Galactocèntric en el qual els braços estan definits, i estimar millor els paràmetres actuals d'aquests braços. Analitzant la distribució en edat dels cúmuls oberts dins dels braços espirals, i calculant la velocitat a la que aquests braços es mouen a partir de l'òrbita dels cúmuls oberts des del seu naixement, hem pogut qüestionar la teoria clàssica d'ona de densitat

com a mecanisme principal de formació de l'estructura espiral a la Via Làctia, afavorint un comportament transitori dels braços espirals.

L'aplicació de les metogologies de *machine learning* a dades en el rang infrarroig ha permès la detecció de 117 446 objectes estel·lars joves (dels quals $\sim$ 90 000 són noves identificacions). Aquest nou catàleg d'objectes estel·lars joves és el més gran i homogeni al pla Galàctic interior.

La implementació de la nostra metodologia per a la cerca de cúmuls oberts en les dades de *Gaia*, basada en un entorn de *Big Data*, ha demostrat ser eficient i estar preparada per a futurs lliuraments de dades de *Gaia*, així com per a altres grans catàlegs. Aquesta tesi ha mostrat que l'ús de *machine learning*, amb un correcte tractament dels recursos computacionals, té un gran camí per recórrer en un futur a l'Astronomia dominat per les dades.

PUBLICATIONS

Complete list of publications at the moment of thesis deposit.

Published articles in this thesis

A. Castro-Ginard et al. (2018). "A new method for unveiling open clusters in Gaia. New nearby open clusters confirmed by DR2." In: *Astronomy and Astrophysics* 618, A59, A59.

A. Castro-Ginard et al. (2019). "Hunting for open clusters in Gaia DR2: the Galactic anticentre." In: *Astronomy and Astrophysics* 627, A35, A35.

Cantat-Gaudin, T. et al. (2020). "Painting a portrait of the Galactic disc with its stellar clusters." In: *Astronomy and Astrophysics* 640, A1, A1.

Kuhn, M. et al. (2020). "SPICY: The Spitzer/IRAC Candidate YSO Catalog for the Inner Galactic Midplane." In: *arXiv e-prints*, arXiv:2011.12961.

A. Castro-Ginard et al. (2020). "Hunting for open clusters in Gaia DR2: 582 new open clusters in the Galactic disc." In: *Astronomy and Astrophysics* 635, A45, A45.

Other published articles

Cantat-Gaudin, T. et al. (2018). "A Gaia DR2 view of the open cluster population in the Milky Way." In: *Astronomy and Astrophysics* 618, A93, A93.

Luri, X. et al. (2018). "Gaia Data Release 2. Using Gaia parallaxes." In: *Astronomy and Astrophysics* 616, A9, A9.

Soubiran, C. et al. (2018). "Open cluster kinematics with Gaia DR2." In: *Astronomy and Astrophysics* 619, A155, A155.

Álvarez Cid-Fuentes, J. et al. (2019). "dislib: Large-scale High Performance Machine Learning in Python." In: *Proceedings of the 15th International Conference of eScience*, pp. 96–105.

Cantat-Gaudin, T. et al. (2019). "Gaia DR2 unravels incompleteness of nearby cluster population: new open clusters in the direction of Perseus." In: *Astronomy and Astrophysics* 624, A126, A126.

Romero-Gómez, M. et al. (2019). "Gaia kinematics reveal a complex lopsided and twisted Galactic disc warp." In: *Astronomy and Astrophysics* 627, A150, A150.

Anders, F. et al. (2020). "The star cluster age function in the Galactic disc with Gaia DR2: Fewer old clusters and a low cluster formation efficiency." In: *arXiv e-prints*, arXiv:2006.01690.

Ramos, P. et al. (2020). "Full 5D characterisation of the Sagittarius stream with Gaia DR2 RR Lyrae." In: *Astronomy and Astrophysics* 638, A104, A104.

Tarricq, Y. et al. (2020). "3D kinematics and age distribution of the Open Cluster population." In: *arXiv e-prints*, arXiv:2012.04017.

Collaboration articles

Collaboration, G. et al. (2018a). "Gaia Data Release 2. Kinematics of globular clusters and dwarf galaxies around the Milky Way." In: *Astronomy and Astrophysics* 616, A12, A12.

Collaboration, G. et al. (2018b). "Gaia Data Release 2. Mapping the Milky Way disc kinematics." In: *Astronomy and Astrophysics* 616, A11, A11.

Collaboration, G. et al. (2018c). "Gaia Data Release 2. Observational Hertzsprung-Russell diagrams." In: *Astronomy and Astrophysics* 616, A10, A10.

Collaboration, G. et al. (2018d). "Gaia Data Release 2. Observations of solar system objects." In: *Astronomy and Astrophysics* 616, A13, A13.

Collaboration, G. et al. (2018e). "Gaia Data Release 2. Summary of the contents and survey properties." In: *Astronomy and Astrophysics* 616, A1, A1.

Collaboration, G. et al. (2018f). "Gaia Data Release 2. The celestial reference frame (Gaia-CRF2)." In: *Astronomy and Astrophysics* 616, A14, A14.

Collaboration, G. et al. (2019). "Gaia Data Release 2. Variable stars in the colour-absolute magnitude diagram." In: *Astronomy and Astrophysics* 623, A110, A110.

Gaia Collaboration et al. (2020a). "Gaia Early Data Release 3: Acceleration of the solar system from Gaia astrometry." In: *arXiv e-prints*, arXiv:2012.02036, arXiv:2012.02036.

Gaia Collaboration et al. (2020b). "Gaia Early Data Release 3: Structure and properties of the Magellanic Clouds." In: *arXiv e-prints*, arXiv:2012.01771, arXiv:2012.01771.

Gaia Collaboration et al. (2020c). "Gaia Early Data Release 3: The Gaia Catalogue of Nearby Stars." In: *arXiv e-prints*, arXiv:2012.02061, arXiv:2012.02061.

Other articles submitted

A. Castro-Ginard et al. (2021). "On the transient nature of the Milky Way spiral arms from open clusters in *Gaia* DR2." In: *submitted to A&A*.

Conference proceedings

Cantat-Gaudin, T. et al. (2019). "A Gaia DR2 view of the open cluster population in the Milky Way." In: *Highlights on Spanish Astrophysics X*. Ed. by B. Montesinos et al., pp. 401–401.

Romero-Gomez, M. et al. (2019). "On the cutting edge of vertical motion: Bending waves and the Galactic warp." In: *The Gaia Universe*, p. 20.

Romero-Gómez, M. et al. (2019). "The complexity and richness of the Galactic disc velocity field unveiled by Gaia DR2." In: *Highlights on Spanish Astrophysics X*. Ed. by B. Montesinos et al., pp. 386–391.

A. Castro-Ginard (2019). "How complete is the open cluster census?" In: *The Gaia Universe*, p. 32.

A. Castro-Ginard et al. (2019). "Detection of new Open Clusters with Gaia." In: *Highlights on Spanish Astrophysics X*. Ed. by B. Montesinos et al., pp. 278–282.

Anders, F. et al. (2020). "Reanalysing the Galactic open-cluster population in light of Gaia DR2." In: *Contributions to the XIV.o Scientific Meeting (virtual) of the Spanish Astronomical Society*, p. 114.

Antoja, T. et al. (2020). "The Sagittarius stream with Gaia data." In: *Contributions to the XIV.o Scientific Meeting (virtual) of the Spanish Astronomical Society*, p. 117.

Ramos, P. et al. (2020). "The Halo-Disc dynamical coupling. Gaia blind detection of the Monoceros and ACS structures." In: *Contributions to the XIV.o Scientific Meeting (virtual) of the Spanish Astronomical Society*, p. 177.

Romero-Gómez, M. et al. (2020). "Dissecting the Galactic bar using Gaia observables and statistical techniques." In: *Contributions to the XIV.o Scientific Meeting (virtual) of the Spanish Astronomical Society*, p. 184.

A. Castro-Ginard et al. (2020). "(Big) Data mining Gaia DR2 to study the Galactic open cluster population." In: *Contributions to the XIV.o Scientific Meeting (virtual) of the Spanish Astronomical Society*, p. 127.

CONTENTS

1 INTRODUCTION

The study of the positions and motions of stars, known as astrometry, has been fundamental in the history of Astronomy. The precise astrometric characterisation of stars offers not just their accurate positions, but many insights into their properties as well as the properties of the large stellar complexes they form, *i.e.* structure and evolution of our own Galaxy. For this reason, the systematic compilation of stellar positions and, when possible, velocities has been a recurrent task throughout History.

Star catalogues are linked to the study of astrometry, both as its main result and the groundwork for later studies. Already in 127 B.C., Hipparcus of Nicaea compiled the first star catalogue which counted with $\sim$ 1000 stars located with a precision of 1°, which let him to derive the first value of the precession of the equinoxes. The Hipparcus' catalogue, re-compiled by Ptolemy and published as part of the Almagest in the 2_{nd}-century, remained the standard star catalogue until the publication of the catalogue by Tycho Brahe in 1598, which counted with precise measurements (1 arcmin) for 1000 stars, still without the aid of a telescope.

Among the many scientific revolutions started by Galileo Galilei, the design of the Galilean telescope in 1609 led to a rapid development of the field of observational Astronomy. However, it was not until 1838 that Friedrich Bessel published the first parallax measurement (0.3 arcsec to 61 Cygni) made using a heliometer, and use it to calculate the distance to a star (parallax was first measured for Vega by Friedrich Struve, but not published). Stellar parallaxes proved to be a good observational resource for the derivation of stellar distances, however, due to the difficulty in measuring them, only parallax measurements for $\sim$ 60 stars were available by the end of the 19_{th} century. Technological improvements during the 20_{th} century allowed for the first space-based astrometric mission, *Hipparcos* (Perryman et al. 1997) launched by the ESA in 1989. *Hipparcos* measured more precise parallaces than any other previous optical telescope, reaching a total of 118 200 parallax measurements with a precision of 1 mas. The *Hipparcos* mission also provided a second catalogue, Tycho-2, with parallaxes measured for $\sim$ 2.5 million stars, with a lower precision. The success of the *Hipparcos* mission in the 20_{th} century, continued with its successor *Gaia* providing a vast amount of parallax measurements (more than 1.3 billion), that will reach a precision of 15 μas at the end of the mission, for the 21_{st} century astronomers. Figure 1.1 shows how the

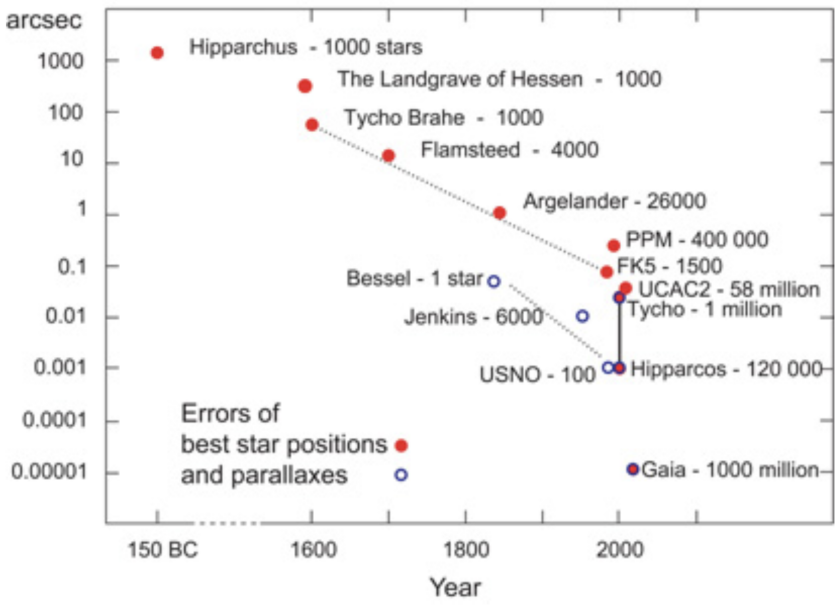


Figure 1.1: Evolution of precision of position (red dots) and parallax (blue circles) measurements overs the years. ESA.

precision in position and parallax measurements has improved over the years.

The success of the *Hipparcos* and *Gaia* missions is not only due to the parallax measurements for a large amount of stars, but also because of the inclusion of proper motion measurements. Proper motions are the projections over the two sky coordinates of the apparent motion of stars (the third component is given by the radial velocity). The movement of stars was already suspected by the first astronomers, but it was not until 1718 that Edmund Halley provided proof of stars' proper motion by noticing that the positions of Sirius, Arcturus and Aldebaran changed since the records of Hipparcus. Eversince, the evolution of proper motion and parallax measurements have gone hand in hand, concluding in the *Hipparcos*, Tycho-2 and *Gaia* catalogues.

THE GAIA MISSION

Gaia is a space-based telescope launched by the European Space Agency in 2013 (ESA, Gaia Collaboration et al. 2016b). The main goal of *Gaia* is to construct the largest three-dimensional map of the Milky Way by measuring very precise positions, parallaxes and proper motions of an unprecedented amount of stars, more than 1.5 billion up to magnitude $G = 20$. As said above, the success of *Gaia* builds on its predecessor mission, *Hipparcos* (Perryman et al. 1997), that measured 118 200 stars with parallaxes and proper motions, already revolutionising the field of Galactic archaeology.

The *Gaia* data releases are scheduled at different stages, each building on the previous one. The first data release (*Gaia* DR1, Gaia Collaboration et al. 2016a) counted with positions and mean photometry in the G band for 1.1 billion sources. The exclusion of parallaxes and proper motions based on *Gaia* data only, was due to the limited observing time of 14 months. However, for stars in common with *Hipparcos* mission, full five-dimensional astrometry was provided using a combination of the *Hipparcos*, Tycho-2 and *Gaia* catalogues. This subset of *Gaia* DR1, known as the Tycho-*Gaia* Astrometric Solution (TGAS, Michalik et al. 2015; Lindegren et al. 2016), with ~ 2 million sources, already enabled scientific applications (to cite some, van der Marel et al. 2016; Helmi et al. 2017; Bovy 2017). The second data release (*Gaia* DR2, Gaia Collaboration et al. 2018), with 22 months of observations, included positions, parallaxes and proper motions and mean magnitudes in three photometric bands (G, G_{BP} and G_{RP}) for 1.3 billion sources, together with radial velocity measurements for 7 million stars. The release scenario has recently changed with the publication of the early third data release (*Gaia* EDR3, Gaia Collaboration et al. 2020a), which contains more precise and more accurate measurements than those of *Gaia* DR2, based on a longer observing time. The future *Gaia* DR3 (scheduled for the first half of 2022) and *Gaia* DR4 (including the observations of the 5 year nominal mission) will include, besides even more precise astrometric and photometric measurements, the addition of mean G, G_{BP} and G_{RP} spectra together with radial velocity measurements for several million stars, and the inclusion of epoch data for all the stars in the release. Beyond the main astrometric and photometric data products, *Gaia* will also provide object classification and astrophysical parameters, stellar atmospheric parameter estimates, non-single stars, Solar system and extragalactic objects.

Gaia is mapping the stellar constituents of the Milky Way, providing a better picture of the Galaxy as a whole and its components (see Fig. 1.2 for an artistic impression). Given the large amount of proper motions, parallaxes and radial velocities of *Gaia* DR2, there have been many significant advances in all fields of Astronomy and in Galactic Astronomy in particular. For instance, the study of different regions of the Galactic disc and stellar halo revealed a structure with an extragalactic origin that has been accreted by the Milky Way, the *Gaia*-Enceladus-Sausage (Helmi et al. 2018). Accretion events leave an imprint in the disc of the Milky Way, and Antoja et al. (2018) found this signature from one of the past collisions with the Sagittarius dwarf galaxy. Further features of the Galactic disc have been also characterised with *Gaia* DR2, such as the Galactic warp (Romero-Gómez et al. 2019) or the bar (Anders et al. 2019). And as a last example, more locally in the Solar neighbourhood, the open cluster population has been re-defined in light of *Gaia* data (Cantat-Gaudin et al. 2018; Castro-Ginard et al. 2018).

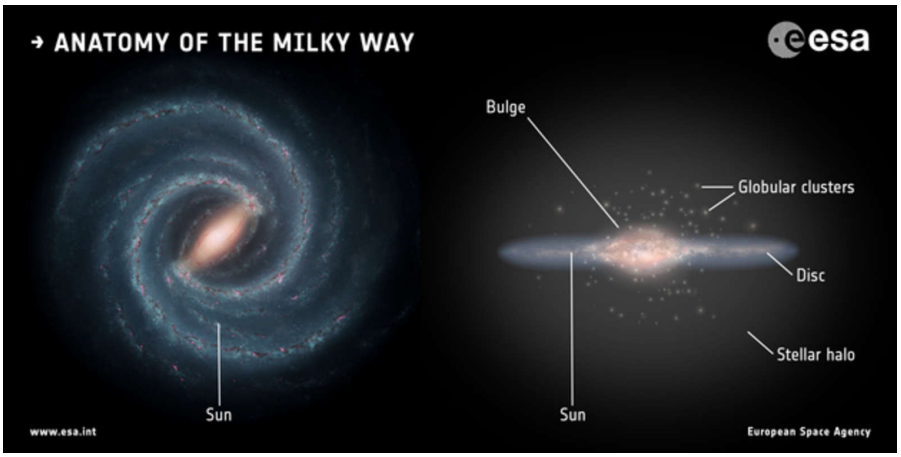


Figure 1.2: Artistic impression of the Milky Way and its main components. ESA.

MACHINE LEARNING

The publication of the second release of *Gaia* data (*Gaia* DR2, Gaia Collaboration et al. 2018), has provided the community with an enormous wealth of data which has revolutionised the field of Galactic archaeology. The more than 1.3 billion stars in the catalogue have also dramatically changed our way to analyse such a large amount of data, boosting the development of machine learning techniques and reinforcing the Big Data era in Astronomy.

Most of the *Gaia* individual measurements will, most likely, never be inspected by a human eye. Thus, there is a need for automatic methodologies to perform tasks such as the extraction of similarities or patterns in the data, the classification of objects in different classes based on similar features, or the estimation of new quantities from the observations. The goal of these automated procedures is not only to speed up the analysis, but to enable tasks that otherwise would be unfeasible. We can find examples of the application of these machine learning tasks already in the *Gaia* catalogue itself, with the computation of stellar parameters such as the effective temperature $T_{\rm eff}$, or line-of-sight extinction A_G and color excess $E(G_{BP} - G_{RP})$ (Andrae et al. 2018). There are also several studies using machine learning-based tasks for various topics based on *Gaia* data, for instance Cantat-Gaudin et al. (2019a) characterised the young cluster population in the Vela-Puppis region using an unsupervised learning methodology. Using supervised learning methodologies, deep learning in particular, Ostdiek et al. (2020) identified stars in the *Gaia* catalogue that could be accreted and not formed in our Galaxy. Necib et al. (2020) used these accreted stars to identify a prograde stellar stream from a massive dwarf galaxy that merged with the Milky Way. By comparing *Gaia* DR2 to simulated data, Mor et al. (2019) explored the star-formation history of the Milky Way, and found a star-formation burst 2-3 Gyr

ago in the Galactic disc. And among the most recent works, Gaia Collaboration et al. (2020b) identified 32 948 white dwarfs in a 100 pc bubble around the Sun using a Random Forest algorithm.

The development of these methodologies requires dedicated infrastructure to deploy the algorithms, due to the high demands on computational power or memory of most of these applications combined with the high data volume. There are several options to enable the Big Data analysis from the infrastructure point of view. There is the case of classic high-performance computing (HPC) centres with libraries dedicated to the analysis of large volumes of data. For instance, the MareNostrum supercomputer (Fig. 1.3), at the Barcelona Supercomputing Centre, consists in a HPC facility that counts with 3 456 nodes with 48 cores each and a total memory of 390 TB. In this thesis, this HPC facility has been combined with a Python-based machine learning library for distributed environments (Álvarez Cid-Fuentes et al. 2019), to detect more than 650 open clusters in the Galactic disc (Castro-Ginard et al. 2020). Another example using a different infrastructure is Mor et al. (2018), who used Apache Spark (Zaharia et al. 2012) and Apache Hadoop, a framework based on the MapReduce paradigm which is very popular in Indrustry applications, to perform fast approximate simulations of the Besançon Galaxy Model (Robin et al. 2003). The power of Apache Spark for astronomical applications was also shown by Zečević et al. (2019), who were able to crossmatch the *Gaia* DR2 and AllWise catalogues in $\sim$ 30 seconds. Garabato Míguez (2020) also used Apache Spark to adapt a type of self-organised maps, a neural network-based algorithm to visualise astrophysical objects with similar characteristics, to identify the astronomical class of objects based on *Gaia* BP/RP spectrophotometry. A last example using yet another different hardware, is Leung et al. (2019) who used a GPU-accelerated computation to simultaneously determine distances and the *Gaia* DR2 zero-point using a deep neural network. All the previous examples are also feasible using cloud computing services (Mor et al. 2020), which remove the need to own the aforementioned infrastructures, opening the access to Big Data analysis to the whole community.

The application of machine learning in a Big Data framework is not limited to *Gaia* DR2. The just published *Gaia* EDR3 (Gaia Collaboration et al. 2020a), includes more precise astrometry and photometry for most of the *Gaia* DR2 sources, and the addition of a number of new sources. However, the full *Gaia* DR3 will include radial velocity measurements for a larger number of stars than in *Gaia* DR2, together with G_{BP}, G_{RP} and G_{RVS} spectra for some stars. It will also include the GAPS (*Gaia* Andromeda Photometric Survey), with photometric time series for all the sources in a 5.5° radius field centred at Andromeda, that will increase the data volume of the catalogue. In the *Gaia* DR4, containing data for the 5 year nominal mission, epoch data

Figure 1.3: MareNostrum Supercomputer, located in an old chapel in the Barcelona Supercomputing Center. BSC.

for all sources will be delivered, reaching a volume of $\sim$ 1 PB for the archive. The trend in increasing the size of the catalogues, which is also true for surveys other than *Gaia*, is making Astronomy become a data-driven field. For instance, the 2MASS (Skrutskie et al. 2006) photometric survey in the near-infrared wavelength regime has reached 10 TB in data volume, or the SDSS (Blanton et al. 2017) which is up to 40 TB. However, the data tsunami is yet to come, with the Vera C. Rubin Observatory (Ivezić et al. 2019) or SKA (Dewdney et al. 2009) expecting to deliver 200 PB and 4.6 EB, respectively.

A particularly well suited field for the application of machine learning methodologies in Galactic surveys, regardless of their nature (astrometric, photometric or spectroscopic), is the field of open clusters (OCs). Already with *Gaia*, this field has gone through a major revolution, re-defining our understanding of these objects.

OPEN CLUSTERS IN THE MILKY WAY

Open clusters are gravitationally bound stellar systems which are composed of tens to thousands of stars mostly located in the Galactic disc, that were born together from the same event (Lada et al. 2003). Stars in an OC share a common location in the space, move at similar velocities, and share the same chemical composition, inherited from their mother giant molecular cloud (see Fig. 1.4). For OCs, parameters such as mean sky position, three-dimensional velocity, age, chemical composition, distance or line-of-sight extinction can be better estimated than for single field stars, making OCs fundamental objects to study a variety of astrophysical problems in our Galaxy. OCs are on their own excellent

Figure 1.4: The Pleiades open cluster. Smithsonian Magazine, Tony Hallas/-Science Faction/Corbis.

laboratories to investigate star-formation and stellar evolution theories, as well as the interactions and dynamical processes among their stars and with the Galactic potential that dissolves them (Friel 2013). In a more general view, taking into account the full OC population, OC are effective tracers of the Galactic disc structure and its properties such as the metallicity gradient of the Milky Way disc (Friel 1995; Netopil et al. 2016; Casamiquela et al. 2017) and its evolution (Friel et al. 2002; Frinchaboy et al. 2013), or tracing the dynamics of the Milky Way spiral arms (Dias et al. 2005; Junqueira et al. 2015).

The field of OCs is a perfect scenario for the application of machine learning tecniques. The fact that they represent clumps in a n-dimensional parameter space provides an excellent testbed for the application of unsupervised clustering algorithms for OC detection and characterisation. In terms of *Gaia* observables, five-dimensional astrometric parameters are available for this purpose $(l, b, \varpi, \mu_{\alpha^*}, \mu_\delta)$, given that radial velocity is only measured for a small subset of the whole *Gaia* DR2. Stars in an OC also follow an isochrone shape in a color-magnitude diagram (CMD), thus offering an opportunity to apply supervised techniques for the recognition of these patterns. Therefore, machine learning algorithms are particularly suited to provide a good characterisation of the OC population in our Galaxy, which is key to enable the aforementioned astrophysical studies.

Before *Gaia*, OCs were mostly detected by searching for overdensities in the sky position, while the other astrometric or photometric-derived parameters, such as proper motions or ages, were compiled from different data sources. Our knowledge of the OC population was summarised in two catalogues which counted with $\sim 2\,500$ common objects (Dias et al. 2002; Kharchenko et al. 2013, hereafter DAML and MWSC, respectively), and often offered a discrepand characterisation

of some objects (objects considered as real OCs or asterisms depending on the catalogue). This heterogeneous compendium from different surveys, each with different uncertainty levels, made the characterisation of these OCs a challenging task and often offered discrepant values for their parameters (Netopil et al. 2015). Furthermore, most of the objects reported in DAML or MWSC are located at less than 2 kpc from the Sun, including the assumption that our knowledge of the OC population was complete up to 1.8 kpc and showing the incompleteness of the OC census at farther distances mainly due to the decreasing luminosity or weaker contrast (cluster overdensity *v.s.* field star population). The situation changed with the publication of *Gaia* DR2, with its more than 1.3 billion sources with five-dimensional phase space coordinates $(l, b, \varpi, \mu_{\alpha^*}, \mu_\delta)$ and mean photometry in three broad bands (G, G_{BP}, G_{RP}). The *Gaia* homogeneous catalogue provides an excellent opportunity to solve the main issues from DAML or MWSC, and re-define the OC census based on precise and accurate new data.

Right after *Gaia* DR2, Cantat-Gaudin et al. (2018) re-visited the OC population reported by DAML and MWSC. The authors were driven by the locations reported in the previous OC catalogues, and used an unsupervised machine learning method, `UPMASK` (Krone-Martins et al. 2014), to characterise the membership probability of the stars belonging to each of the OCs. The main result was the characterisation of only 1 169 OCs out of $\sim$ 3 000. Even though some clusters were already flagged as dubious or asterisms in DAML or MWSC, and other were too distant or to reddened to be seen by *Gaia*, numerous clusters were found to be not physical groupings. Furthermore, the authors serendipitously discovered 60 new OCs in the vicinity of known ones, also showing the need for dedicated studies searching for these objects, and challenging the assumption of completeness up to 1.8 kpc.

The systematic search for unknown OCs in *Gaia* started even before the publication of *Gaia* DR2, with Castro-Ginard et al. (2018, presented in Chapter 2) scanning the bright Solar vicinity contained in the TGAS subset of *Gaia* DR1. In Castro-Ginard et al. (2018) we presented an unsupervised machine learning method to search for stellar overdensities in the five-dimensional astrometric space of *Gaia* data $(l, b, \varpi, \mu_{\alpha^*}, \mu_\delta)$, which we combined with a supervised learning method to recognise the isochrone pattern imprinted by OCs in a CMD. This methodology has been succesfully applied to the Galactic disc, in *Gaia* DR2, to discover more than 650 OCs clusters Castro-Ginard et al. (2018, 2019, 2020), and it has been followed by similar studies applying machine learning techniques for the same purpose. To cite some, Cantat-Gaudin et al. (2019b) detected 41 new OCs in the direction of Perseus using a Gaussian mixture model to detect overdensities in three-dimensions (parallax and proper motions), Sim et al. (2019) were able to detect 207 OCs by visually inspecting proper

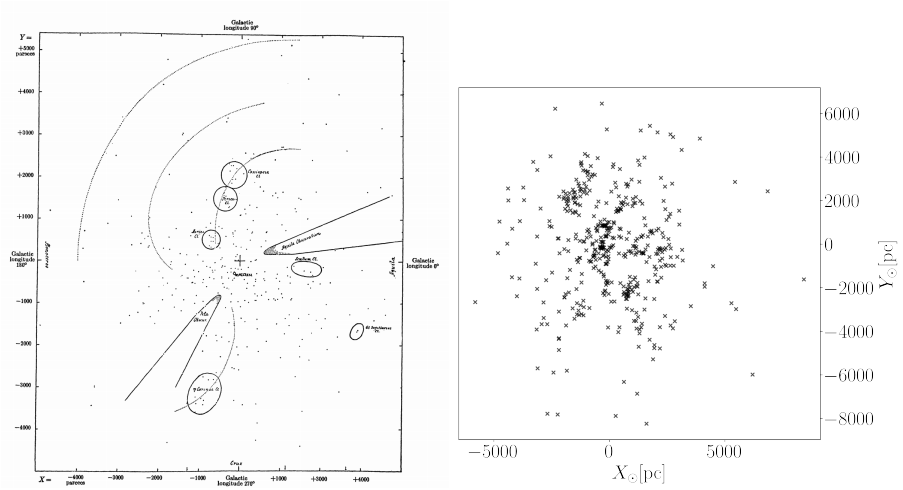


Figure 1.5: Known OC population at different epochs. *Left pannel*: Figure 8 from Trumpler (1930). Projection of OCs on the Galactic disc. The Sun is marked with a cross, dots represent the OCs and the dotted lines trace spiral structure centred at the Sun. *Right pannel*: Projection of the most up-to-date young OC population ($\leq$ 30 Myr, Castro-Ginard et al., submitted).

motion diagrams, Liu et al. (2019) reported the detection of 2 443 overdensities of which 76 were unknown OCs by using a friends-of-friends algorithm on three-dimensional regions of the sky, and Kounkel et al. (2019, 2020), who used `HDBSCAN` to detect groupings with filamentary shape within 3 kpc from the Sun. Recently, Hunt et al. (2020) compared the detection efficiency of the most used techniques, confirming that no single technique is able to detect all the existing structures.

The homogeneous characterisation of OCs and the detection of new ones using the *Gaia* DR2, allowed for the re-definition of the OC census in the solar neighbourhood. The most complete catalogue includes 2017 OCs with five-dimensional mean astrometric parameters and membership probabilities for their member stars based on *Gaia* data (Cantat-Gaudin et al. 2020b). They have also estimated astrophysical parameters for a large fraction of the total number of OCs, *i.e.* age, distance and line-of-sight extinction, computed from *Gaia* photometry in the G, G_{BP} and G_{RP} bands, in combination with parallaxes, using an artificial neural network trained on a set of reference clusters (Cantat-Gaudin et al. 2020b, Appendix a). For the sixth astrometric dimension, Tarricq et al. (2020) compiled radial velocity measurements from *Gaia* RVS and different spectroscopic ground-based surveys, *e.g. Gaia*-ESO (Randich et al. 2013), APOGEE (Ahumada et al. 2019), GALAH (Buder et al. 2018), OCCASO (Casamiquela et al. 2016), being able to report radial velocities for 1 385 OCs.

Open clusters as tracers of Galactic structure

OCs have been used to trace the Galactic disc structure for almost a century. Already in 1930, Trumpler (1930) attempted to describe the spiral structure of the Milky Way using OCs as tracers. The left pannel of Fig. 1.5 shows the distribution of the OC population known at the time, with spiral arms sketched on top. At that time, however, the incompleteness in the knowledge of the OC population led to a misleading view of the Galactic structure, with the spiral arms centred at the Sun. Improvements in the OC census allowed a better interpretation of it. This is shown in the right pannel of Fig. 1.5 where the young population of OCs ($\leq$ 30 Myr) is represented, highlighting the presence of Galactocentric spiral arms.

The nature of the spiral arms has been also debated since the 1960's. Lin et al. (1964) proposed that spiral arms are quasi-stationary density waves rotating around the Galactic centre with a constant pattern speed. This classical model is known as the density wave model, and its imprint has been detected in galaxies such as NGC 1566 (Shabani et al. 2018). Alternatively, Toomre (1964) proposed that spiral arms are short-lived, transient and recurrent structures, formed from a superposition of different density waves. For the case of the Milky Way, no concensus has been reached regarding which is the formation mechanism of its spiral arms. Studies using OCs (previous to *Gaia*), tended to explain spiral arms as classical density waves (Dias et al. 2005; Junqueira et al. 2015). However, in light of *Gaia* DR2, studies using the kinematic substructure in the Solar neighbourhood agreed with the transient nature of the arms (Hunt et al. 2018; Quillen et al. 2018).

The publication of *Gaia* DR2 has represented an increase in our knowledge of the OC population, and it has also allowed for the homogeneous compilation of its mean astrometric and astrophysical parameters. This new OC catalogue sets the groundwork for the study of the Milky Way spiral arms, providing new insights in the aforementioned debates.

THESIS OVERVIEW

This thesis is organised in three parts, corresponding to the presentation and application of the methodology to blindly search for OCs, the exploitation of the generated catalogue in terms of structure and evolution of the Galactic disc, and the conclusions. There is also a fourth part which corresponds to the Appendices, which include additional papers produced during the thesis.

In Part i, three Chapters describe the methodology used to detect new OCs in *Gaia* data. Chapter 2 is the description of the methodlogy, and a first application on the TGAS data. In Chapter 3 the methodology is applied to data from a region around the Galactic anticentre in *Gaia* DR2, and adapted to work with different density regions. The Chapter 4 presents the application of the methodology to the whole Galactic disc in *Gaia* DR2.

In the Part ii, which consists of a single Chapter 5, we present the use of the OC discovered in Part i (together with the previously known OCs) in a Galactic context. We use these OCs as tracers to distinguish as far as possible among different formation mechanisms of the spiral arms of the Milky Way.

In Part iii we present a summary of the work in the previous Parts. Also the conclusions from the individual Chapters and the future perspectives for work are discussed.

There are two additional works which are included in the form of Appendices. In Appendix a, we present a work where the astrophysical parameters for OCs which are relevant for Part ii are computed. In Appendix b we present a work where a full machine learning pipeline is devised to detect and characterise new young stellar objects in the near-infrared/infrared data from the Spitzer Space Telescope.

A flow chart representing the flux of information among different parts of the thesis is shown in Fig. 1.6.

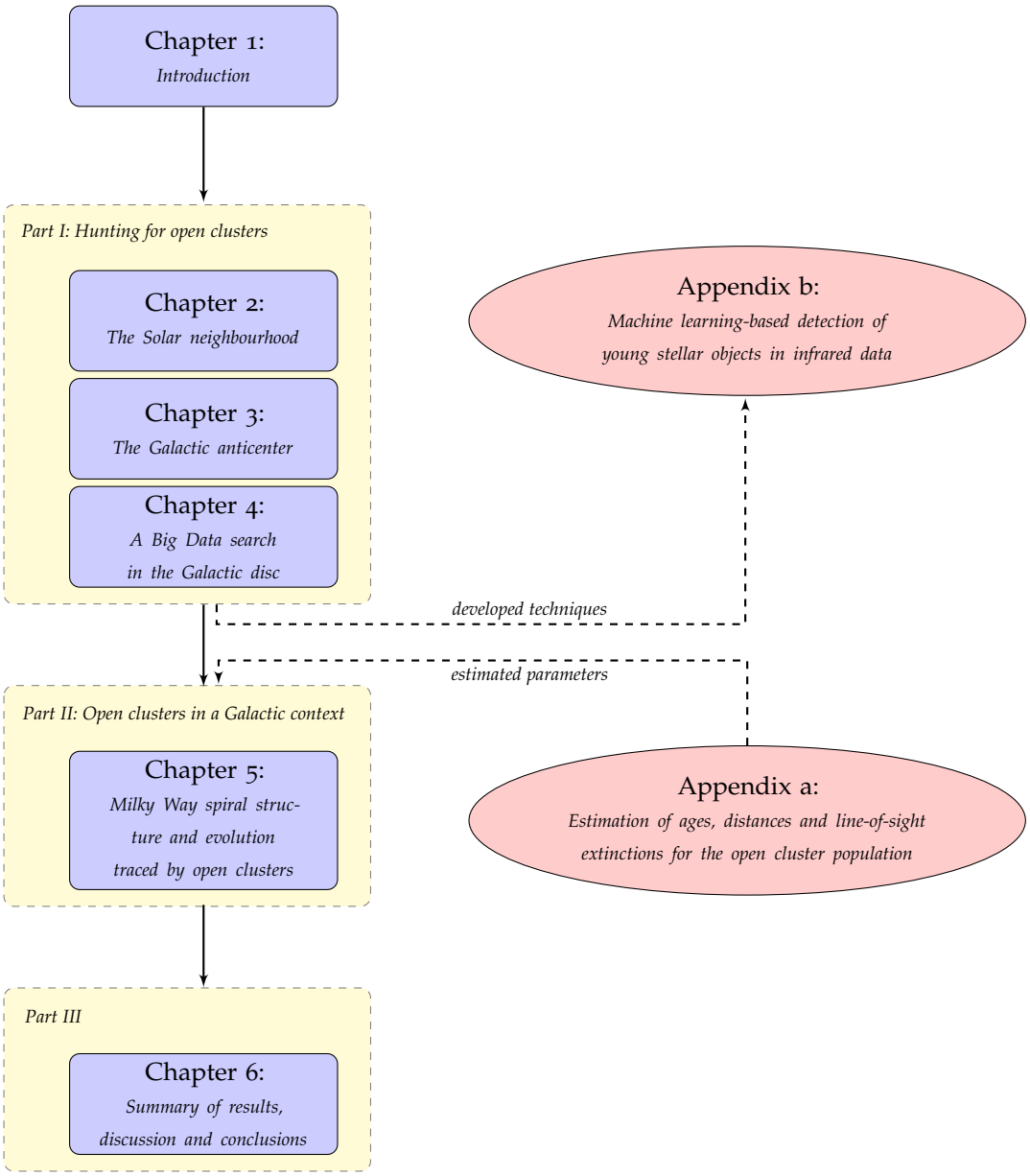


Figure 1.6: Overview of the thesis. Blue squares represent different chapters, which are organised in three main parts. Red circles are the appendices. Arrows show the information flux among the different parts of the thesis.

Part I

HUNTING FOR OPEN CLUSTERS

2

THE SOLAR NEIGHBOURHOOD

This Chapter contains the published version of Castro-Ginard et al. (2018, A&A, 618, A59).

The devise and development of a methodology based on different machine learning techniques is presented. The methodology implements a full machine learning pipeline, from the data preparation to the interpretation of the results, including both unsupervised and supervised learning. As a first step, the methodology is implemented to work with Tycho-*Gaia* Astrometric Solution data (TGAS, Michalik et al. 2015; Lindegren et al. 2016), the subset of the *Gaia* DR1 (Gaia Collaboration et al. 2016a) which contains parallax and proper motion measurements, together with sky positions, for around 2 million sources up to magnitude $G = 12$.

To search for statistical overdensities in the five dimensional astrometric parameter space $(l, b, \varpi, \mu_{\alpha^*}, \mu_\delta)$, we apply the clustering algorithm `DBSCAN` (Ester et al. 1996) . The overdensities found are classified into mere statistical overdensities[1] or real physical overdensities (assuming to correspond to OCs). The physical overdensities are recognised as such if the OC member stars follow an isochrone pattern in a CMD. For the CMDs, the 2MASS photometry (Skrutskie et al. 2006) was used due to the unavailability of *Gaia*'s G_{BP} and G_{RP} in its first data release. The recognition of the isochrone is automatically done using an artificial neural network (`ANN`, Hinton 1989) previously trained on data from well-characterised OCs (Gaia Collaboration et al. 2017).

The application of the methodology to the TGAS dataset provided 23 new OC candidates, which were later confirmed using the *Gaia* DR2 astrometric and phometric data (Gaia Collaboration et al. 2018). These new OCs, most of them closer than 1 kpc due to the bright limiting magnitude of TGAS, challenged the idea that the OC census was complete up to 1.8 kpc (Kharchenko et al. 2013) setting the path for numerous future OC searches in the Galactic disc.

1 Star groupings that are statistically more compact than random agrupations, but do not represent physical stellar groups.

A&A 618, A59 (2018)
https://doi.org/10.1051/0004-6361/201833390
© ESO 2018

Astronomy & Astrophysics

A new method for unveiling open clusters in *Gaia*

New nearby open clusters confirmed by DR2

A. Castro-Ginard[1], C. Jordi[1], X. Luri[1], F. Julbe[1], M. Morvan[1,2], L. Balaguer-Núñez[1], and T. Cantat-Gaudin[1]

[1] Dept. Fisica Quantica i Astrofisica, Institut de Ciències del Cosmos (ICCUB), Universitat de Barcelona (IEEC-UB), Martí Franquès 1, 08028 Barcelona, Spain
e-mail: acastro@fqa.ub.edu
[2] Mines Saint-Etienne, Institut Henri Fayol, 42023 Saint-Etienne, France

Received 8 May 2018 / Accepted 11 June 2018

ABSTRACT

Context. The publication of the *Gaia* Data Release 2 (*Gaia* DR2) opens a new era in astronomy. It includes precise astrometric data (positions, proper motions, and parallaxes) for more than 1.3 billion sources, mostly stars. To analyse such a vast amount of new data, the use of data-mining techniques and machine-learning algorithms is mandatory.
Aims. A great example of the application of such techniques and algorithms is the search for open clusters (OCs), groups of stars that were born and move together, located in the disc. Our aim is to develop a method to automatically explore the data space, requiring minimal manual intervention.
Methods. We explore the performance of a density-based clustering algorithm, DBSCAN, to find clusters in the data together with a supervised learning method such as an artificial neural network (ANN) to automatically distinguish between real OCs and statistical clusters.
Results. The development and implementation of this method in a five-dimensional space (l, b, ϖ, μ_{α^*}, μ_δ) with the Tycho-Gaia Astrometric Solution (TGAS) data, and a posterior validation using *Gaia* DR2 data, lead to the proposal of a set of new nearby OCs.
Conclusions. We have developed a method to find OCs in astrometric data, designed to be applied to the full *Gaia* DR2 archive.

Key words. surveys – open clusters and associations: general – astrometry – methods: data analysis

1. Introduction

The volume of data in the astronomical catalogues is continuously increasing with time, and therefore its analysis is becoming a highly complex task. In this context, the *Gaia* mission, with the publication of its first data release (*Gaia* DR1, Gaia Collaboration 2016) containing positions for more than one billion sources, opened a new era in astronomy. In spite of this large number of stars, however, full five-parameter astrometric data, that is, positions, parallax, and proper motions ($\alpha, \delta, \varpi, \mu_{\alpha^*}, \mu_\delta$) are available only for a relatively small subset. This subset is the *Tycho-Gaia* Astrometric Solution (TGAS Lindegren et al. 2016; Michalik et al. 2015), which provides a good starting point to devise and test scientific analysis tools in preparation for the larger releases, and in particular for the recently published second *Gaia* data release (*Gaia* DR2, Gaia Collaboration 2018). In *Gaia* DR2, precise five-parameter astrometric data for more than 1.3 billion stars are available, together with three-band photometry. The analysis of such a vast amount of data is simply not possible with the usual techniques that require a manual supervision, and has to rely on the use of data-mining techniques and machine-learning algorithms. In this paper we develop a set of such techniques, allowing an automatic exploration of the data space for the detection of open clusters (OCs); we apply them to TGAS and we check the validity of the results with the DR2 data, in preparation for its application to the full dataset.

The analysis tools developed in this paper are designed for the automated detection of OCs. According to the currently accepted scenarios of star formation, most of the stars are born in groups from giant molecular clouds (see for instance Lada et al. 1993). Such groups, of up to a few thousand stars, can lose members or even completely dissolve due to internal and close external encounters with stars and gas clouds in their orbits in the Galactic disc. Open clusters, being the fundamental building blocks of galaxies, are key objects for several astrophysical aspects: (a) very young OCs are informative of the star formation mechanism (the fragmentation of the gas clouds, the time sequence of formation, the initial mass function (IMF)), (b) young OCs trace the star forming regions (young clusters are seen near their birth place), (c) the evaporation of OC stars into the field stellar population (by studying the internal kinematics and the mass segregations), (d) intermediate and old OCs allow for the study of chemical enrichment of the galactic disc due to more precise determination of ages than for field stars (gradients with galactocentric distance and age can be analysed), (e) the stellar structure and evolution (colour magnitude diagrams (CMDs) provide empirical isochrones to compare with the theoretical models). The most updated and complete compilations of known OCs are those in Dias et al. (2002) and Kharchenko et al. (2013)[1]. Both lists are internally homogeneous in their determination of mean proper motions, distances, reddening and ages, but there is no full agreement between them on which group of stars is considered a cluster or an asterism. In total, there are about 2500 known OCs, most of them detected as stellar overdensities in the sky and confirmed through proper motions and/or CMDs. About 50% of the OCs in these samples are closer than 2 kpc and about 90% are closer than 5 kpc.

[1] Supplemented by Schmeja et al. (2014) and Scholz et al. (2015).

Certainly, our knowledge of OCs beyond 1–2 kpc is rather incomplete due to the decreasing angular size and luminosity of the clusters with distance and the obscuration by the interstellar dust. Froebrich (2017) identified 125 compact (distant) and so-far unknown OCs using deep high-resolution near-infrared (NIR) surveys, again by identifying overdensities in the spatial distribution confirmed as OCs using CMDs.

The recently released *Gaia* DR2 provides an ideal dataset for the detection of so-far unknown OCs. Identifying clustering of objects in a multidimensional space (positions, proper motions, parallaxes and photometry) allows for a much more efficient detection of these objects than simply using the usual two-dimensional (2D) (sky positions) approach. With this purpose in mind we have devised a method to systematically search for OCs in *Gaia* data in an automatic way and we have, as an initial validation step, applied it to the TGAS subset of *Gaia* DR1 (Gaia Collaboration 2016). Although the 2 million stars in TGAS have a relatively bright limiting magnitude of ∼12, the inclusion of the proper motions and parallaxes allows us to detect sparse or poorly populated clusters that have so far gone undetected in the solar neighbourhood[2]. Importantly, the inclusion of additional dimensions and the better precision of the data increases the statistical significance of the overdensities. These overdensities are detected using a density-based clustering algorithm named DBSCAN (Ester et al. 1996), which has been previously used to find spatial overdensities (Caballero & Dinis 2008) or cluster membership determination (Wilkinson et al. 2018; Gao et al. 2014, 2017); they are subjected to a confirmation step using a classification algorithm based on an artificial neural network (Hinton 1989) to recognise isochrone patterns on CMDs. The thus-detected candidate OCs are finally validated by hand using *Gaia* DR2 (Gaia Collaboration 2018) photometric data, in order to confirm the validity of the methodology in view of its application to the full *Gaia* DR2 archive in an upcoming paper.

This paper is organised as follows: in Sect. 2, we describe the clustering algorithm used. In Sect. 3, we optimise the choice of the values of the algorithm parameters by applying it to a simulated dataset. In Sect. 2.3, the neural network classification algorithm used to discriminate between real OCs and detections due to random noise is described. In Sect. 4, we discuss the results of the method when applied to the TGAS dataset, materialised in a list of 31 OC candidates. Finally, these candidates are manually validated using *Gaia* DR2 photometric datasetta in Sect. 5, allowing to us confirm most of them. Conclusions are presented in Sect. 6.

2. Methods

The methodology used to identify groups of stars as possible new OCs is sketched in Fig. 1. Starting from the whole TGAS catalogue and after applying a preprocessing step (see Sect. 2.1), an unsupervised clustering algorithm named DBSCAN[3] detects statistical clusters (see Sect. 2.2) in the data. After removing the OCs already catalogued in MWSC, an Artificial Neural Network[3] is applied to automate the distinction between statistical clusters and physical OCs, based on a CMD built using the photometric data from the 2MASS catalogue.

[2] For instance Röser et al. (2016) discovered nine OCs within 500 pc from the Sun based on proper-motion analysis using a combination of *Tycho*-2 and URAT1 catalogues. The existence of still-undiscovered nearby OCs cannot therefore be discarded.

[3] Algorithm from the scikit-learn python package (Pedregosa et al. 2011).

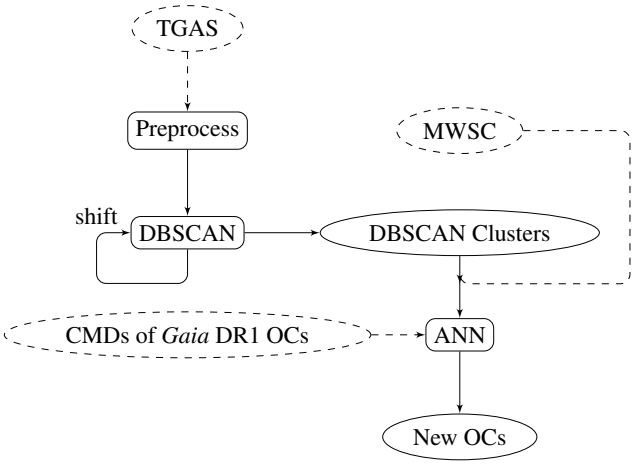


Fig. 1. Flow chart of the method applied to find OCs. Solid boxes represent code, solid ellipses represent generated catalogues, and dashed ellipses represent external catalogues.

2.1. Preprocessing

Most of the catalogued OCs are found in the Galactic disc ($|b| < 20$ deg), for example, 96% of the clusters from the Dias catalogue (Dias et al. 2002) and 94% from the MWSC (Kharchenko et al. 2013) lie in that region. We therefore explore the Milky Way disc scanning all longitudes in the region ±20 deg in latitude. In addition, we remove stars with extreme proper motions and large or negative parallaxes. This helps in the determination of the DBSCAN parameter ϵ (see Sect. 2.2) with almost no loss of generality because these conditions would make any OC easily detectable. A star with the following values is rejected by the algorithm: $|\mu_{\alpha^*}|, |\mu_\delta| > 30$ mas yr^{-1}, $\varpi < 0$ mas and $\varpi > 7$ mas.

The resulting sky area of study is further divided into smaller regions, rectangles of size L deg, where the clustering algorithm is to be applied. The reason for this division is twofold. On the one hand, it saves computational time because the volume of the data in the region is much smaller. On the other hand, the DBSCAN algorithm needs a starting point to define an averaged density of stars in the region; with smaller regions this average is more representative than if we take the whole sky, where the density can significantly vary from one region to another. Once we have the sky divided into rectangles, to avoid the redundant detection of split clusters that might be spread over more than one of these regions or may be in the intersection of two regions, any cluster found with at least one star on the edge of the rectangle is rejected. To deal with the border conflicts the rectangles are shifted $L/3$ and $2L/3$ and the algorithm is run one more time for each shift. During these shifts, the algorithm explores regions where $|b| > 20$ deg, so clusters in that region might appear. The clusters found in the second or third run are then only taken into account if none of its members is in any cluster of the previous runs; in this way we ensure that no clusters are missed or detected more than once because they are on the borders of the regions.

The last step in the preprocessing is the scaling of the star parameters used by DBSCAN. The algorithm makes use of the distance between sources in the N-dimensional space to define if the stars are clustered or not. Because there is no dimension preferred in the five-dimensional (5D) parameter space (l, b, ϖ, μ_{α^*}, μ_δ), we standardise the parameters (rescale them to

mean zero and variance one) so that their weights in the process are equalised.

2.2. DBSCAN

Once the region of the search is defined and the average distance between stars in the parameter space is determined, an automatic search for groups of stars that form an overdensity in the 5D space is started.

The clustering algorithm DBSCAN (Ester et al. 1996) is a density-based algorithm that makes use of the notion of distance between two sources in the data to define a set of nearby points as a cluster; it has the advantage over other methods of being able to find arbitrarily shaped clusters. An OC naturally falls in the following description: groups of stars with a common origin, meaning that they share a common location (l, b, ϖ) and motion $(\mu_{\alpha^*}, \mu_\delta)$. The TGAS (Lindegren et al. 2016) data set contains precise information for these five parameters, so one can define the distance between two stars (i and j) as

$$d(i,j) = \sqrt{(l_i - l_j)^2 + (b_i - b_j)^2 + (\varpi_i - \varpi_j)^2 + (\mu_{\alpha^*,i} - \mu_{\alpha^*,j})^2 + (\mu_{\delta,i} - \mu_{\delta,j})^2}. \quad (1)$$

The choice of this euclidean distance is due to its simplicity, although a distance with specific weights on the different parameters, in order to optimise the search for different kinds of clusters (rich or poor, sparse or compact, etc.) or to take into account the uncertainities of each value, could be investigated. We also note that the distance is calculated with the standardised values of these parameters.

The definition of a DBSCAN cluster depends on two paramters: ϵ and *minPts*. A hypersphere of radius ϵ is built centred on each source, and if the number of sources that fall inside the hypersphere is greater than or equal to the pre-set *minPts,* the points are considered to be clustered. This definition of cluster allows us to make the distinction between three types of sources in the data set: i) core points, sources that have a number of neighbours (within the hypersphere of radius ϵ) greater than or equal to *minPts*, ii) members, sources that do not have these neighbours in their hyperspheres but fall in the hypersphere of a core point, and iii) field stars, sources than do not fulfil any of the two previous conditions. For an intuitive 2D description of a cluster in DBSCAN, see Fig. 2.

Determination of the ϵ and *minPts* parameters

Therefore the DBSCAN algorithm depends only on two parameters, the minimum number of sources (*minPts*) to consider that a cluster exists and the radius (ϵ) of the hypersphere in which to search for these *minPts* sources. In order to determine the optimum value of *minPts* for OC detection, the algorithm is tested with a simulated sample and a set of the values that perform best is chosen (see Sect. 3). In particular, the determination of ϵ is crucial for the efficiency of the detection, and the selected values can affect the number and shape of the clusters found.

Aiming to reduce the free input parameters, we have implemented an automated determination of the ϵ value that best fits the data on a given region. Since a cluster is a concentration of stars in the parameter space, the distance of each star belonging to a cluster to its $k_{\rm th}$ nearest neighbour should be smaller than the average distance between stars belonging to the field (Fig. 3). Our determination of ϵ, taking advantage of this fact, is as follows:

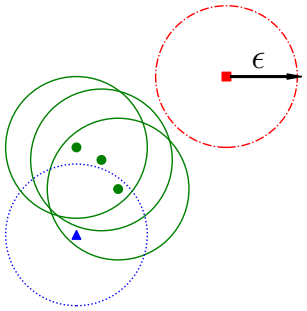


Fig. 2. Schematic representation of a DBSCAN cluster with *minPts* = 3. Points in green represent core points, each point has *minPts* points in its (green solid) hypersphere. The blue triangle represents a member point, it does not have *minPts* in its (blue dashed) hypersphere but it is reached by a core point. The red square represents a field star; it does not have any other point in its (red dash-dot) hypersphere. All the hyperspheres have radius equal to ϵ.

- Compute the $k_{\rm th}$ nearest-neighbour distance (kNND) histogram for each region and store its minimum as $\epsilon_{k\rm NN}$.
- Generate a new random sample, of the same number of stars, according to the distribution of each astrometric parameter estimated using a Gaussian kernel density estimator. Subsequently, compute the kNND histogram for these stars and store the minimum value as $\epsilon_{\rm rand}$. Since we are generating random samples, the minimum number of the kNND distribution will vary upon each realisation; in order to minimise this effect we store the average over 30 repetitions of this step: $\epsilon_{\rm rand}$.
- Finally, to obtain the most concentrated stars (which will be considered as the candidate members of the OC) and minimise the contamination from field stars, the choice of the parameter is $\epsilon = (\epsilon_{k\rm NN} + \epsilon_{\rm rand})/2$.

Figure 3 shows a real distribution of seventh-nearest neighbour distance ($7_{\rm th}$-NND) around the cluster NGC 6633 (in blue) together with a random resampled $7_{\rm th}$-NND histogram (in orange) with the choice of ϵ in that region (red line); the peak belonging to the cluster is well separated from field stars through ϵ. In addition, the figure shows the histogram of distances to the seventh-nearest neighbour of each star in the NGC 6633 cluster (in green), where the members are taken from Gaia Collaboration (2017).

The choice of the value for k has to be related to the expected members of the cluster. Here, since *minPts* determines the minimum members of a cluster, the value for k is set to $k = minPts - 1$. Two free parameters (L, *minPts*) are left to be optimised using simulations (see Sect. 3).

2.3. Identification of open clusters

At this point, when DBSCAN has found a list of candidate OCs, the method needs to be refined to distinguish real OCs from the statistical clusters (random accumulation of points). This step is an automatisation of what is usually done by visual inspection; plot the CMD of the sky region and see if the clusterized stars follow an isochrone. We treat this as a pattern-recognition problem, where artificial neural networks (ANNs) with a multilayer perceptron architecture have been shown to be a good approach (Bishop 1995; Duda et al. 2000). Similar problems, such as the identification of globular clusters (Brescia et al. 2012) or a selection for quasi stellar objects (QSOs; Yèche et al. 2010), have also been solved using a multilayer perceptron.

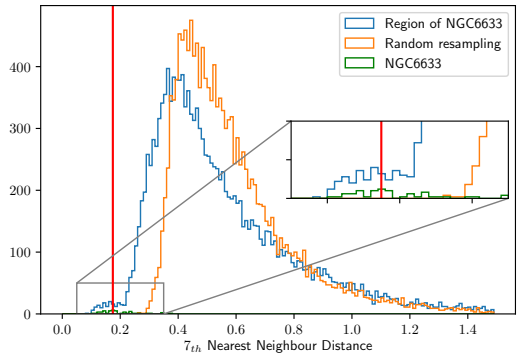


Fig. 3. Histogram of the $7_{\rm th}$-NNDs of the region around the cluster NGC 6633. The blue line shows the $7_{\rm th}$-NND histogram of all the stars in that sky region in TGAS. Orange line shows the $7_{\rm th}$-NND histogram of one realization of a random resample. Green line shows the $7_{\rm th}$-NND histogram for the listed members of NGC 6633 (more visible in the zoom plot). The red line corresponds to the chosen value of ϵ in this region. The plot was made with the parameters $L = 14$ deg and $minPts = 8$.

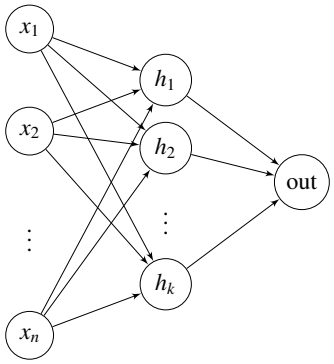


Fig. 4. Schematic representation of a multilayer perceptron with one hidden layer. The x_i values represent the input data. The h_i labels represent neurons in the hidden layer.

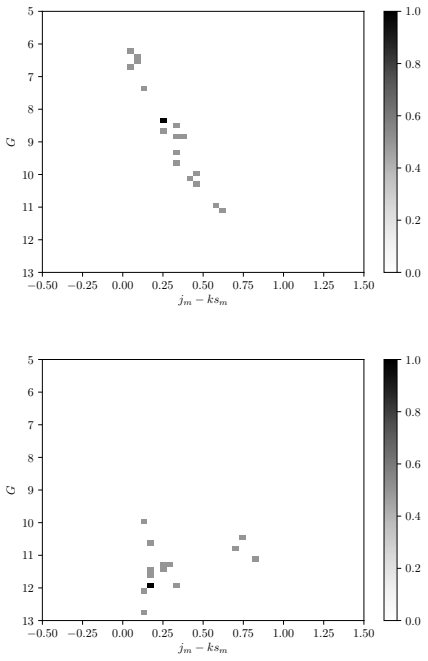


Fig. 5. Examples of training data for the ANN classificator. The upper plot corresponds to a density map of a CMD of a subset of the members of Coma Berenices. The lower plot is the density map of a CMD of a cluster found by DBSCAN that we labelled as noise. In both cases, the colours represent the value of each pixel, and this is the input of the ANN model.

2.3.1. Artificial neural networks

Artificial neural networks are computing models that try to mimic how a biological brain works. In particular, the multilayer perceptron consists in a set of at least three layers of nodes (neurons) capable of classifying a given input feature vector into the class it belongs.

Figure 4 shows a schematic representation of a multilayer perceptron with one hidden layer. The left-most (input) layer represents the set of input features $\{x_1, x_2, \ldots, x_n\}$. This is followed by the hidden layer, where each hidden neuron (labeled as h_i) weights the received input from the previous layer as $\nu_i = \omega_{i1}x_1 + \omega_{i2}x_2 + \cdots + \omega_{in}x_n$, and responds according to an activation function, in our case we use a hyperbolic tangent activation function

$$y(\nu_i) = \tanh(\nu_i), \tag{2}$$

which is then passed to the output layer that performs the classification.

2.3.2. Data preparation

Artificial neural networks are supervised classification algorithms that require a pre-classified learning sample to train them. In our case, the data used to train the model are the OCs taken from Gaia Collaboration (2017). These clusters are well-characterised; they have a reasonable number of members and show clear isochrones in the CMD, and are the target of our pattern-recognition algorithm. Furthermore, they have the same astrometric uncertainties as our data so they are representative of our problem. In order to train the model, and to increase the size of the training set, several subsets of these OC member stars are randomly selected and plotted in a CMD to serve as patterns. Moreover, CMDs that do not correspond to clusters are also needed as examples of negatives for the training. In this case, we inspect the output from DBSCAN (for pairs of $(L, minPts)$ that were not used in the detection step) and select sets of clusterized stars not following any isochrone.

Figure 5 shows two examples of training data sets for the model. The upper plot corresponds to members of the Coma Berenices cluster listed in Gaia Collaboration (2017). The members are randomly chosen to form a set of ten sub-clusters, each one with characteristics similar to those found by DBSCAN. The CMD of these sub-clusters is then converted to a density map so that the value of each pixel can be used as the input for the ANN. A density map of one of these sub-clusters is shown in the upper plot. The lower plot corresponds to non-clusters for the training on negative identifications.

2.3.3. Performance of the classification

The ANN classificator is trained with a total of 296 images, containing a balanced relation between CMDs from true (real) OCs

and CMDs from field stars. For performance estimation purposes, this whole set is divided into a training and a test set, containing 67% and 33% , respectively. The test CMDs are classified with a precision of a 97.95% to the right class (OC or field stars). Even though the model is then trained with all the 296 CMDs, the precision reached in the test set is only an estimation of the upper limit because the ANN has learnt from the OCs in *Gaia* DR1 listed in Gaia Collaboration (2017). The detection of new OCs is then limited to have the same characteristics as those in Gaia Collaboration (2017), where there are a total of 19 nearby OCs with ages ranging from 40 to 850 Myr, and no significant differential extinction. A training set that is larger and wider in terms of characteristics of the OCs needs to be built in order to apply the method to the *Gaia* DR2 data.

3. Simulations

A simulation of TGAS-like data is used to test the clustering method and set the optimal parameters to detect as many clusters as possible with a minimum of false positives.

As described in Arenou et al. (2017), the simulation consists in astrometric data from *Tycho*-2 stars taken as nominal where errors coming from the AGIS solution have been added. The proper motions used for the simulation are those from *Tycho*-2; to prevent their dispersion from spuriously increasing when adding the TGAS errors, they were "deconvolved" using Eq. 10 from Arenou & Luri (1999). In the case of the parallaxes, for nearby stars, the simulated value is a weighted average of "deconvolved" Hipparcos parallaxes, while for the more distant stars, it is taken from the photometric parallax in the Pickles & Depagne (2011) catalogue. The simulation of the TGAS-like errors follows the description from Michalik et al. (2015), which is based on the algorithms from Lindegren et al. (2012). In short, this dataset is very representative of the real TGAS dataset that we use both in terms of its distribution of parameters (taken from Tycho) and its astrometric errors (generated to be as close as possible to the TGAS ones).

The OCs are added to this dataset *a posteriori*, simulated using the *Gaia* Object Generator (GOG; Luri et al. 2014) (see details of how they are simulated in Roelens 2013). For each cluster, the stars with $G > 12$ are filtered out due to the limiting magnitude in TGAS. Moreover, the simulation provides true values for the astrometric parameters to which observational errors are added. Using the uncertainties published in the TGAS catalogue, a normal random number is drawn centred in the true value, to compute the observed quantities.

Choice of the parameters

Selection of the best parameters to run the algorithm is made in terms of noise and efficiency. Their definition, in terms of true positive rate (tp), false positive rate (fp), and false negative rate (fn), is fp/tp for noise and fn/tp for the efficiency.

In order to find the pairs of parameters that best perform, the algorithm was run over several pairs of (L, *minPts*). The sweep over this parameter space allowed us to select the set of pairs of parameters that are less contaminated by spurious clusters. Figure 6 shows the performance of each pair of (L, *minPts*) for the investigated pairs. The reddest pixel represents the best performing pairs of parameters while the bluest pixels represent the worst performing pairs. In the best case, with noise around ~0.25, we are introducing one spurious cluster in the detection every four real clusters, while in the worst case, we have a noise around 0.5. An efficiency of 0.25 means that we do not detect one out of four real clusters. The selection was made in an attempt to find a balance between noise and efficiency; the black box in Fig. 6 represents the selected pairs of (L, *minPts*), which are $L \in [12, 16]$ and *minPts* $\in [5, 9]$.

4. Results

The whole method is run over the TGAS data to obtain a list of OC candidates. First, the DBSCAN algorithm is applied to the preselected data (see Sect. 2.1) with the optimal values for the parameters $L = \{12, 13, 14, 15, 16\}$ and *minPts* $= \{5, 6, 7, 8, 9\}$. This results in a list of clusterized stars, including real clusters already catalogued, non-catalogued possible clusters, and noise. Although the clusters that are already catalogued are useful to verify that the algorithm is capable of finding real clusters, they are discarded (see Fig. 1). To do this, all the clusters found by DBSCAN whose centre lies within a box of 2 deg × 2 deg centred in a cluster present in the MWSC catalogue are discarded. In this way, we ensure a list composed only of new cluster candidates. Röser et al. (2016) published a list of nine nearby OCs using proper motions from a combination of Tycho-2 with URAT1 catalogues. We did not include these clusters in the "cross-match with known clusters" step, in order to use them to check the method.

The classification of these clusters into probable OC candidates and statistical clusters is done with the ANN algorithm. The model is trained with CMDs from real clusters (see Sect. 2.3.2) with the photometric data from 2MASS and TGAS, and it is capable of identifying isochrone patterns in CMDs. The isochrone patterns identified by the ANN model are based on those of the OCs listed in Gaia Collaboration (2017). Only the clusters found to follow an isochrone with a confidence level higher than 90% are selected.

Table 1 lists 31 open cluster candidates resulting from the application of the above-described algorithms. We include the mean sky position, proper motions, and parallaxes of the identified members. We do not provide uncertainties because the data have been superseeded by *Gaia* DR2. Because the method is run over 25 different pairs of parameters (L, *minPts*), the final list is sorted by the number of appearances of the clusters in the different pairs of parameters. The value $N_{\rm found}$ indicates how many times the cluster has been found for the used pairs of (L, *minPts*).

As mentioned above, we did not include the OCs in Röser et al. (2016) in the list of previously known clusters and therefore we expect some overlap with our candidates. This is the case for our UBC1 and UBC12, which are RSG4 and RSG3, respectively.

Of the other seven clusters, RSG2 was not found, possibly due its high galactic latitude and its high μ_δ mean, which is -29.54 mas yr^{-1}. Because our preprocessing removes stars with $|\mu_{\alpha^*}|, |\mu_\delta| > 30$ mas yr^{-1} (see Sect. 2.1) and due to the proper motion uncertainties in the TGAS catalogue, we may lose part of the members and, so, the algorithm does not consider the surviving members as a cluster. On the other hand, the criterium to match our candidates with the list of known OCs is purely positional (within a box of 2 deg × 2 deg). We do not impose a match in proper motions and/or parallaxes because of the large differences between the values quoted in MWSC and Dias et al. (2002), which makes us doubt the reliability of some values. This criterion discards candidate clusters that are in the vicinity of know clusters, and this is the case of RSG1, RSG5, RSG6, RSG7, RSG8 and RSG9. Our candidate list is therefore not complete, especially at very low latitudes where the density of known clusters increases.

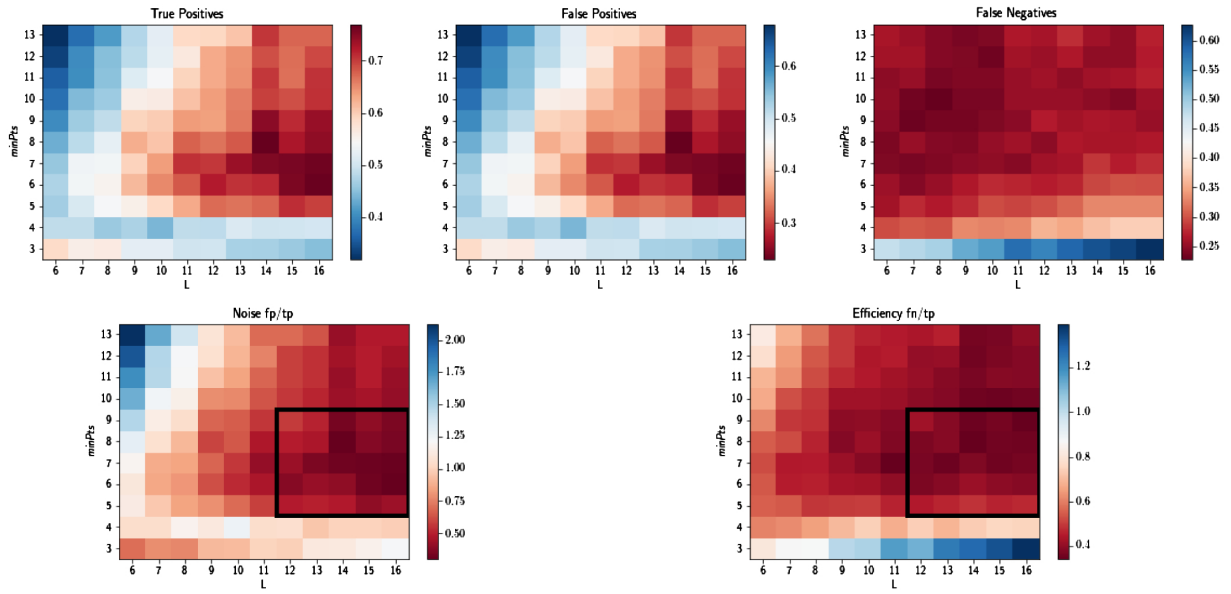


Fig. 6. Performance of the algorithm with a different set of parameters ($L, minPts$) tested with simulated data. *Top panels*: true positive (*left*), false positive *(middle*), and false negative (*right*) rates. We highlight the inversion of the colour bar in the true positive rate to always represent the reddest pixels as the best performing pair of parameters. *Bottom panels*: noise (*left*) and efficiency (*right*). The black box encloses the area of pixels corresponding to the selected pairs of parameters.

UBC7 shares proper motions and parallaxes with Collinder 135. It is located at 2.3 deg from the quoted Collinder 135 centre and for this reason it is not matched in our step to discard already known clusters. Figure 7 shows a cone search of 10 deg centred in UBC7 where a pattern in the data is clearly visible. This pattern is an artefact of the Gaia scanning law in the 14 month mission of Gaia DR1. UBC7 is located where two stripes cross and this, together with the fact that their stars also share parallaxes and proper motions, leads to its detection as a separate cluster. This is an indication that the inhomogeneities in the sky coverage of TGAS data might lead to the detection of spurious clusters. Collinder 135 is not detected by DBSCAN because their members lie in a region not well covered by the observations. Furthermore, they are more spread than UBC7 and they are not recognised as a group. It could be that Collinder 135 is larger than quoted in the literature and includes UBC7.

5. Validation using *Gaia* DR2

Gaia DR2 provides an excellent set of data for the confirmation of our candidate members because of the improved precision of the astrometric parameters, the availability of those parameters for the stars down to ~21 mag, and the availability of precise G, G_{BP} and G_{RP} photometry.

In order to validate each cluster, we run our method again with a set of DR2 objects selected in a region around its centre (a cone search of 1 or 2 deg depending on the mean parallax of the cluster). The determination of the ϵ parameter for DBSCAN is now more complicated due to the higher density of stars in the *Gaia* DR2 data, reaching, in some studied cases, ~150 000 stars in that region. Because our goal here is simply to validate the already found candidates (not detecting new OCs) and thus validate our method, we apply a set of cuts in the data. These cuts are mainly in magnitude and parallax to increase the contrast between the cluster and field populations, to avoid large uncertainties, and to discard distant stars (our candidates being detected with TGAS data, the clusters are necessarily nearby; see Fig. 8).

Figure 9 shows an example of UBC1 in the TGAS (top panels) and *Gaia* DR2 (bottom panels) data. Left plots show the spatial distribution of the member stars found in each data set; in the TGAS case, this shows a squared area of 10 deg × 10 deg whilst in *Gaia* DR2, it is a cone search of 2 deg. The middle plots show the members in the proper motion space and we can see that in *Gaia* DR2 data the stars are more compact. The major difference is in the rightmost plots where a CMD is shown for both cases, one using photometry from 2MASS (top) and one using only *Gaia* data (bottom). The much better quality of the *Gaia* photometric data (both plots share the same stars for $G \leq 12$) allows us to see the isochrone pattern that the member stars follow with greatly improved clarity.

We are able to re-detect, and thus confirm, a high percentage of the listed OCs using DBSCAN in a region around the cluster. Table 2 lists the confirmed OCs. The clusters that we consider as confirmed are those which share most of the stars with those previously found in TGAS. See plots similar to Fig. 9 in Appendix A for all the OCs. *Gaia* DR2 includes mean radial velocities for stars brighter than 12 mag. In Table 2 we include the mean radial velocity for the OCs derived from the identified members.

The non-confirmed clusters are UBC15, UBC16, UBC18, UBC22, UBC23, UBC24, UBC25, UBC28, UBC29 and UBC30. They are all in the second half of Table 1, which means that they are the least-frequently found ($N_{\rm found} < 5$) within the explored parameters ($L, minPts$). The criteria followed in order to sort the list of candidates is reasonable; 100% of the clusters with $N_{\rm found} \geq 5$ are confirmed, while for $N_{\rm found} < 5$, 59% are confirmed. As a whole, we are able to confirm ~70% of the proposed candidates; this is within the expected performance

Table 1. List of the 31 open cluster candidates.

Name	α (deg)	δ (deg)	l (deg)	b (deg)	ϖ (mas)	μ_{α^*} (mas yr^{-1})	μ_δ (mas yr^{-1})	$N_{\rm found}$
UBC1[a]	287.83	56.62	87.30	19.77	3.04	−2.80	3.69	27
UBC2	4.90	46.38	117.22	−16.13	1.62	−5.95	−5.67	24
UBC3	283.74	12.29	44.29	4.80	0.53	−1.57	−2.31	21
UBC4	60.73	35.23	161.37	−12.97	1.74	−0.08	−5.36	21
UBC5	238.65	−47.66	331.90	4.63	1.61	−7.21	−4.80	18
UBC6	343.87	51.14	105.06	−7.65	1.35	−7.46	−4.54	15
UBC7[b]	106.64	−37.54	248.52	−13.36	3.67	−9.43	7.03	14
UBC8	84.65	56.99	155.06	13.35	2.17	−3.35	−3.24	13
UBC9	276.60	26.42	54.48	16.84	2.80	−0.12	−5.31	12
UBC10	324.20	60.86	101.34	6.43	0.99	−1.73	−3.15	10
UBC11	246.61	−60.17	326.80	−7.69	2.15	−0.25	−7.34	10
UBC12[c]	126.11	−8.39	231.65	16.32	2.32	−8.19	4.47	6
UBC13	121.24	4.14	217.71	18.23	1.75	−7.22	−1.48	5
UBC14	295.01	3.21	41.43	−9.29	1.33	0.56	−1.76	5
UBC15	268.05	−25.89	3.35	0.30	0.77	1.06	−1.38	4
UBC16	143.77	−27.40	258.09	17.91	1.93	−4.67	2.15	3
UBC17	83.15	−1.57	205.11	−18.20	2.70	−0.02	−0.41	3
UBC18	97.59	−39.65	247.88	−20.72	1.40	0.91	6.70	2
UBC19	56.63	29.93	162.35	−19.22	2.70	2.39	−4.56	2
UBC20	278.66	−13.77	18.77	−2.59	0.50	−0.13	−2.13	2
UBC21	130.06	−21.06	244.72	12.45	1.18	−6.13	2.40	2
UBC22	90.00	14.14	194.46	−4.62	0.66	0.06	−2.93	1
UBC23	252.57	−4.79	13.50	24.14	1.76	−4.41	−6.76	1
UBC24	256.48	1.26	21.39	23.91	2.02	−3.66	−1.65	1
UBC25	257.20	−17.50	4.98	13.31	1.20	−4.20	−4.87	1
UBC26	285.49	22.05	53.83	7.66	1.63	2.07	−5.44	1
UBC27	294.30	15.57	51.98	−2.72	0.85	−1.36	−5.90	1
UBC28	332.41	66.51	107.78	8.53	1.02	−4.34	−3.39	1
UBC29	129.43	−16.54	240.57	14.58	1.21	−6.38	2.13	1
UBC30	3.15	73.14	120.08	10.49	1.12	2.10	0.62	1
UBC31	61.06	32.14	163.74	−15.04	2.85	3.69	−5.04	1

Notes. The parameters are the mean of the members found with TGAS. $N_{\rm found}$ refers to the times each cluster has been found within the explored parameters ($L, minPts$). UBC stands for University of Barcelona Cluster. [(a)]is RSG4 in Röser et al. (2016). [(b)]probably related to Collinder 135. [(c)]is RSG3 in Röser et al. (2016).

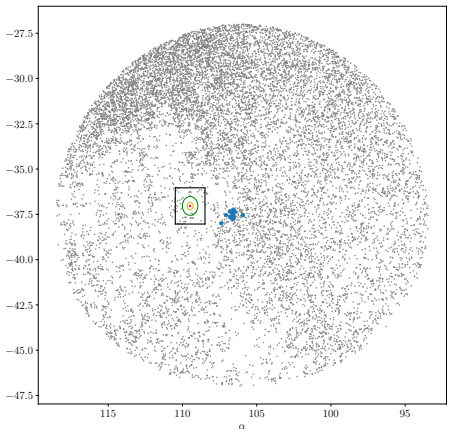


Fig. 7. Cone search of 10 deg centred in UBC7 in the TGAS data with more than 120 photometric observations. Blue dots represent members of UBC7. The red, yellow, and green circles represent the r_0, r_1 and r_2 radius in the MWSC catalogue for Collinder 135. The black box is the 2 deg × 2 deg zone where all candidate clusters are considered as known clusters. The visible stripes on the data are due to the *Gaia* scanning law.

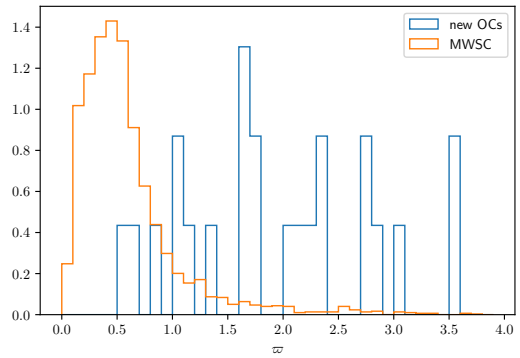


Fig. 8. Normalized parallax distribution of the found OCs (blue) and the ones listed in MWSC (orange). The newly detected OCs are closer than most of the catalogued clusters in MWSC.

limits obtained in the simulations, where we have around 25% and 50% in terms of noise (see Sect. 3).

In the following sections we make comments on some of the confirmed clusters.

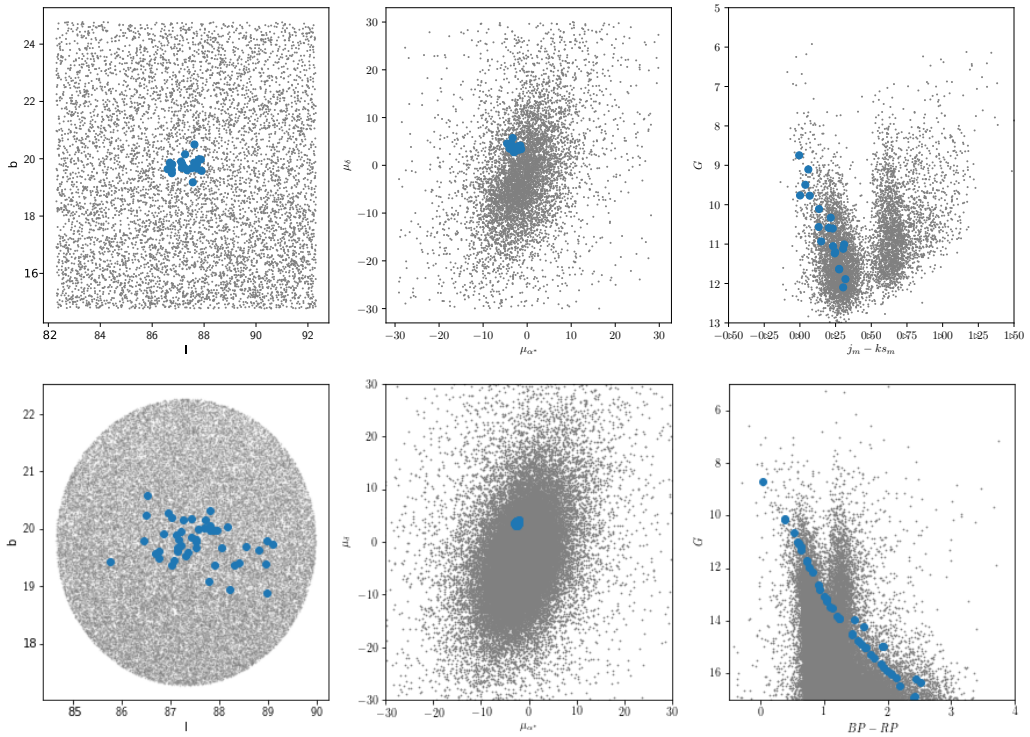


Fig. 9. Visualisation of UBC1 from Table 1. *Top panels left plot*: position of the member stars (blue) along with field stars (grey) in a 10 deg × 10 deg area in TGAS data. *Middle plot*: same stars in the proper motion space. *Right plot*: CMD of the stars in the field using photometry from *Gaia* and 2MASS; member stars follow an isochrone. *Bottom panels*: equivalent for *Gaia* DR2 data. The major difference is in the CMD, where the members detected in *Gaia* DR2 are clearly following an isochrone due to the better quality of the photometric *Gaia* data.

5.1. General comments

The confirmed OCs are distributed on the Galactic disc, and they tend to be at galactic latitudes $|b| > 5$ deg. Figure 10 shows the distribution of the found OCs together with the ones listed in MWSC. They are also nearby compared to those in MWSC (see Fig. 8), most of them within 1 kpc with the exception of UBC3, UBC6, and UBC27 which are detected with parallaxes of 0.58 ± 0.04 mas, 0.67 ± 0.01 mas and 0.88 ± 0.03 mas, respectively.

5.2. UBC1 and UBC12

As mentioned in Sect. 4, UBC1 and UBC12 are RSG4 and RSG3, respectively, in Röser et al. (2016). They are located at about 330 and 430 pc, respectively. There is relatively good agreement in terms of proper motions of RSG3. On the contrary, for RSG4, the values are significantly discrepant at the level of 12σ.

5.3. UBC3

UBC3 is also a poor cluster located at about 1.7 kpc, the farthest cluster among our confirmed candidates. The presence of stars in the red clump area indicates an intermediate age cluster. There are only two stars with radial velocity in DR2 and both are in disagreement. One of those stars is also discordant in terms of its position in the CMD. This could be indicative of a non-membership.

5.4. UBC4, UBC19, and UBC31

UBC19 and UBC31 have proper motions and parallaxes compatible with being substructures of the association Per OB2, if we accept sizes of more than 8 deg for the association. Whether or not they are part of Per OB2 should be investigated through a deep study of a large area. UBC19 has a celestial position near to Alessi Teustch 10 cluster in Dias et al. (2002), but their proper motions do not match. UBC4 has similar parameters but lies slightly farther at about 570 pc.

5.5. UBC7 and Collinder 135

Gaia DR2 data allow us to study UBC7 and Collinder 135 at fainter magnitudes than TGAS. The DR2 data do not show the scanning law pattern that TGAS shows, and still we see two concentrations on the sky (see Fig. 11) with slightly different mean proper motions and parallaxes. The values of the mean and error of the mean for UBC7 are $(\mu_{\alpha^*}, \mu_\delta) = (-9.74 \pm 0.02,\ 6.99 \pm 0.02)$ mas yr^{-1} and $\varpi = 3.563 \pm 0.006$ mas and for Collinder 135 are $(\mu_{\alpha^*}, \mu_\delta) = (-10.09 \pm 0.02,\ 6.20 \pm 0.03)$ mas yr^{-1} and $\varpi = 3.310 \pm 0.004$ mas (computed with the members found with the method described in this paper). To discard possible artefacts due to effects of regional systematic error (Lindegren et al. 2018), we have used the photometry and inspected the CMDs. The sequences overlap, revealing the fact that both clusters have the same age or very similar. When apparent magnitudes are converted into absolute magnitudes using the individual parallaxes of the stars, the overlap of the two

Table 2. List of the confirmed OCs.

Name	α (deg)	δ (deg)	l (deg)	b (deg)	ϖ (deg)	μ_{α^*} (mas yr^{-1})	μ_δ (mas yr^{-1})	$V_{\rm rad}$ (km s^{-1})	N ($N_{V_{\rm rad}}$)
UBC1	288.00 (0.84)	56.83 (0.63)	87.55 (0.74)	19.76 (0.35)	3.05 (0.02)	−2.49 (0.25)	3.69 (0.24)	−21.46 (2.36)	47 (14)
UBC2	5.80 (0.84)	46.59 (0.34)	117.89 (0.62)	−15.99 (0.32)	1.74 (0.03)	−6.34 (0.12)	−5.03 (0.13)	−9.73 (2.22)	23 (4)
UBC3	283.77 (0.16)	12.34 (0.22)	44.35 (0.24)	4.79 (0.12)	0.58 (0.04)	−0.60 (0.08)	−1.36 (0.09)	−7.25 (13.54)	29 (2)
UBC4	60.96 (1.07)	35.35 (0.74)	161.42 (1.05)	−12.75 (0.50)	1.64 (0.05)	−0.75 (0.13)	−5.72 (0.13)	3.67 (1.65)	44 (3)
UBC5	238.42 (0.74)	−47.72 (0.41)	331.74 (0.56)	4.68 (0.32)	1.78 (0.01)	−6.69 (0.15)	−4.18 (0.09)	−14.91 (−)	29 (1)
UBC6	343.95 (0.48)	51.19 (0.19)	105.13 (0.29)	−7.63 (0.21)	0.67 (0.01)	−4.64 (0.06)	−4.90 (0.08)	−31.64 (1.51)	76 (3)
UBC7	106.92 (0.61)	−37.74 (0.65)	248.80 (0.71)	−13.25 (0.42)	3.56 (0.05)	−9.74 (0.19)	6.99 (0.20)	16.42 (4.71)	77 (21)
UBC8	84.36 (0.86)	57.16 (0.54)	154.83 (0.64)	13.30 (0.36)	2.05 (0.03)	−3.14 (0.17)	−3.99 (0.16)	−5.96 (3.94)	103 (21)
UBC9	276.64 (0.41)	26.40 (0.39)	54.48 (0.40)	16.80 (0.38)	2.87 (0.02)	0.60 (0.16)	−5.35 (0.18)	−17.98 (3.12)	25 (6)
UBC10a	324.46 (1.36)	61.75 (0.95)	102.03 (1.02)	7.02 (0.55)	1.07 (0.01)	−2.14 (0.11)	−3.03 (0.12)	−23.12 (−)	43 (1)
UBC10b	326.87 (0.96)	61.10 (0.47)	102.49 (0.36)	5.75 (0.55)	1.01 (0.01)	−3.46 (0.09)	−1.86 (0.10)	−46.90 (−)	40 (1)
UBC11	246.16 (1.91)	−59.94 (0.87)	326.81 (1.15)	−7.39 (0.61)	2.13 (0.04)	−0.30 (0.37)	−6.78 (0.28)	−18.18 (5.35)	44 (4)
UBC12	126.13 (0.65)	−8.56 (0.47)	231.81 (0.71)	16.24 (0.41)	2.21 (0.05)	−8.27 (0.20)	4.07 (0.28)	31.34 (−)	19 (1)
UBC13	120.90 (0.79)	3.60 (1.14)	218.04 (1.02)	17.68 (0.99)	1.60 (0.04)	−7.76 (0.19)	−1.16 (0.21)	22.91 (5.48)	36 (6)
UBC14	294.80 (0.58)	3.64 (1.01)	41.70 (1.06)	−8.91 (0.52)	1.30 (0.02)	0.14 (0.16)	−2.09 (0.20)	−9.85 (−)	46 (1)
UBC17a	83.38 (0.22)	−1.58 (0.86)	205.23 (1.04)	−18.01 (1.06)	2.74 (0.04)	1.59 (0.27)	−1.20 (0.35)	18.96 (7.64)	180 (18)
UBC17b	83.35 (0.76)	−1.54 (0.94)	205.18 (0.95)	−18.02 (0.79)	2.36 (0.04)	0.05 (0.17)	−0.16 (0.24)	33.19 (4.41)	103 (4)
UBC19	56.48 (0.37)	29.91 (0.22)	162.25 (0.24)	−19.32 (0.32)	2.39 (0.11)	2.71 (0.53)	−5.19 (0.27)	31.38 (3.46)	34 (2)
UBC21	130.35 (0.81)	−20.68 (0.94)	244.56 (1.10)	12.87 (0.55)	1.12 (0.02)	−6.51 (0.22)	2.48 (0.17)	− (−)	47 (0)
UBC26	285.24 (0.69)	21.92 (0.74)	53.61 (0.86)	7.80 (0.49)	1.66 (0.03)	2.01 (0.17)	−5.18 (0.21)	6.79 (17.43)	64 (2)
UBC27	294.31 (0.25)	15.58 (0.25)	52.00 (0.24)	−2.73 (0.25)	0.88 (0.03)	−0.82 (0.07)	−6.22 (0.08)	− (−)	65 (0)
UBC31	61.11 (1.21)	32.76 (1.13)	163.33 (1.04)	−14.55 (1.14)	2.70 (0.07)	3.77 (0.22)	−5.43 (0.24)	22.74 (5.73)	84 (12)
UBC32	279.43 (0.66)	−14.04 (0.93)	18.87 (0.96)	−3.38 (0.60)	3.56 (0.04)	−1.75 (0.26)	−9.26 (0.29)	−21.58 (7.24)	60 (14)

Notes. The parameters are the mean (and standard deviation) of the members found with *Gaia* DR2. We also include radial velocity for those stars available. N refers to the number of members found (and members to compute mean radial velocity).

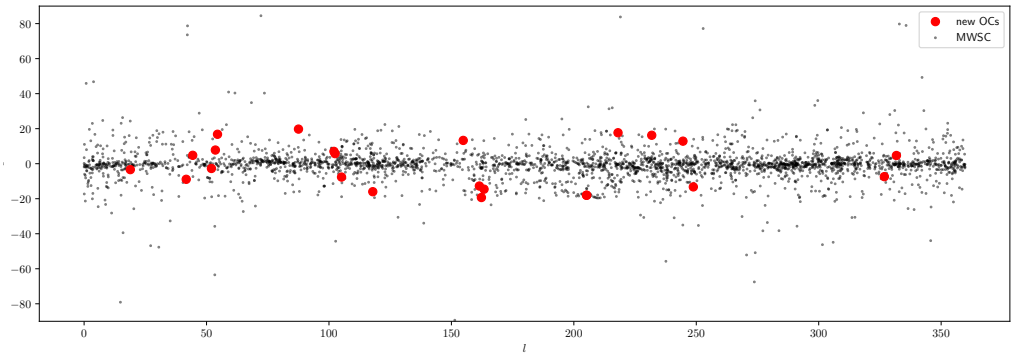


Fig. 10. Spatial distribution in (l, b) of the found OCs (red) together with the ones listed in MWSC (black). The confirmed OCs tend to be at latitudes $|b| > 5$ deg.

sequences is even greater. This confirms that the difference in parallax is a true difference and not an artefact.

Given the differences in proper motions and parallaxes and given the separation in the sky, we therefore conclude that UBC7 and Collinder 135 are two distinct groups, most probably formed in the same process given the similarity of their ages.

5.6. UBC10

This is a rather sparse cluster according to the members derived for the analysis in an area of 1 deg radius with *Gaia* DR2. In addition, the celestial position and parallax of this cluster indicate a potential relationship with the Cep OB2 association. Therefore, we have explored a larger area of 2 deg and there are several subgroups of proper motions and parallaxes certainly distributed towards the position of Cep OB2. A global analysis of an even larger area would confirm or discard the existence of new subgroups in this association.

5.7. UBC17

The large sample of stars of *Gaia* DR2 with respect to TGAS has revealed two groups of proper motions and parallaxes. The distances and proper motions relate them to the Ori OB1 association. Exploring a larger area of 2 deg we can identify ACCC19, Collinder 170, and sigma Ori clusters. This is an indication of the rich structure of the region and so a global analysis of an even larger area encompassing the whole Ori OB1 association is needed, which is, however, beyond the scope of this paper.

5.8. UBC32

UBC20 TGAS DBSCAN candidate cluster was located at a parallax of about 0.5 mas. However, during the analysis of *Gaia* DR2, although such a cluster was not found, a clear detection at a parallax of 3.5 mas has been revealed. It is poor and sparse, and

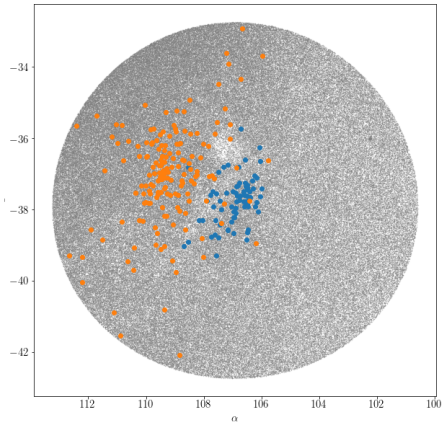


Fig. 11. Cone search of 5 deg in the area of UBC7 (blue) and Collinder 135 (orange). The grey dots correspond to the stars brighter than $G = 17$ mag with more than 120 photometric observations in *Gaia* DR2 data. We have checked that the lower stellar density between the two clusters only appears for parallaxes smaller than 1.5 mas, meaning that it is caused by dust in the background and does not impact our results for the clusters.

decentred with respect to the studied area towards lower galactic latitudes.

6. Conclusions

We have designed, implemented, and tested an automated data-mining system for the detection of OCs using astrometric data. The method is based on i) DBSCAN, an unsupervised learning algorithm to find groups of stars in a N-dimensional space (our implementation uses five parameters l, b, ϖ, μ_{α^*}, μ_δ) and ii) an ANN trained to distinguish between real OCs and spurious statistical clusters by analysis of CMDs. This system is designed to work with minimal manual intervention for its application to large datasets, and in particular to the *Gaia* second data release, *Gaia* DR2.

In this paper, we have tuned and tested the performance of the method by running it using the simulated data and the TGAS dataset, which is small enough to manually check the results. This execution has generated a list of detections that, after removal of know OCs from MWSC, contains 31 new candidates. Using *Gaia* DR2 data we manually examined these candidates and confirmed around 70% of them as OCs, with 100% success in $N_{\rm found} > 5$. In addition, in the confirmation step, we are able to spot richer structures, in particular regions that require further study.

From this exercise, we have confirmed that our method can reliably detect OCs. We have also shown that the TGAS data contain some artefacts due to the nature of the *Gaia* scanning law. We expect these effects to be much reduced (but not completely removed) in *Gaia* DR2, which includes the observations of 22 months of data and where the sky coverage is much more uniform (see Lindegren et al. 2018). Also, the bright limiting magnitude of TGAS prevented the detection of distant (and therefore faint) clusters, which will be detected with the much deeper *Gaia* DR2 data.

Finally, the method leads to reliable results, but we have also identified some limitations. On the one hand, the representativeness of the training dataset for the ANN is crucial to distinguish real and non-real OCs, and we need to build a wider and more realistic training set of CMDs of OCs to use with *Gaia* DR2. On the other hand, since OCs appear more compact or more sparse depending on their distance, there is not a universal value of the ϵ parameter in DBSCAN that can allow the detection of all of them. Therefore, this parameter needs to be adapted to the different possible characteristics of OCs in DR2.

Acknowledgements. This work has made use of results from the European Space Agency (ESA) space mission *Gaia*, the data from which were processed by the *Gaia Data Processing and Analysis Consortium* (DPAC). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement. The *Gaia* mission website is http://www.cosmos.esa.int/gaia. The authors are current or past members of the ESA *Gaia* mission team and of the *Gaia* DPAC. This work was supported by the MINECO (Spanish Ministry of Economy) through grant ESP2016-80079-C2-1-R (MINECO/FEDER, UE) and ESP2014-55996-C2-1-R (MINECO/FEDER, UE) and MDM-2014-0369 of ICCUB (Unidad de Excelencia "María de Maeztu"). This research has made use of the TOPCAT (Taylor 2005). This research has made use of the VizieR catalogue access tool, CDS, Strasbourg, France. The original description of the VizieR service was published in A&AS, 143, 23.

References

Arenou, F., & Luri, X. 1999, in Harmonizing Cosmic Distance Scales in a Post-HIPPARCOS Era, eds. D. Egret & A. Heck, ASP Conf. Ser., 167, 13
Arenou, F., Luri, X., Babusiaux, C., et al. 2017, A&A, 599, A50
Bishop, C. M. 1995, Neural Networks for Pattern Recognition (New York, NY, USA: Oxford University Press, Inc.)
Brescia, M., Cavuoti, S., Paolillo, M., Longo, G., & Puzia, T. 2012, MNRAS, 421, 1155
Caballero, J. A., & Dinis, L. 2008, Astron. Nachr., 329, 801
Dias, W. S., Alessi, B. S., Moitinho, A., & Lépine, J. R. D. 2002, A&A, 389, 871
Duda, R. O., Hart, P. E., & Stork, D. G. 2000, Pattern Classification, 2nd edn. (Wiley-Interscience)
Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. 1996, in Proc. of the Second International Conf. on Knowledge Discovery and Data Mining, KDD'96 (AAAI Press), 226
Froebrich, D. 2017, MNRAS, 469, 1545
Gaia Collaboration (Brown, A. G. A., et al.) 2016, A&A, 595, A2
Gaia Collaboration (van Leeuwen, F. et al.) 2017, A&A, 601, A19
Gaia Collaboration (Brown, A. G. A., et al.) 2018, A&A, 616, A1
Gao, X.-H., Chen, L., & Hou, Z.-J. 2014, Chin. Astron. Astrophys., 38, 257
Gao, X. H., Wang, C., Gu, X. Q., & Xu, S. K. 2017, Acta Astron. Sin., 58, 46
Hinton, G. 1989, Artif. Intell., 40, 185
Kharchenko, N. V., Piskunov, A. E., Schilbach, E., Röser, S., & Scholz, R.-D. 2013, A&A, 558, A53
Lada, E. A., Strom, K. M., & Myers, P. C. 1993, in Protostars and Planets III, eds. E. H. Levy & J. I. Lunine, 245
Lindegren, L., Lammers, U., Hobbs, D., et al. 2012, A&A, 538, A78
Lindegren, L., Lammers, U., Bastian, U., et al. 2016, A&A, 595, A4
Lindegren, L., Hernandez, J., Bombrun, A., et al. 2018, A&A, 616, A2
Luri, X., Palmer, M., Arenou, F., et al. 2014, A&A, 566, A119
Michalik, D., Lindegren, L., & Hobbs, D. 2015, A&A, 574, A115
Pedregosa, F., Varoquaux, G., Gamfort, A., et al. 2011, J. Mach. Learn. Res., 12, 2825
Pickles, A., & Depagne, E. 2011, VizieR Online Data Catalog: VI/135
Roelens, M. 2013, Gaia Capabilities for the Study of Open Clusters, Universitat de Barcelona, 2013, 16, http://archives.esf.org/coordinating-research/research-networking-programmes/physical-and-engineering-sciences-pen/current-research-networking-programmes/gaia-research-for-european-astronomy-training-great/scientific-activities.html
Röser, S., Schilbach, E., & Goldman, B. 2016, A&A, 595, A22
Schmeja, S., Kharchenko, N. V., Piskunov, A. E., et al. 2014, A&A, 568, A51
Scholz, R.-D., Kharchenko, N. V., Piskunov, A. E., Röser, S., & Schilbach, E. 2015, A&A, 581, A39
Taylor, M. B. 2005, in Astronomical Data Analysis Software and Systems XIV eds. P. Shopbell, M. Britton, & R. Ebert, ASP Conf. Ser., 347, 29
Wilkinson, S., Merín, B., & Riviere-Marichalar, P. 2018, A&A, 618, A12
Yèche, C., Petitjean, P., Rich, J., et al. 2010, A&A, 523, A14

Appendix A: Colour-magnitude diagrams of the identified open clusters

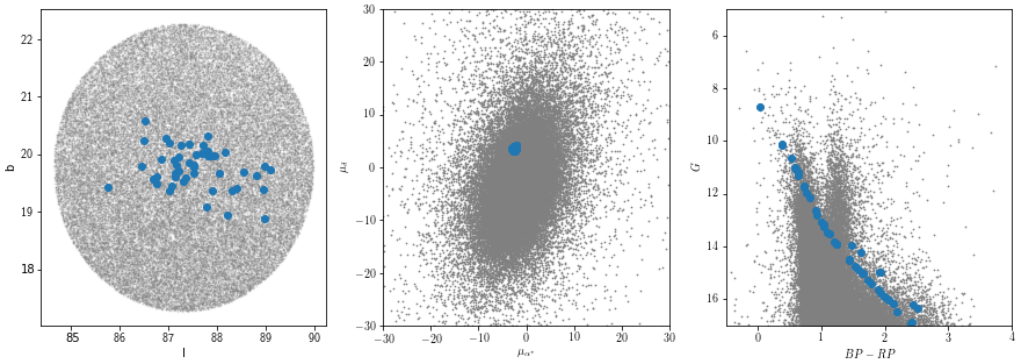


Fig. A.1. Member stars (blue) together with field stars (grey) for UBC1 in (l, b) (*left panel*) and in proper motion space (*middle panel*). The CMD shows the sequence of the identified members (outlining an empirical isochrone) (*right panel*).

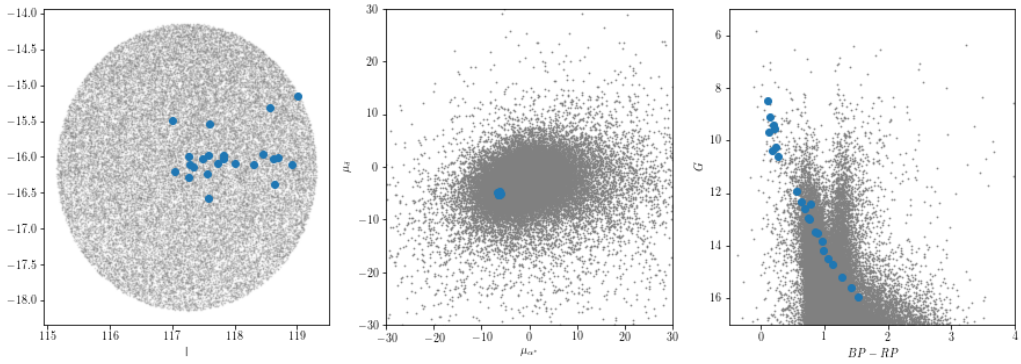


Fig. A.2. As in Fig. A.1 but for UBC2.

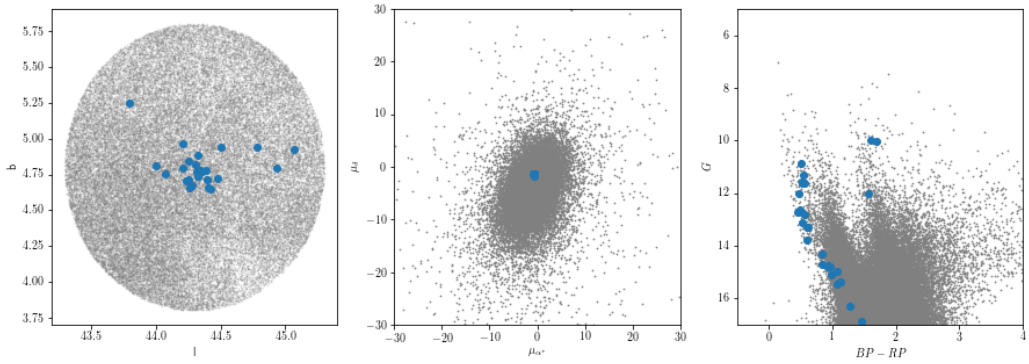


Fig. A.3. As in Fig. A.1 but for UBC3.

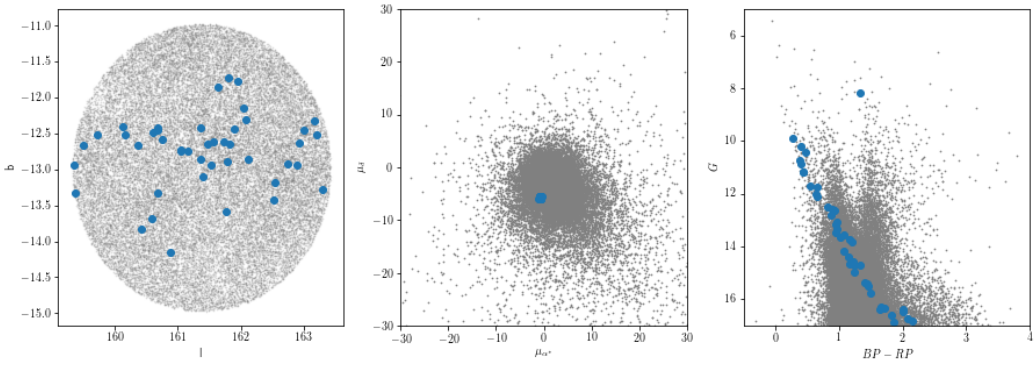


Fig. A.4. As in Fig. A.1 but for UBC4.

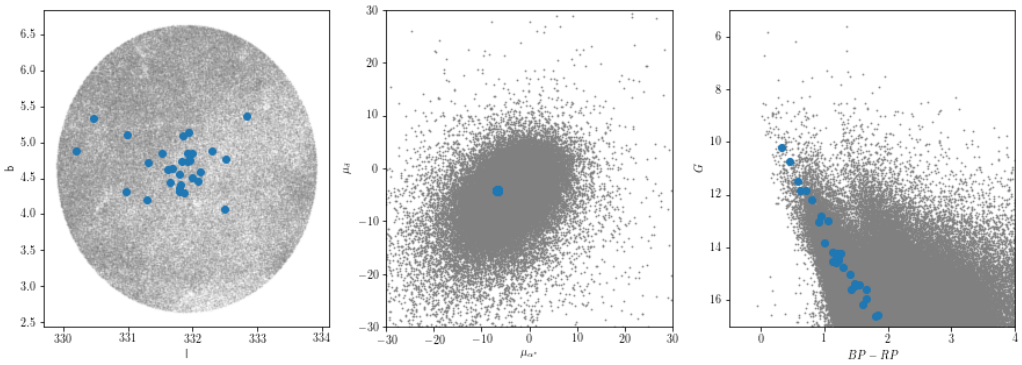


Fig. A.5. As in Fig. A.1 but for UBC5.

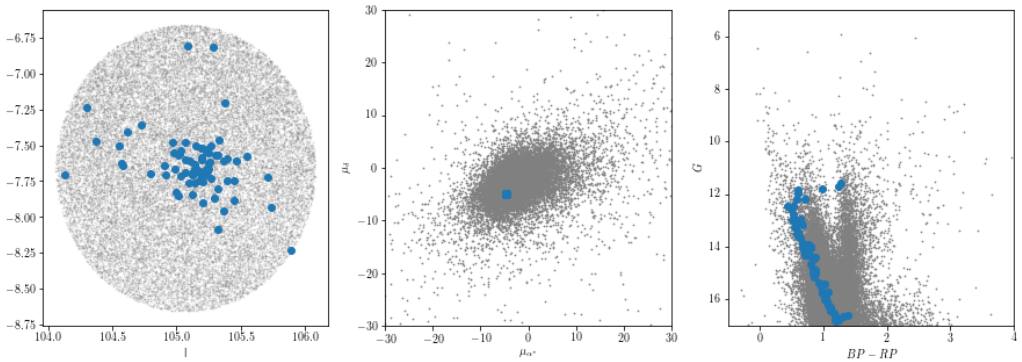


Fig. A.6. As in Fig. A.1 but for UBC6.

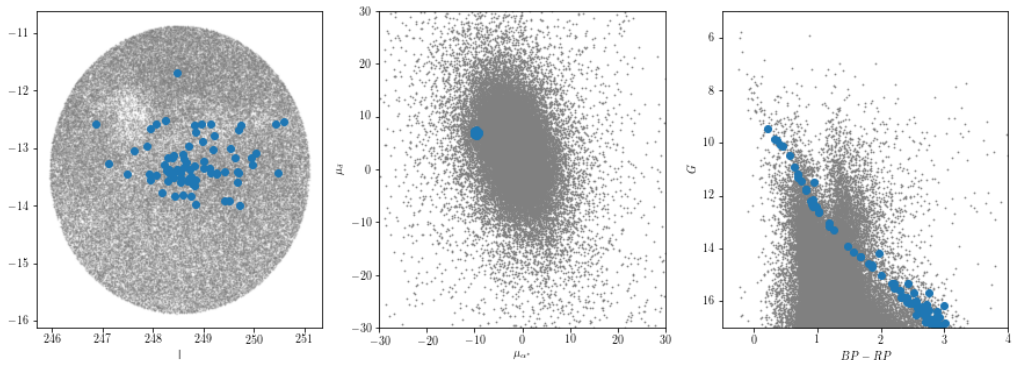


Fig. A.7. As in Fig. A.1 but for UBC7.

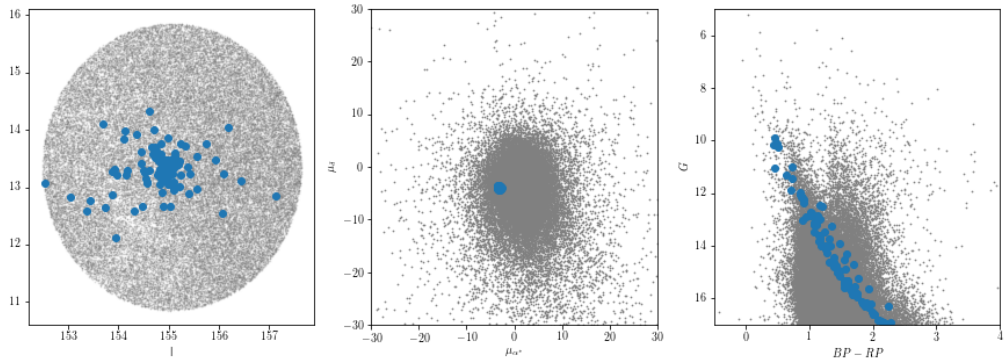


Fig. A.8. As in Fig. A.1 but for UBC8.

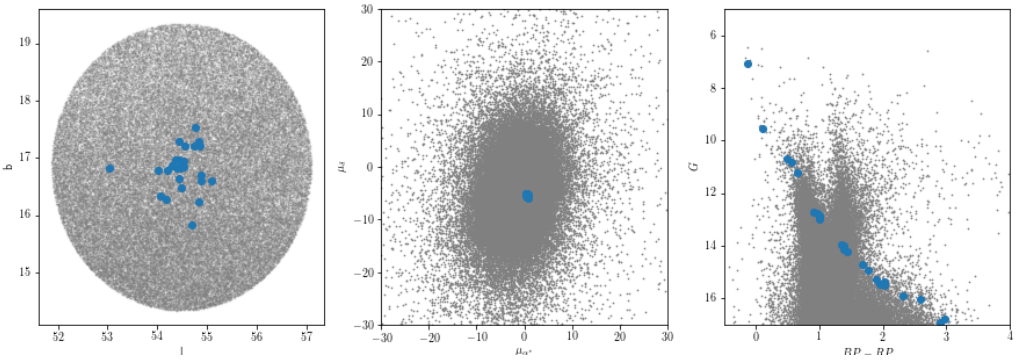


Fig. A.9. As in Fig. A.1 but for UBC9.

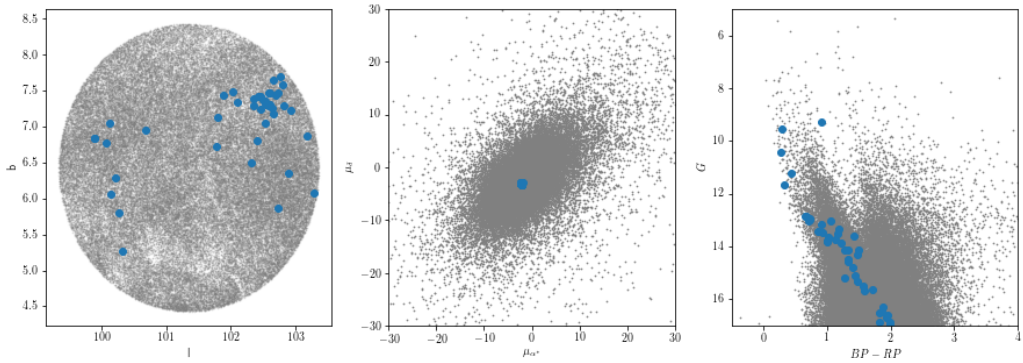


Fig. A.10. As in Fig. A.1 but for UBC10a.

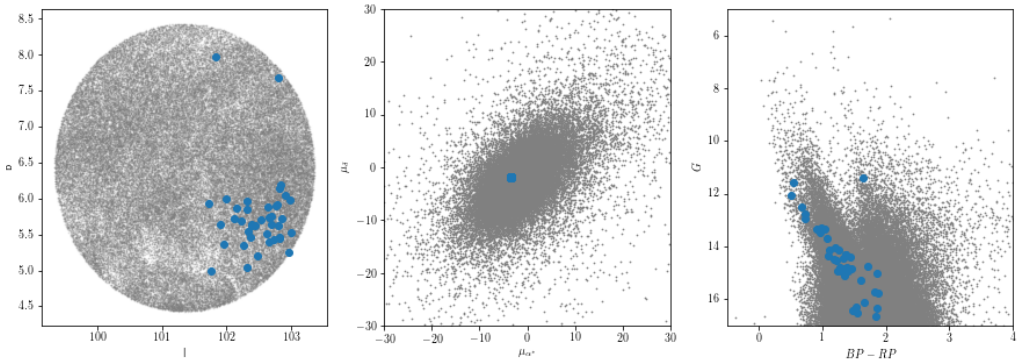


Fig. A.11. As in Fig. A.1 but for UBC10b.

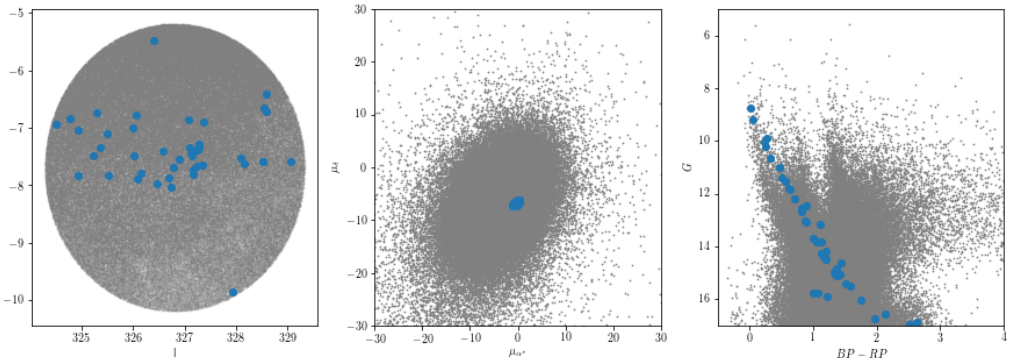


Fig. A.12. As in Fig. A.1 but for UBC11.

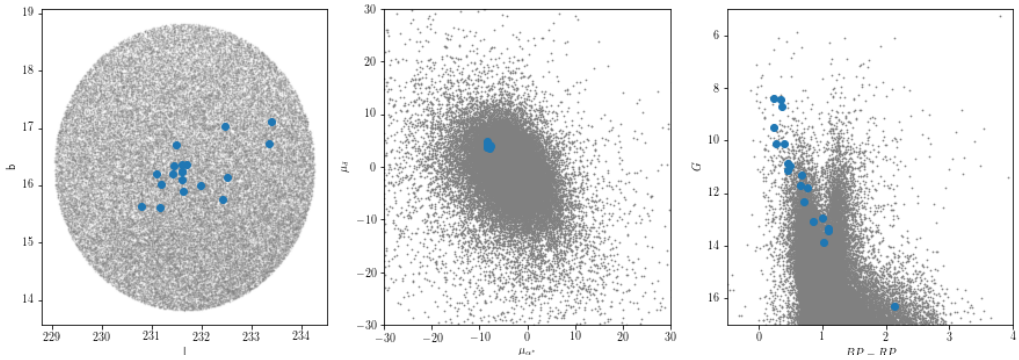


Fig. A.13. As in Fig. A.1 but for UBC12.

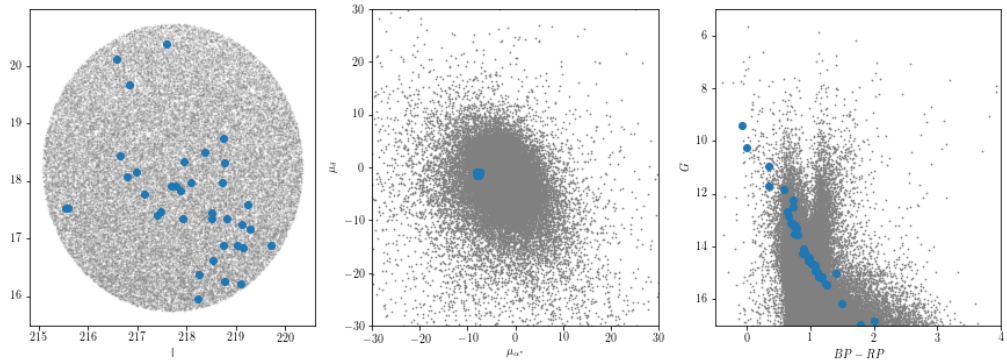


Fig. A.14. As in Fig. A.1 but for UBC13.

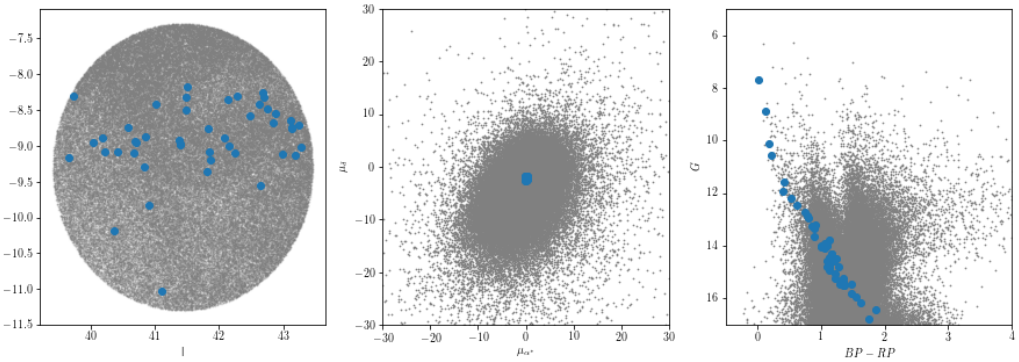


Fig. A.15. As in Fig. A.1 but for UBC14.

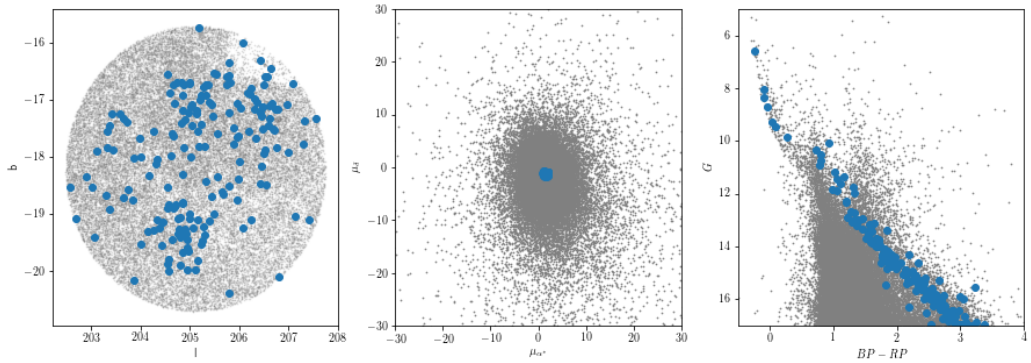


Fig. A.16. As in Fig. A.1 but for UBC17a.

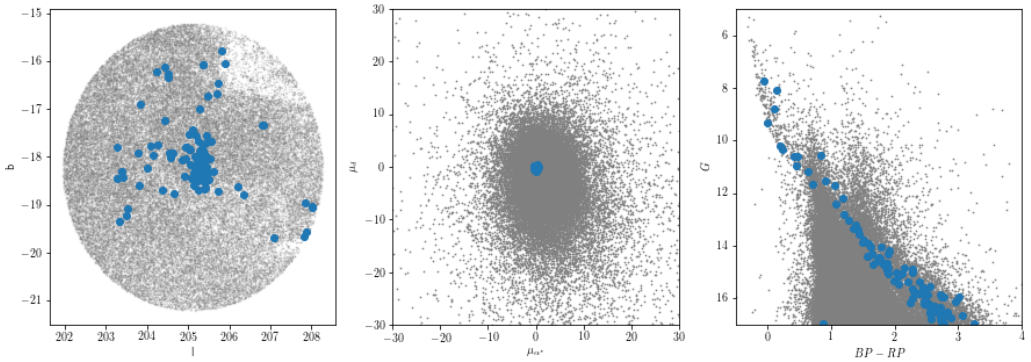


Fig. A.17. As in Fig. A.1 but for UBC17b.

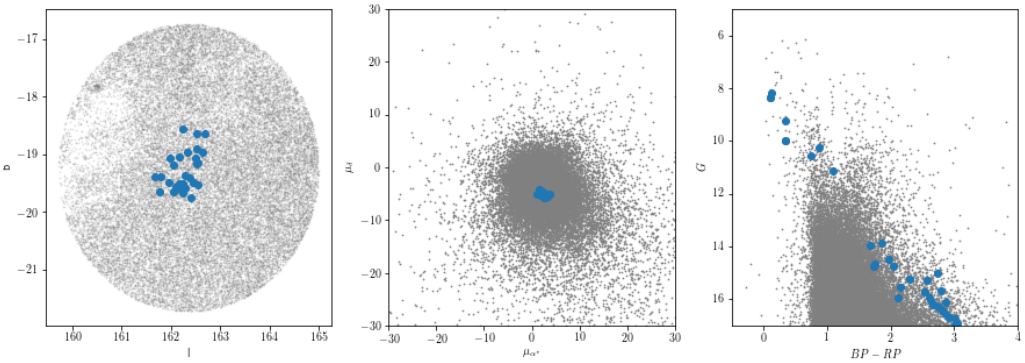


Fig. A.18. As in Fig. A.1 but for UBC19.

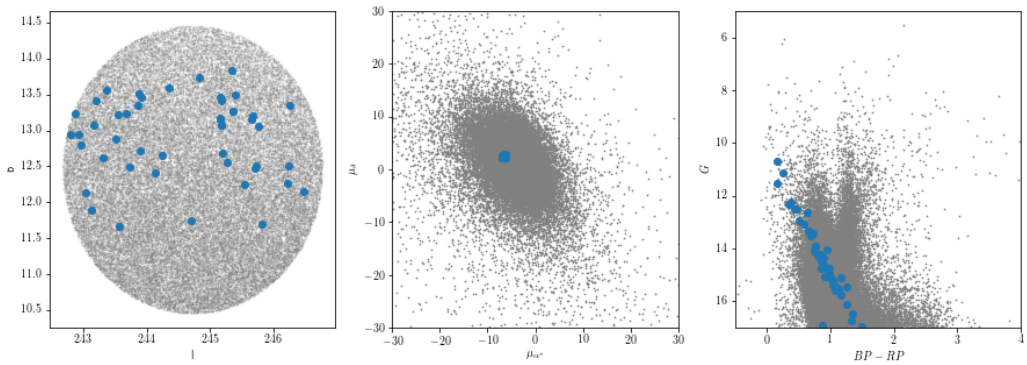


Fig. A.19. As in Fig. A.1 but for UBC21.

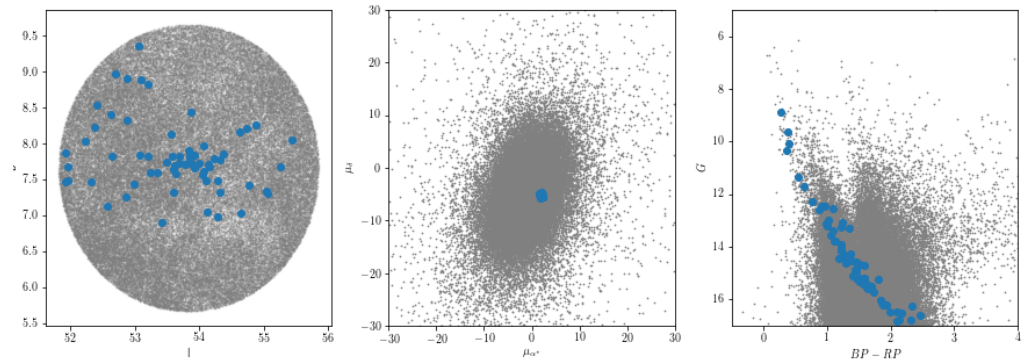


Fig. A.20. As in Fig. A.1 but for UBC26.

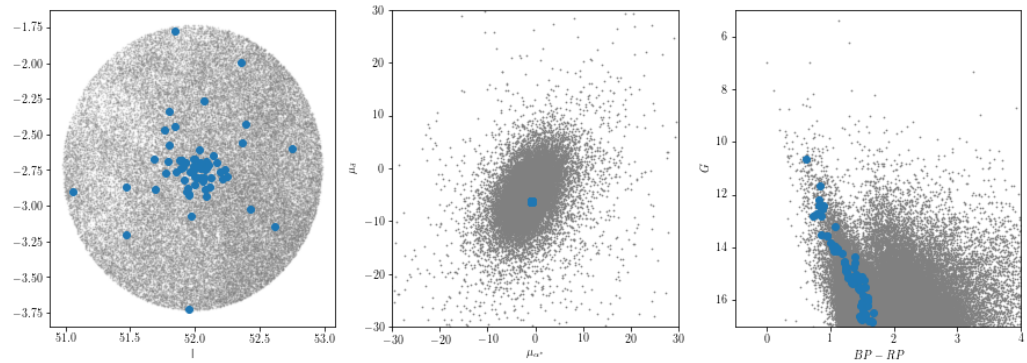


Fig. A.21. As in Fig. A.1 but for UBC27.

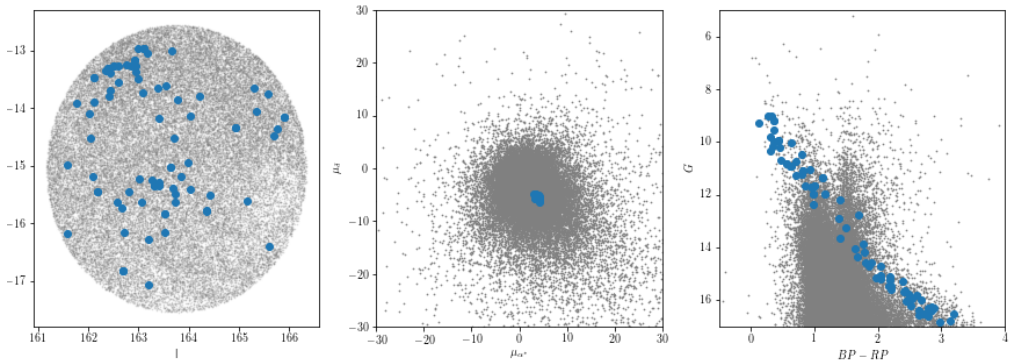


Fig. A.22. As in Fig. A.1 but for UBC31.

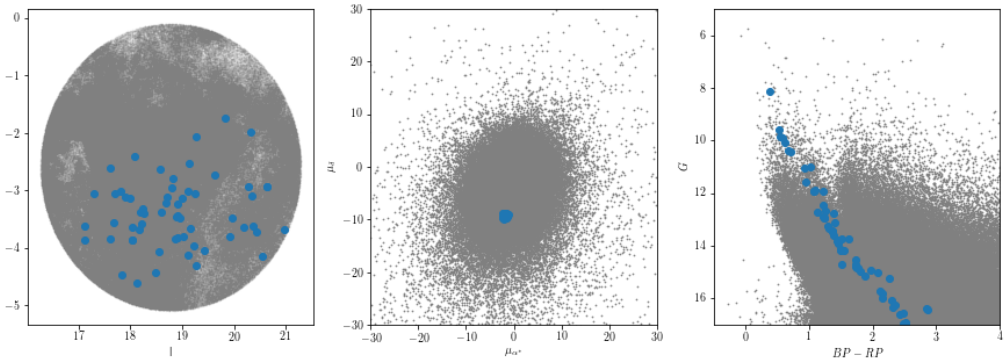


Fig. A.23. As in Fig. A.1 but for UBC32.

3

THE GALACTIC ANTICENTER

This Chapter contains the published version of Castro-Ginard et al. (2019, A&A, 627, A35).

The content of the Chapter corresponds to the adaptation of the methodology presented in Chapter 2 to deal with the analysis of the *Gaia* DR2 (Gaia Collaboration et al. 2018).

The publication of *Gaia* DR2, with its unprecedented volume and precision of astrometric and photometric data (1.3 billion sources, with a parallax uncertainty of 0.1 mas), represented a huge step forward in the study of our Galaxy. The usual methodologies had to be adapted to work with high differences in the density for different regions of the Galaxy that are present in the *Gaia* catalogue. In this Chapter, we search for open clusters in a particularly low density disc region, the Galactic anticentre and the Perseus arm ($120^\circ \leq l \leq 205^\circ$ and $-10^\circ \leq b \leq 10^\circ$). This first study allowed us to know the strengths and limitations of our methodology, and prepared the methodology to be applied to the whole Galactic disc.

In this study, we detected 53 new OCs, some of them independently discovered by Cantat-Gaudin et al. (2019b) using a different methodology and showing the complementarity of different methodologies for the same purpose. This number of new OCs represented an increase of 22% on the previously known census in this region, and included objects closer than 2 kpc which as in Chapter 2 challenge the idea of completeness of the OC population in that volume.

A&A 627, A35 (2019)
https://doi.org/10.1051/0004-6361/201935531
© ESO 2019

Astronomy & Astrophysics

Hunting for open clusters in *Gaia* DR2: the Galactic anticentre⋆

A. Castro-Ginard, C. Jordi, X. Luri, T. Cantat-Gaudin, and L. Balaguer-Núñez

Dept. Física Quàntica i Astrofísica, Institut de Ciències del Cosmos (ICCUB), Universitat de Barcelona (IEEC-UB), Martí i Franquès 1, 08028 Barcelona, Spain
e-mail: acastro@fqa.ub.edu

Received 25 March 2019 / Accepted 14 May 2019

ABSTRACT

Context. The *Gaia* Data Release 2 (DR2) provided an unprecedented volume of precise astrometric and excellent photometric data. In terms of data mining the *Gaia* catalogue, machine learning methods have shown to be a powerful tool, for instance in the search for unknown stellar structures. Particularly, supervised and unsupervised learning methods combined together significantly improves the detection rate of open clusters.
Aims. We systematically scan *Gaia* DR2 in a region covering the Galactic anticentre and the Perseus arm ($120° \leq l \leq 205°$ and $-10° \leq b \leq 10°$), with the goal of finding any open clusters that may exist in this region, and fine tuning a previously proposed methodology and successfully applied to TGAS data, adapting it to different density regions.
Methods. Our methodology uses an unsupervised, density-based, clustering algorithm, DBSCAN, that identifies overdensities in the five-dimensional astrometric parameter space $(l, b, \varpi, \mu_{\alpha^*}, \mu_\delta)$ that may correspond to physical clusters. The overdensities are separated into physical clusters (open clusters) or random statistical clusters using an artificial neural network to recognise the isochrone pattern that open clusters show in a colour magnitude diagram.
Results. The method is able to recover more than 75% of the open clusters confirmed in the search area. Moreover, we detected 53 open clusters unknown previous to *Gaia* DR2, which represents an increase of more than 22% with respect to the already catalogued clusters in this region.
Conclusions. We find that the census of nearby open clusters is not complete. Different machine learning methodologies for a blind search of open clusters are complementary to each other; no single method is able to detect 100% of the existing groups. Our methodology has shown to be a reliable tool for the automatic detection of open clusters, designed to be applied to the full *Gaia* DR2 catalogue.

Key words. surveys – open clusters and associations: general – astrometry – methods: data analysis

1. Introduction

The popularity of machine learning (ML) techniques used to analyse astronomical data is growing, as is the volume of astronomical catalogues. The use of these techniques is mandatory to extract meaningful insight from big data sets such as the second data release of the ESA *Gaia* astrometric mission (*Gaia* DR2, Gaia Collaboration 2016, 2018), which contains more than 550 GB[1] of data, including precise astrometry (Lindegren et al. 2018) and excellent photometry (Evans et al. 2018), among other products, for more than 1.3×10^9 sources down to magnitude $G = 21$ mag. This unprecedented volume of extremely precise data reveals unseen details in the structure of our galaxy.

Open clusters (OCs) are considered as fundamental objects in our understanding of the structure and evolution of the Milky Way disc. The stars of an OC were born and move together; i.e. in terms of *Gaia* observables, they share $(l, b, \varpi, \mu_{\alpha^*}, \mu_\delta)$ and follow a specific pattern in a colour-magnitude diagram (CMD) $(G, G_{\rm BP}, G_{\rm RP})$. That they can represent overdensities in five-dimensional astrometric space can be exploited by unsupervised learning algorithms to either characterise known OCs when looking for new member stars (Gao 2018a,b; Cantat-Gaudin et al. 2018), or to detect new overdensities in the parameter space (Castro-Ginard et al. 2018; Cantat-Gaudin et al. 2019). Supervised learning methods can help in determining whether a group of stars is an OC by identifying the isochrone pattern of its member stars in a CMD, due to the common age of its members. In the OC domain, *Gaia* DR2 represents a perfect scenario for the application of ML methods to both its detection and characterisation.

Our understanding of the OC population has dramatically changed with *Gaia* DR2. A pre-*Gaia* census of the OC population counted around 3000 objects (Dias et al. 2002; Kharchenko et al. 2013; Froebrich et al. 2007; Schmeja et al. 2014; Scholz et al. 2015; Röser et al. 2016) compiled from heterogeneous data sources, making the characterisation of OC parameters a difficult task. After the publication of *Gaia* DR2, Cantat-Gaudin et al. (2018) revisited the OC population using a ML based unsupervised membership determination algorithm. This resulted in the compilation of a homogeneous OC catalogue of 1229 objects, including some serendipitously detected OCs and discarding some objects listed in previous catalogues. These well-determined members and mean astrometrical parameters from the *Gaia* DR2 data allowed the kinematical study of these objects (Soubiran et al. 2018) and the derivation of ages and physical parameters (Bossini et al. 2019). Additionally, the combination of ML techniques and *Gaia* DR2 data triggered the detection of new OCs. The discovery of nearby OCs

⋆ Table 2 is only available at the CDS via anonymous ftp to cdsarc.u-strasbg.fr (130.79.128.5) or via http://cdsarc.u-strasbg.fr/viz-bin/qcat?J/A+A/627/A35

[1] https://www.cosmos.esa.int/web/gaia/dr2

(Castro-Ginard et al. 2018; Cantat-Gaudin et al. 2019), where the census was thought to be complete, showed the necessity to keep exploring the sky for new objects.

In Castro-Ginard et al. (2018, hereafter CG18) we presented a method for the automatic detection of OCs in the *Gaia* data. The method consists in the application of an unsupervised clustering algorithm, DBSCAN, that looks for overdensities in the astrometric five-dimensional space $(l, b, \varpi, \mu_{\alpha^*}, \mu_\delta)$. Once the overdensities are detected, we classify them as either random statistical overdensities or real OCs by identifying the isochrone pattern of OC member stars in a CMD using an artificial neural network (ANN). The method has proved to be successful in the detection of OCs in the TGAS data (Lindegren et al. 2016; Michalik et al. 2015), which were later validated in the *Gaia* DR2 data. In this paper we apply the methodology to a region of the sky around the Galactic anticentre with the aim of increasing our knowledge of the OC population in that region, and fine tuning the methodology for its planned future application in an all sky blind search.

The paper is organised as follows. Section 2 briefly describes the methodology used, which is discussed in detail in CG18. The data set used for the detection is described in Sect. 3. The proposal of new OCs and some comments on the results found are in Sect. 4. Finally, concluding remarks are summarised in Sect. 5.

2. Methodology

This section briefly describes the methodology used in CG18, where our approach to detect OCs in the *Gaia* DR2 data is explained in detail. The method consists of three parts: a preprocessing step, where the data is prepared to be exploited; a density-based clustering algorithm, DBSCAN (Ester et al. 1996), used to look for overdensities in the five-dimensional astrometric data; and a classification of the resulting clusters into real OCs and random statistical clusters using an ANN (Hinton 1989) to recognise the isochrone pattern of the cluster member stars in a CMD.

In the preprocessing step the sky area of study is divided into smaller regions, rectangles of size $L \times L$ deg, in order to compute a representative average star density of the region used to search for overdensities. In each rectangle the parameters used to perform the clustering analysis $(l, b, \varpi, \mu_{\alpha^*}, \mu_\delta)$, are standardised (re-scaled to have zero mean and variance of one) to avoid a preferred dimension and to balance the importance of each dimension on the clustering process.

The detection of statistical clusters is done using the DBSCAN[2] algorithm, which is a density-based algorithm that uses the notion of distance between stars to define close stars as a cluster. The statistical distance between two stars is computed as the Euclidean distance in the standardised five-dimensional parameter space. The reasons for the choice of DBSCAN are twofold. Firstly, it is able to detect arbitrarily shaped clusters, so it accounts, for instance, for the effects of the projection of a cluster location into a two-dimensional sky (l and b). Secondly, it requires only two input parameters: *minPts*, the minimum number of stars needed to be considered a cluster, and ϵ, the radius of the hyper-sphere where we search for these *minPts* stars. The parameter ϵ is automatically computed in each rectangle, assuming that the distance between neighbours in a cluster is smaller than that between field stars (see Sect. 2.2 of CG18).

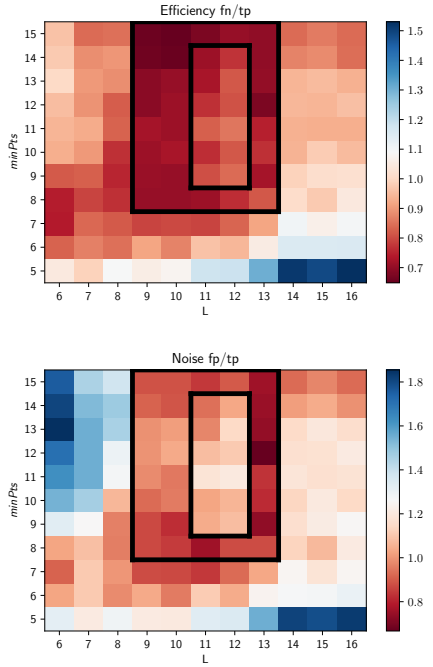


Fig. 1. Pairs of parameters $(L, minPts)$ explored. *Top plot*: efficiency of each pair (false negative – true positive rate). *Bottom plot*: noise (false positive – true positive rate). The redder the pixel, the better the performance of the algorithm in terms of OC detection. The parameters selected, considered optimal for OC detection, are inside the black lines.

The values for the parameters $(L, minPts)$ are set using *Gaia* DR2-like simulated data (see Sect. 3 in CG18), where in this case we added the errors[3] at the time of *Gaia* DR2. Several combinations of optimal parameters $(L, minPts)$ were selected in order to assess the resulting performance of the algorithm; in this case we chose 28 pairs of $(L, minPts)$. Figure 1 shows the pairs of parameters explored and the chosen combination inside the black lines, whose values range within $L \in [9°, 13°]$ and $minPts \in [8, 15]$. These parameters were selected to try to find a balance between low noise and good efficiency, defined as the false positive – true positive ratio for the noise and false negative – true positive for the efficiency.

After the clustering process, the resulting clusters can be either real OCs or random statistical clusters. These two types can be differentiated by the pattern followed by the cluster member stars on a CMD. The classification into real OCs or random statistical clusters is done with an ANN[2] that is able to identify the characteristic shape of isochrones in CMDs corresponding to real OCs. To train the ANN we used CMDs from OCs from the most homogeneous OC catalogue to date (see details in Cantat-Gaudin et al. 2018), which also has the advantage of being compiled from *Gaia* DR2 data so it is representative of the OCs we expect to detect, and with similar photometric errors. The training set consists of a sample of 1229 real OCs. In addition we used data augmentation techniques so the volume of the training set was increased by randomly selecting member stars to create a set of subclusters from each of these catalogued OCs. On the negative identification side, we used CMDs from random

[2] Algorithm from the scikit-learn python package (Pedregosa et al. 2011).

[3] Implementation provided by PyGaia package: https://github.com/agabrown/PyGaia

field stars on the same field as the 1229 OCs to avoid location biases.

As a last step, and to ensure the selection only of newly detected OCs, we removed the already catalogued OCs. We improved this step with respect to CG18 thanks to the compilation of the catalogue by Cantat-Gaudin et al. (2018). In this case positional arguments were used in order to match a found OC with a catalogued one. An OC was considered to be already catalogued if the mean parameters $(l, b, \varpi, \mu_{\alpha^*}, \mu_\delta)$ of its members was compatible within 2σ of the mean parameters of the catalogued OC in Cantat-Gaudin et al. (2018). We did not make a cross-identification with other catalogues such as Dias et al. (2002), Kharchenko et al. (2013) and Bica et al. (2019), and others, due to the inhomogeneous data sources they are compiled from.

3. Data

The *Gaia* catalogue, in its second data release (*Gaia* DR2, Gaia Collaboration 2018), provides precise five-dimensional astrometric data (positions, parallax and proper motions) together with magnitudes in three photometric broad bands ($G, G_{\rm BP}$, and $G_{\rm RP}$) for more than 1.3 billion sources up to $G = 21$ mag. In this work we focus on a region located at the disc ($b \in [-10°, 10°]$) near the Galactic anticentre ($l \in [120°, 205°]$) down to magnitude $G = 17$ mag, where we find a total of 8 715 057 sources with mean standard uncertainties of 0.07 mas for the parallax and 0.1 mas yr^{-1} for proper motions.

We fixed the search region in the Galactic disc because the expectation to find OCs decreases at higher altitudes. For instance, around 93% of the OCs catalogued in Cantat-Gaudin et al. (2018) are at $|b| < 10°$, and around 99% are located at $|b| < 20°$; similar numbers are found in the catalogues of Dias et al. (2002) and Kharchenko et al. (2013) with 96% and 94% of the OCs located at $|b| < 20°$. Moreover, initially the search region was as wide as $|b| < 40°$, but an exploratory analysis of the results of our method showed that the detection of clusters tends to be less reliable at $|b| > 10°$. This effect is shown in Fig. 2; the clusters at $|b| > 10°$ are detected fewer times within the 28 pairs of $(L, minPts)$ explored than those located at the disc, decreasing the reliability of the candidate. In addition, clusters detected outside the disc increase in size with Galactic latitude, so with decreasing stellar density. Since there is no physical reason for this and although we cannot discard that some of these detected clusters may be real, we interpret that the determination of the ϵ parameter for such low density regions is not accurate, and therefore we decided to limit the final search region to the disc, defined as $|b| < 10°$.

The reason for the choice of the region near the Galactic anticentre is twofold. On the computational side, the limited volume of data due to the manageable density of stars in the anticentre direction facilitates its analysis, while keeping the richness of the data up to $G = 17$ mag. On the astrophysical side, objects at a greater distance can be reached due to the moderate extinction caused by interstellar dust, compared to the Galactic centre direction. The search region also covers the area recently studied by Cantat-Gaudin et al. (2019) with *Gaia* DR2 data; they have found 41 new clusters and note that the region $l \in [140°, 160°]$ seems to be devoid of OCs.

4. Results

The method described in Sect. 2 is applied to the *Gaia* DR2 data, focused on a region around the anticentre, i.e. $120° \leq l \leq 205°$, and in the disc, $-10° \leq b \leq 10°$. This results in the detection of 53 OCs that were unknown previous to *Gaia* DR2, which represent an increase of ∼22% with respect to the reference catalogue.

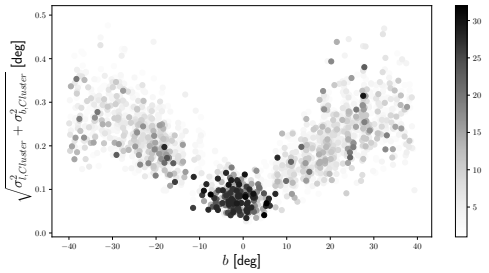


Fig. 2. Cluster size as a function of the Galactic latitude (b). The greyscale represents how many times each cluster is found within the pairs of $(L, minPts)$ explored. High latitude clusters are detected fewer times and are larger in size.

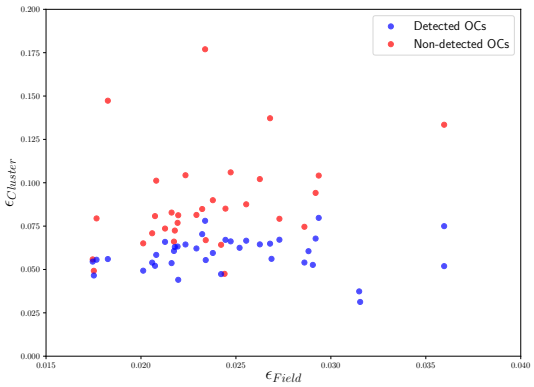


Fig. 3. Parameter ϵ computed for detected (blue) and non-detected (red) OCs in Cantat-Gaudin et al. (2018) as a function of the ϵ computed for the whole field, corresponding to $L = 13°$ and $minPts = 9$.

4.1. Determination of a detection

We can assess the detection criteria by comparing the detected and non-detected OCs from the existing catalogues. In our region of search, Cantat-Gaudin et al. (2018) report 240 OCs of which we were able to recover 182, i.e. ∼76% of the already known OCs. The reason for the non-detection of the remaining ∼24% OCs is related to the contrast of the OC with respect to the field, as seen by the DBSCAN algorithm.

Figure 3 shows a distribution of the ϵ parameter computed for each of the 240 OCs, including the detected and non-detected OCs for $L = 13°$ and $minPts = 9$. The computation of the ϵ parameter, as explained in Sect. 2.2 of CG18, was done via a data-driven approach; for the interpretation of the parameter the whole data set used has to be taken into account and not just the physical properties of the OCs (or the field). In this case, the key factor that enables the detection of the OC is the OC-field contrast in terms of compactness. We see from Fig. 3 that only clusters with low values of ϵ (high contrast) are detected. This is confirmed by the fact that the re-application of the method detects most of the undetected OCs when increasing the contrast with respect to the field by localising the search area to a cone search centred at the targeted OC instead of the large rectangle.

4.2. Proposal of new OCs

The application of our method to the described data set gave us an initial list of 491 OC candidates. The Monte Carlo-type analysis (application of the method for several optimal pairs of parameters) allowed us to assess the reliability of these detections by the number of times each cluster was found. In order to clean the initial list from false positives, we manually inspected each of the OC candidates and tried to re-detect the candidate in a smaller field (cone search around the centre of the targeted OC) where the OC field contrast is higher. This re-detection was done using the DBSCAN algorithm again in a cone search region centred on the targeted OC[4]. The decision on the proposal of the candidate as an OC is made based on the reliability of the candidate and its re-detection.

After this manual step, 53 of these 491 candidates were validated and proposed as OCs. The reason why only 53 OCs were validated is related to the low complexity of the ANN architecture, and the low volume of training data available (based on *Gaia* DR2 data only). This can give false positive identifications, i.e. incorrect identification, of a stellar structure as an OC. In our manual validation step we were conservative, tending to accept as OCs only those groups without a sparse distribution in the sky, with greater compactness in proper motions and parallax, and with better defined sequences in the CMD. This step may have introduced a strong bias in the selection and rejected true clusters. With the improved proper motions and parallaxes of *Gaia* DR3 and a more populated training data set, it will be possible to repeat the analysis to fainter magnitudes and will produce fewer dubious cases. Even though the method is devised to require minimal user intervention, this is an important step as the exploitation of the *Gaia* data in terms of blind search for stellar structures is at its initial stages, so a robust OC catalogue needs to be built to reliably train an automatic detection procedure.

A final list of 53 OCs is proposed, divided into class A and class B depending on the reliability of the candidate. Positions (α, δ) and (l, b) together with mean parameters $(\varpi, \mu_{\alpha^*}, \mu_\delta)$ and mean $V_{\rm rad}$ when available can be found in Table 1 for each of the new OCs, which also includes the computed apparent size of the OC and its estimated distance with a one-sigma (asymmetric) confidence interval. A list of the detected members for all the reported OCs is available in Table 2[5].

4.3. Comments on the detected OCs

The newly found OCs are distributed along the Galactic anticentre direction as shown in Fig. 4, where green crosses represent OCs found in this work, blue triangles are the already catalogued OCs in Cantat-Gaudin et al. (2018) and yellow boxes are the OCs in Cantat-Gaudin et al. (2019). It is worth noting that in a region around $l \sim 140°$ the density of OCs decreases in terms of catalogued clusters and of newly detected ones. This confirms the findings in Cantat-Gaudin et al. (2019) that this region seems to be devoid of OCs. The low OC density is better seen in Fig. 5, where an X-Y projection is shown with the Sun at $(0, 0)$, and it seems to be pointing in the direction of the Perseus arm (Local and Perseus arms follow the model of Reid et al. 2014). This region of relatively low density was first reported as a lack of OB stars in the *Gaia* DR2 data, and dubbed the Gulf of Camelopardalis[6].

The strategy we used to detect OCs relies on the OC field contrast, which is able to detect those OCs with the highest contrast. This may result in a detection bias towards the more compact objects. Figure 6 shows the radius of the detected OC as a function of its distance, which is computed as $1/\overline{\varpi}$ (Luri et al. 2018) given the low parallax relative error ($\overline{\sigma_\varpi} \sim 0.04$ mas corresponding to 3−16% in parallax relative error). The size range of the objects found increases with distance, limiting our detection to very compact objects in a close neighbourhood. The mean size of the detected OCs is $\sigma_l, \sigma_b \sim 0.08°$, and corresponds to an apparent size of $\theta \sim 0.11°$. Our detection limit seems to be at a cluster apparent size of $\theta = 0.2°$.

In terms of estimated distance, we find 6 new OCs within 1 kpc (the closest one at around 645 pc) and 27 within 1.8 kpc, to be added to the 23 found by Castro-Ginard et al. (2018) and the 31 by Cantat-Gaudin et al. (2019) in that distance range, further supporting the claim that more objects are yet to be discovered in this volume, especially with the combination of the excellent *Gaia* data and ML algorithms in future all-sky searches. This challenges the statement that the OC census is complete up to 1.8 kpc (Kharchenko et al. 2013; Matsunaga et al. 2018; Piskunov et al. 2018).

From the kinematical point of view, the reported OCs have a mean dispersion of $\mu_{\alpha^*}, \mu_\delta \sim 0.2$ mas yr^{-1}, computed from the found member stars. This corresponds to a mean tangential velocity dispersion of $\sim$2.2 km s^{-1}. Only 30 of the 53 reported OCs have a radial velocity measurement available in *Gaia* DR2, 11 of which have more than two measurements ($N_{V_{\rm rad}} > 2$). The large $\sigma_{V_{\rm rad}}$ for six of them may indicate the presence of binaries or non-members. Cross-matching with external surveys dedicated to radial velocity estimation, such as APOGEE (Majewski et al. 2017), does not add information (only one star was found in common between the two catalogues). The little information on radial velocities makes it difficult to characterise OC members free of contamination from field stars.

The photometric information is included when deciding if a CMD matches a real OC or not. This is done using an ANN trained with CMDs from the 1229 OCs in Cantat-Gaudin et al. (2018) (see Sect. 2), so the expected isochrone patterns are similar to those present in the training set. The ages of the reference clusters used in the training span from 40 Myr to 1.5 Gyr, so objects accepted by the ANN are in that age range. No estimation of photometric derived quantities is done here, only to mention that 25 of the 53 reported OCs have stars evolved beyond the main sequence, representing the oldest population of the found clusters. In Fig. 7 a few examples of detected OCs are shown, four class A and one class B, showing different ages. Together with the distribution in the five astrometric parameters $(\alpha, \delta, \varpi, \mu_{\alpha^*}, \mu_\delta)$, the rightmost plots show the CMDs for each example OC.

4.4. Matches with other catalogues

The candidates have been cross-matched with known catalogues of OCs (Dias et al. 2002; Kharchenko et al. 2013). These catalogues contain around 2000 and 3000 known stellar structures, respectively. However, some of these structures have recently been found not to be real OCs (Han et al. 2016; Kos et al. 2018; Cantat-Gaudin et al. 2018; Angelo et al. 2019). Moreover, the catalogues were both compiled from heterogeneous data

[4] For an internal check, we studied the areas centred on the new OCs with UPMASK (Krone-Martins & Moitinho 2014) and we confirmed our findings in 96% of the cases.

[5] Available online at VizieR service.

[6] https://www.cosmos.esa.int/web/gaia/iow_20180614

Table 1. Proposed OCs ordered by increasing l.

Name	α (deg)	δ (deg)	l (deg)	b (deg)	θ (deg)	ϖ (mas)	d (kpc)	μ_{α^*} (mas yr^{-1})	μ_δ (mas yr^{-1})	$V_{\rm rad}$ (km s^{-1})	N ($N_{V_{\rm rad}}$)
Class A											
UBC 33	7.39(0.18)	60.49(0.08)	120.24(0.09)	−2.27(0.08)	0.12	0.63(0.03)	$1.6^{+0.08}_{-0.07}$	−0.94(0.09)	−0.46(0.06)	−(−)	43(0)
UBC 34 [a]	11.8(0.22)	66.75(0.15)	122.51(0.09)	3.89(0.15)	0.17	1.55(0.04)	$0.64^{+0.02}_{-0.02}$	−5.02(0.31)	−3.1(0.36)	−12.02(−)	41(1)
UBC 35 [a]	15.1(0.11)	55.41(0.09)	124.21(0.07)	−7.44(0.09)	0.11	0.79(0.05)	$1.27^{+0.08}_{-0.07}$	−4.46(0.2)	−1.94(0.15)	−31.58(0.68)	70(3)
UBC 36	16.47(0.06)	59.64(0.06)	124.76(0.03)	−3.18(0.06)	0.07	0.47(0.04)	$2.15^{+0.19}_{-0.16}$	−1.21(0.19)	−0.46(0.14)	−50.68(5.09)	27(2)
UBC 37 [a]	20.95(0.38)	70.58(0.12)	125.64(0.13)	7.88(0.12)	0.17	1.33(0.06)	$0.75^{+0.04}_{-0.03}$	−6.13(0.36)	2.08(0.27)	−25.02(1.52)	82(2)
UBC 38 [a]	18.73(0.11)	60.5(0.07)	125.82(0.06)	−2.24(0.06)	0.09	0.79(0.04)	$1.27^{+0.06}_{-0.06}$	−2.45(0.13)	−1.81(0.12)	87.09(−)	56(1)
UBC 39	19.79(0.12)	61.02(0.07)	126.29(0.06)	−1.67(0.07)	0.09	0.48(0.03)	$2.09^{+0.16}_{-0.14}$	−1.23(0.08)	−0.13(0.12)	−(−)	45(0)
UBC 40	22.63(0.06)	60.24(0.04)	127.77(0.03)	−2.27(0.04)	0.05	0.4(0.03)	$2.48^{+0.19}_{-0.16}$	−1.01(0.24)	−0.56(0.13)	−(−)	27(0)
UBC 41	23.23(0.09)	59.79(0.05)	128.13(0.04)	−2.66(0.05)	0.06	0.38(0.04)	$2.62^{+0.28}_{-0.23}$	−0.76(0.29)	−0.73(0.22)	−42.02(−)	47(1)
UBC 42 [a]	26.14(0.11)	58.74(0.05)	129.79(0.06)	−3.42(0.05)	0.07	0.45(0.03)	$2.23^{+0.16}_{-0.14}$	−0.93(0.15)	−1.01(0.13)	−(−)	55(0)
UBC 43 [a]	28.1(0.09)	58.65(0.06)	130.8(0.05)	−3.29(0.06)	0.08	0.28(0.04)	$3.54^{+0.56}_{-0.43}$	−2.37(0.13)	−0.44(0.12)	−43.7(2.34)	73(2)
UBC 44	31.11(0.1)	54.36(0.06)	133.53(0.06)	−7.01(0.06)	0.08	0.35(0.04)	$2.84^{+0.32}_{-0.26}$	−2.2(0.24)	−0.23(0.23)	−38.03(0.98)	47(5)
UBC 45 [a]	33.75(0.1)	58.45(0.04)	133.7(0.05)	−2.67(0.04)	0.07	0.63(0.04)	$1.59^{+0.1}_{-0.09}$	−1.02(0.16)	−1.53(0.15)	−(−)	31(0)
UBC 46	33.69(0.15)	57.31(0.11)	134.03(0.08)	−3.76(0.11)	0.14	0.4(0.03)	$2.52^{+0.21}_{-0.18}$	−0.82(0.21)	−1.14(0.22)	−(−)	65(0)
UBC 47	42.0(0.09)	63.8(0.06)	135.37(0.05)	3.78(0.05)	0.07	0.65(0.04)	$1.54^{+0.09}_{-0.08}$	1.19(0.25)	−1.12(0.17)	−10.43(−)	24(1)
UBC 48 [a]	39.07(0.19)	50.05(0.16)	139.64(0.14)	−9.39(0.15)	0.2	1.36(0.05)	$0.73^{+0.03}_{-0.03}$	2.5(0.31)	−2.5(0.26)	−14.04(9.46)	49(3)
UBC 49	60.22(0.12)	59.19(0.06)	145.14(0.07)	4.75(0.04)	0.09	0.34(0.05)	$2.97^{+0.56}_{-0.41}$	−1.77(0.13)	−1.33(0.14)	−14.29(−)	47(1)
UBC 50 [a]	51.5(0.13)	51.08(0.1)	146.11(0.08)	−4.7(0.1)	0.13	0.8(0.04)	$1.25^{+0.07}_{-0.06}$	2.03(0.18)	−6.78(0.21)	−8.78(0.3)	52(2)
UBC 51	59.67(0.17)	52.56(0.09)	149.24(0.09)	−0.47(0.1)	0.14	0.88(0.03)	$1.14^{+0.04}_{-0.04}$	−0.23(0.23)	−1.37(0.28)	−0.22(−)	34(1)
UBC 52	64.74(0.13)	52.37(0.11)	151.65(0.11)	1.47(0.07)	0.13	0.41(0.04)	$2.43^{+0.26}_{-0.22}$	−0.86(0.12)	0.58(0.1)	−27.83(7.35)	32(2)
UBC 53	59.82(0.09)	47.4(0.06)	152.68(0.06)	−4.33(0.06)	0.08	0.6(0.04)	$1.67^{+0.13}_{-0.11}$	0.67(0.12)	−2.92(0.15)	−18.13(7.71)	47(3)
UBC 54	64.72(0.19)	46.44(0.15)	155.8(0.14)	−2.77(0.14)	0.2	0.88(0.05)	$1.14^{+0.08}_{-0.07}$	3.33(0.23)	−3.79(0.3)	−15.46(0.46)	143(2)
UBC 56	69.88(0.14)	47.53(0.12)	157.43(0.12)	0.53(0.09)	0.15	1.11(0.04)	$0.9^{+0.03}_{-0.03}$	1.62(0.28)	−4.01(0.25)	−(−)	72(0)
UBC 57	62.96(0.1)	42.72(0.05)	157.48(0.06)	−6.32(0.06)	0.09	0.48(0.05)	$2.08^{+0.23}_{-0.19}$	3.19(0.22)	−2.24(0.19)	5.24(0.23)	36(3)
UBC 58 [a]	68.41(0.13)	40.5(0.1)	161.94(0.11)	−4.98(0.09)	0.14	0.95(0.06)	$1.05^{+0.07}_{-0.06}$	2.03(0.41)	−3.41(0.46)	1.0(−)	39(1)
UBC 59	82.24(0.12)	48.04(0.09)	162.06(0.1)	7.44(0.08)	0.12	0.38(0.04)	$2.62^{+0.35}_{-0.27}$	0.69(0.24)	−2.0(0.26)	−29.73(9.06)	76(5)
UBC 60 [a]	68.13(0.2)	39.5(0.13)	162.54(0.16)	−5.81(0.13)	0.2	1.47(0.05)	$0.68^{+0.02}_{-0.02}$	3.62(0.43)	−5.73(0.36)	−9.52(13.66)	71(8)
UBC 61	75.06(0.15)	36.27(0.15)	168.55(0.15)	−3.72(0.12)	0.19	0.75(0.05)	$1.33^{+0.1}_{-0.09}$	2.1(0.14)	−2.17(0.12)	10.1(0.93)	52(2)
UBC 62 [a]	76.11(0.12)	35.82(0.08)	169.42(0.09)	−3.32(0.09)	0.13	0.83(0.05)	$1.21^{+0.08}_{-0.07}$	0.36(0.18)	−3.75(0.19)	−(−)	94(0)
UBC 63	79.67(0.09)	37.82(0.08)	169.49(0.08)	0.16(0.07)	0.11	0.65(0.04)	$1.54^{+0.1}_{-0.09}$	1.12(0.18)	−3.56(0.17)	−(−)	26(0)
UBC 65 [a]	82.18(0.15)	34.32(0.14)	173.53(0.15)	−0.15(0.1)	0.18	0.78(0.06)	$1.28^{+0.1}_{-0.09}$	−1.48(0.14)	−4.67(0.2)	−(−)	79(0)
UBC 66 [a]	78.58(0.1)	31.72(0.09)	173.95(0.09)	−4.1(0.09)	0.13	0.91(0.04)	$1.09^{+0.05}_{-0.04}$	0.52(0.21)	−1.48(0.23)	−(−)	27(0)
UBC 67 [a]	81.87(0.06)	33.53(0.06)	174.05(0.05)	−0.79(0.06)	0.08	0.47(0.03)	$2.14^{+0.15}_{-0.13}$	0.45(0.13)	−2.71(0.13)	−(−)	38(0)
UBC 68	91.17(0.1)	36.77(0.07)	175.21(0.07)	7.39(0.08)	0.1	0.43(0.05)	$2.32^{+0.32}_{-0.25}$	−0.5(0.22)	−1.69(0.23)	−(−)	54(0)
UBC 69 [a]	84.77(0.08)	28.4(0.1)	179.7(0.1)	−1.5(0.06)	0.12	0.71(0.04)	$1.42^{+0.09}_{-0.08}$	−0.13(0.18)	−3.82(0.22)	−(−)	44(0)
UBC 70 [a]	91.06(0.07)	31.61(0.06)	179.72(0.06)	4.81(0.06)	0.09	0.48(0.05)	$2.07^{+0.23}_{-0.19}$	−0.72(0.16)	−3.27(0.13)	14.57(0.79)	60(2)
UBC 72	90.99(0.09)	26.65(0.08)	184.02(0.09)	2.34(0.08)	0.12	0.52(0.04)	$1.93^{+0.16}_{-0.14}$	0.36(0.13)	−0.01(0.15)	30.35(0.63)	77(3)
UBC 74	95.47(0.07)	22.41(0.06)	189.7(0.06)	3.9(0.06)	0.09	0.35(0.05)	$2.82^{+0.45}_{-0.34}$	1.09(0.11)	−2.62(0.13)	43.98(1.53)	65(3)
UBC 75 [a]	83.77(0.07)	15.71(0.09)	190.02(0.09)	−9.02(0.07)	0.11	0.67(0.05)	$1.5^{+0.11}_{-0.1}$	0.26(0.17)	−2.4(0.2)	5.95(−)	57(1)
UBC 76	89.0(0.11)	17.34(0.08)	191.2(0.08)	−3.87(0.1)	0.13	0.57(0.02)	$1.75^{+0.07}_{-0.06}$	0.14(0.14)	−1.11(0.09)	−(−)	24(0)
UBC 78 [a]	85.75(0.13)	13.72(0.1)	192.74(0.12)	−8.41(0.1)	0.16	0.91(0.05)	$1.1^{+0.06}_{-0.05}$	0.64(0.35)	−3.65(0.33)	27.84(20.2)	62(2)
UBC 80	91.64(0.09)	8.75(0.1)	199.97(0.1)	−5.84(0.09)	0.14	0.45(0.04)	$2.22^{+0.24}_{-0.2}$	−0.48(0.09)	−1.01(0.21)	−(−)	30(0)
UBC 81 [a]	96.35(0.07)	11.15(0.06)	200.05(0.06)	−0.62(0.06)	0.09	0.58(0.03)	$1.71^{+0.09}_{-0.08}$	−1.11(0.15)	−0.94(0.14)	−(−)	49(0)
UBC 82	95.89(0.08)	8.38(0.1)	202.3(0.1)	−2.31(0.09)	0.13	0.42(0.04)	$2.38^{+0.28}_{-0.23}$	1.31(0.09)	−2.3(0.17)	12.69(0.58)	36(3)
UBC 83	97.56(0.1)	7.36(0.12)	203.96(0.11)	−1.33(0.12)	0.16	0.48(0.06)	$2.11^{+0.29}_{-0.23}$	−1.08(0.14)	0.39(0.05)	−(−)	51(0)
Class B											
UBC 84	15.42(0.11)	61.73(0.09)	124.14(0.05)	−1.11(0.09)	0.1	0.37(0.03)	$2.73^{+0.27}_{-0.23}$	−1.55(0.26)	−0.97(0.18)	−(−)	55(0)
UBC 85	18.68(0.18)	57.86(0.11)	126.04(0.09)	−4.87(0.11)	0.15	0.36(0.03)	$2.78^{+0.29}_{-0.24}$	−3.69(0.15)	−0.54(0.33)	−(−)	33(0)
UBC 86	33.03(0.16)	57.61(0.07)	133.6(0.1)	−3.58(0.06)	0.11	0.34(0.04)	$2.93^{+0.4}_{-0.31}$	−0.85(0.16)	−0.95(0.15)	−39.91(17.43)	71(4)
UBC 87	60.51(0.1)	56.42(0.07)	147.09(0.07)	2.77(0.06)	0.09	0.38(0.03)	$2.62^{+0.25}_{-0.21}$	0.77(0.15)	−1.31(0.16)	−(−)	36(0)
UBC 88	58.18(0.18)	45.94(0.15)	152.76(0.14)	−6.17(0.14)	0.2	1.0(0.06)	$1.0^{+0.06}_{-0.05}$	−1.36(0.33)	−2.95(0.27)	−(−)	88(0)
UBC 89 [a]	81.22(0.14)	37.57(0.08)	170.4(0.09)	1.03(0.1)	0.14	0.88(0.06)	$1.13^{+0.08}_{-0.07}$	0.39(0.23)	−4.27(0.22)	59.54(−)	64(1)
UBC 90	97.21(0.04)	14.92(0.05)	197.11(0.05)	1.87(0.04)	0.06	0.34(0.05)	$2.96^{+0.5}_{-0.37}$	1.23(0.14)	−1.38(0.16)	49.63(−)	53(1)

Notes. The parameters shown are the mean and standard deviation for the (N) members found, the computed apparent size (θ) and estimated distance (d) with one-sigma confidence interval; radial velocity is included when available and is computed with $N_{V_{\rm rad}}$ members. The name follows the numeration started in CG18. [a]Coincidence with COIN clusters.

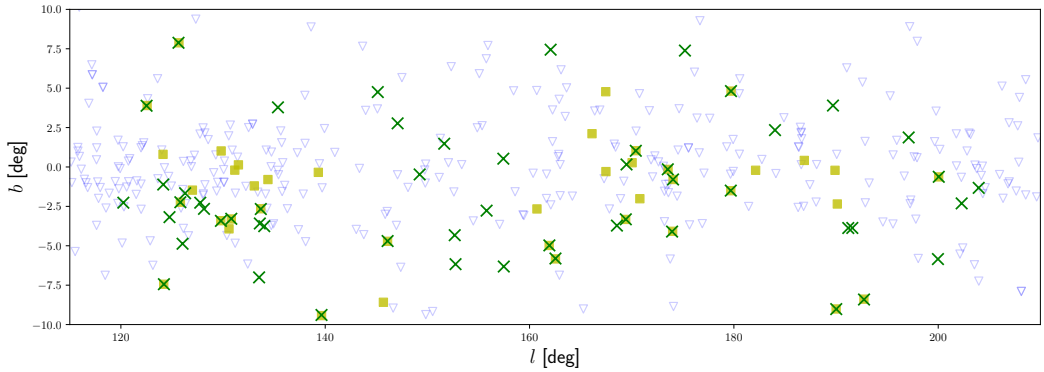


Fig. 4. Spatial distribution (l, b) of the detected (green crosses) OCs, together with the already catalogued ones (blue triangles) in Cantat-Gaudin et al. (2018) and the COIN-*Gaia* clusters (yellow boxes) (Cantat-Gaudin et al. 2019).

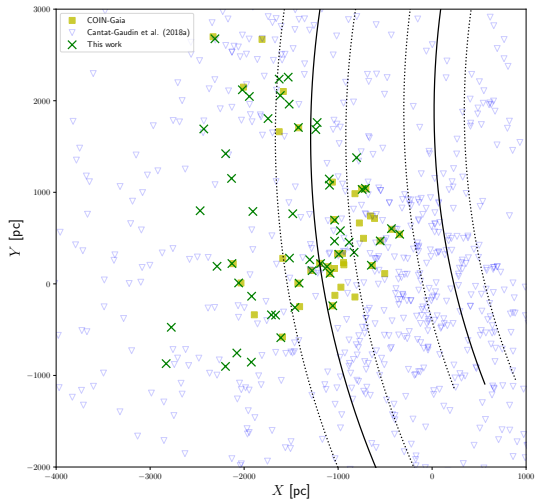


Fig. 5. X-Y projection of the detected OCs (green crosses) together with already catalogued OCs (blue triangles) and COIN-*Gaia* clusters (yellow boxes). Black lines represent the Local and the Perseus arms, plotted following the model in Reid et al. (2014). The Sun is at $(0, 0)$.

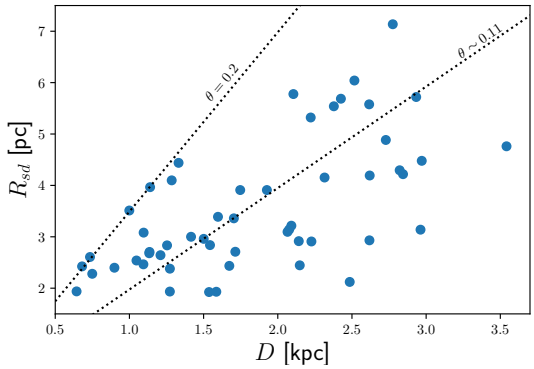


Fig. 6. Radius (computed from the standard deviation in l and b as $\sqrt{\sigma_l^2 + \sigma_b^2}$) as a function of distance for each of the reported 53 OCs. Dotted lines represent the limiting cluster apparent size and the mean apparent size, $\theta = 0.2°$ and $\theta \sim 0.11°$, respectively.

sources, making the identification of an OC less reliable beyond positional arguments. We consider an OC to be positionally matched to a catalogued one if their centres lie within a circle of radius $r = 0.5°$. We find 17 candidates whose centres are identified with one object either from Dias et al. (2002) or Kharchenko et al. (2013); however, none of the identifications are compatible in the rest of the astrometric mean parameters $(\varpi, \mu_{\alpha^*}, \mu_\delta)$, with the closest pair differing by $\sim 8\sigma$ in at least one parameter. However, we find UBC 84 near the association Cas OB1, to which it may be related due to the extended region of association.

A recent list of 10 978 star clusters, associations, and candidates in the Milky Way has been published by Bica et al. (2019). Our list of candidates was cross-matched and only UBC 90 and UBC 44 are near one entry in the catalogue, Teutsch 20 and Patchick 12, respectively. Teutsch 20 and Patchick 12 are not listed in any of the other studied catalogues. Moreover, the quoted distance for Teutsch 20 is 2.54 ± 0.05 kpc (Guo et al. 2018) and we find UBC 90 at 2.94 kpc, which is not compatible within errors. For the case of Patchick 12, we found no record of its mean astrometric parameters in the literature.

As said before, Cantat-Gaudin et al. (2019) recently found 41 OCs located in roughly the same area of the sky, exploring the data of *Gaia* DR2. We find 21 OCs in common that share the five astrometric parameters (see Table 1). The other 20 OCs were not detected in our blind search, but we were able to recover them by increasing the OC field contrast when running DBSCAN in a cone search centred on the targeted OC. This shows that ML methods are complementary to each other, with none of the explored methods being able to detect 100% of the existing structures.

5. Conclusions

We use the methodology described in CG18 to systematically explore the *Gaia* DR2 archive to search for unknown OCs in the anticentre direction. The method is a fully automated data mining task that uses an unsupervised clustering algorithm, DBSCAN, to find groups of stars that share common $(l, b, \varpi, \mu_{\alpha^*}, \mu_\delta)$ and decide whether or not they are real OCs based on an isochrone pattern recognition in the CMD using an ANN.

We can assess the overall performance in terms of the detection of already existing OCs. In this case, the method is able to find more than 75% of the confirmed OCs in the search region. Most of the remaining ~24% of the clusters not found are

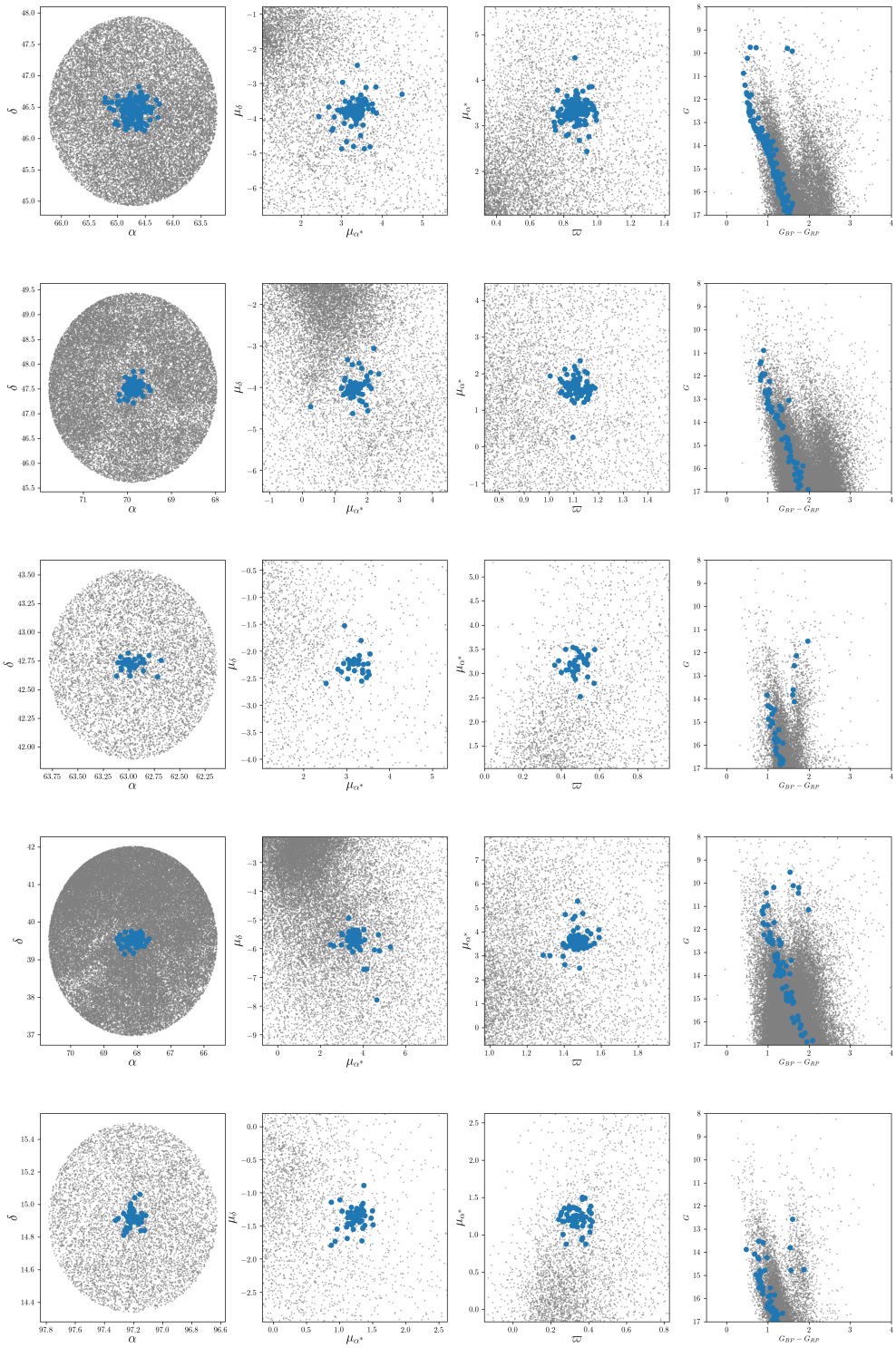


Fig. 7. Five examples of the 53 detected OCs. The blue dots represent the detected members, while grey dots represent field stars. *Leftmost plots*: position of the OC in (α, δ). *Inner left plots*: $(\mu_{\alpha^*}, \mu_\delta)$ distribution, whilst *inner right plots*: (ϖ, μ_{α^*}) distribution. *Rightmost plots*: CMD of each OC. The plotted OCs are, *from top to bottom*: UBC 54, UBC 56, UBC 57, UBC 60, and UBC 90. The first four clusters are class A and the last one is a class B cluster (see Table 1).

recovered when the search is focused on the targeted OC. This suggests that our method works better for the OCs whose OC field contrast is high, and may be biased towards the more compact objects when the distance decreases.

The application of the whole methodology leads to the report of 53 new OCs in a region covering the Galactic anticentre and the Perseus arm in the *Gaia* DR2 data ($120° \leq l \leq 205°$ and $-10° \leq b \leq 10°$), which represents an increase of more than 22% with respect to the OCs catalogued in this area. Moreover, 28 of the detected OCs are closer than 2 kpc, suggesting that there may be more groups to be detected in this volume.

The density of OCs decreases in a region near $l \sim 140°$. Very few OCs are found in this region, including already catalogued OCs and the newly reported OCs. This region has been named the Gulf of Camelopardalis, and it reveals a complex structure of the second Galactic quadrant whose mapping was only recently made possible by *Gaia* DR2 data, and still deserves further study.

The application of our methodology in the search regions shows that the census of OCs may not be complete. Moreover, other similar methodologies exploring the same region are able to find more groups not detected via our method, while they missed some groups detected here. We conclude that a blind search using a single detection method is not able not recover all the existing stellar structures, and that different ML algorithms for this purpose are complementary to each other.

The design of the whole methodology, requiring minimal manual intervention, means that its application to a big data set such as the whole *Gaia* DR2 is possible. The planned future exploitation of the *Gaia* archive in terms of blind search of OCs would represent a huge increase to the known OC population.

Acknowledgements. This work has made use of results from the European Space Agency (ESA) space mission *Gaia*, the data from which were processed by the *Gaia* Data Processing and Analysis Consortium (DPAC). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement. The *Gaia* mission website is http://www.cosmos.esa.int/gaia. The authors are current or past members of the ESA *Gaia* mission team and of the *Gaia* DPAC. This work was supported by the MINECO (Spanish Ministry of Economy) through grant ESP2016-80079-C2-1-R (MINECO/FEDER, UE) and MDM-2014-0369 of ICCUB (Unidad de Excelencia "María de Maeztu"). This research has made use of the TOPCAT (Taylor 2005). This research has made use of the VizieR catalogue access tool, CDS, Strasbourg, France. The original description of the VizieR service was published in A&AS 143, 23. ACG thanks Dr. Laia Casamiquela for her useful comments.

References

Angelo, M. S., Santos, J. F. C., Corradi, W. J. B., & Maia, F. F. S. 2019, A&A, 624, A8
Bica, E., Pavani, D. B., Bonatto, C. J., & Lima, E. F. 2019, AJ, 157, 12
Bossini, D., Vallenari, A., Bragaglia, A., et al. 2019, A&A, 623, A108
Cantat-Gaudin, T., Jordi, C., Vallenari, A., et al. 2018, A&A, 618, A93
Cantat-Gaudin, T., Krone-Martins, A., Sedaghat, N., et al. 2019, A&A, 624, A126
Castro-Ginard, A., Jordi, C., Luri, X., et al. 2018, A&A, 618, A59
Dias, W. S., Alessi, B. S., Moitinho, A., & Lépine, J. R. D. 2002, A&A, 389, 871
Ester, M., Kriegel, H. P., Sander, J., & Xu, X. 1996, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96 (AAAI Press), 226
Evans, D. W., Riello, M., De Angeli, F., et al. 2018, A&A, 616, A4
Froebrich, D., Scholz, A., & Raftery, C. L. 2007, MNRAS, 374, 399
Gaia Collaboration (Prusti, T., et al.) 2016, A&A, 595, A1
Gaia Collaboration (Brown, A. G. A., et al.) 2018, A&A, 616, A1
Gao, X. 2018a, ApJ, 869, 9
Gao, X.-H. 2018b, Ap&SS, 363, 232
Guo, J.-C., Zhang, H.-W., Zhang, H.-H., et al. 2018, Res. Astron. Astrophys., 18, 032
Han, E., Curtis, J. L., & Wright, J. T. 2016, AJ, 152, 7
Hinton, G. 1989, Artif. Intell., 40, 185
Kharchenko, N. V., Piskunov, A. E., Schilbach, E., Röser, S., & Scholz, R.-D. 2013, A&A, 558, A53
Kos, J., de Silva, G., Buder, S., et al. 2018, MNRAS, 480, 5242
Krone-Martins, A., & Moitinho, A. 2014, A&A, 561, A57
Lindegren, L., Lammers, U., Bastian, U., et al. 2016, A&A, 595, A4
Lindegren, L., Hernández, J., Bombrun, A., et al. 2018, A&A, 616, A2
Luri, X., Brown, A. G. A., Sarro, L. M., et al. 2018, A&A, 616, A9
Majewski, S. R., Schiavon, R. P., Frinchaboy, P. M., et al. 2017, AJ, 154, 94
Matsunaga, N., Bono, G., Chen, X., et al. 2018, Space Sci. Rev., 214, 74
Michalik, D., Lindegren, L., & Hobbs, D. 2015, A&A, 574, A115
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, J. Mach. Learn. Res., 12, 2825
Piskunov, A. E., Just, A., Kharchenko, N. V., et al. 2018, A&A, 614, A22
Reid, M. J., Menten, K. M., Brunthaler, A., et al. 2014, ApJ, 783, 130
Röser, S., Schilbach, E., & Goldman, B. 2016, A&A, 595, A22
Schmeja, S., Kharchenko, N. V., Piskunov, A. E., et al. 2014, A&A, 568, A51
Scholz, R. D., Kharchenko, N. V., Piskunov, A. E., Röser, S., & Schilbach, E. 2015, A&A, 581, A39
Soubiran, C., Cantat-Gaudin, T., Romero-Gómez, M., et al. 2018, A&A, 619, A155
Taylor, M. B. 2005, in Astronomical Data Analysis Software and Systems XIV, eds. P. Shopbell, M. Britton, & R. Ebert, ASP Conf. Ser., 347, 29

4

A BIG DATA SEARCH IN THE GALACTIC DISC

This Chapter contains the published version of Castro-Ginard et al. (2020, A&A, 635, A45). This paper was marked as a research highlight by Astronomy & Astrophysics on March 2020.

We applied the methodology described in Chapter 2 and adapted to *Gaia* data in Chapter 3 to search for unknown OCs in the whole Galactic disc (defined as $|b| < 20°$) up to magnitude $G = 17$. The analysis of this region of the *Gaia* data, which counted with 122 727 809 stars with complete five-dimensional astrometric data and photometry for the three *Gaia* bands, represented a huge challenge due to the amount of data. Therefore, the analysis was carried out in a high-performance computing environment[1] using Big Data techniques.

For the clustering part of the algorithm, the `DBSCAN` code was parallelized to deal with very crowded regions. The parallelisation was done using an approach based on graphs, which takes into account the data dependencies when building the graph to later distribute and perform the computation (Tejedor et al. 2017; Álvarez Cid-Fuentes et al. 2019).

Once the statistical overdensities are found, the recognition of real physical OCs among them is done using an upgraded ANN. This new Deep Learning ANN uses several convolutional layers to automatically characterise the meaningful features that will define an isochrone in a CMD, and this allowed us to build a more robust recognition of OCs.

The results presented in this work, 582 new OCs in the Galactic disc, represent a huge increase on the knowledge of the OC population which counted with 1229 OCs known previous to (and confirmed by) *Gaia* (Cantat-Gaudin et al. 2018).

1 MareNostrum at the Barcelona Supercomputer Center. This HPC cluster is ranked as world TOP500. https://www.bsc.es/marenostrum

A&A 635, A45 (2020)
https://doi.org/10.1051/0004-6361/201937386
© ESO 2020

Astronomy & Astrophysics

Hunting for open clusters in *Gaia* DR2: 582 new open clusters in the Galactic disc⋆

A. Castro-Ginard[1], C. Jordi[1], X. Luri[1], J. Álvarez Cid-Fuentes[2], L. Casamiquela[3], F. Anders[1], T. Cantat-Gaudin[1], M. Monguió[1], L. Balaguer-Núñez[1], S. Solà[2], and R. M. Badia[2]

[1] Dept. Física Quàntica i Astrofísica, Institut de Ciències del Cosmos (ICCUB), Universitat de Barcelona (IEEC-UB), Martí i Franquès 1, 08028 Barcelona, Spain
e-mail: acastro@fqa.ub.edu
[2] Barcelona Supercomputing Center (BSC), Barcelona, Spain
[3] Laboratoire d'Astrophysique de Bordeaux, Univ. Bordeaux, CNRS, B18N, allée Geoffroy, Saint-Hilaire 33615, Pessac, France

Received 20 December 2019 / Accepted 18 January 2020

ABSTRACT

Context. Open clusters are key targets for studies of Galaxy structure and evolution, and stellar physics. Since the *Gaia* data release 2 (DR2), the discovery of undetected clusters has shown that previous surveys were incomplete.
Aims. Our aim is to exploit the Big Data capabilities of machine learning to detect new open clusters in *Gaia* DR2, and to complete the open cluster sample to enable further studies of the Galactic disc.
Methods. We use a machine-learning based methodology to systematically search the Galactic disc for overdensities in the astrometric space and identify the open clusters using photometric information. First, we used an unsupervised clustering algorithm, DBSCAN, to blindly search for these overdensities in *Gaia* DR2 $(l, b, \varpi, \mu_{\alpha^*}, \mu_\delta)$, and then we used a deep learning artificial neural network trained on colour–magnitude diagrams to identify isochrone patterns in these overdensities, and to confirm them as open clusters.
Results. We find 582 new open clusters distributed along the Galactic disc in the region $|b| < 20°$. We detect substructure in complex regions, and identify the tidal tails of a disrupting cluster UBC 274 of ∼3 Gyr located at ∼2 kpc.
Conclusions. Adapting the mentioned methodology to a Big Data environment allows us to target the search using the physical properties of open clusters instead of being driven by computational limitations. This blind search for open clusters in the Galactic disc increases the number of known open clusters by 45%.

Key words. surveys – open clusters and associations: general – astrometry – methods: data analysis

1. Introduction

Since the publication of the second data release of the ESA mission *Gaia* (*Gaia* DR2; Gaia Collaboration 2016, 2018), which contains more than 1.3 billion stars with precise astrometric measurements (positions, parallax, and proper motions) and integrated photometry for three broad bands (G, G_{BP}, and G_{RP}), among other data products, the study of open clusters (OCs) has been revolutionised and the OC population redefined in statistical terms.

Open clusters are fundamental objects in galaxies that allow us to understand the structure and evolution of the Milky Way. They are groups of stars that are gravitationally bound and born in the same event and therefore stars in an OC share a common position and proper motion $(l, b, \varpi, \mu_{\alpha^*}, \mu_\delta)$ as well as initial chemical composition and age. The possibility to reliably estimate the ages and distances of OCs, compared to the estimation on individual stars, makes them a useful tool for studying several topics in astrophysics. Young OCs allow us to the derive the initial mass function (IMF) and trace star forming regions, providing useful information on star forming mechanisms. Intermediate to old OCs contain information about the processes occurring in the Galactic disc that disrupt these stellar structures and drive the evolution of the disc. All OCs are also indispensable to constrain stellar structure and evolutionary models. To enable most of these studies, a complete and homogeneous census of the OC population needs to be built.

Many published studies were aimed at detecting new OCs and accurately determining membership probability. Shortly after the publication of *Gaia* DR2, Cantat-Gaudin et al. (2018) was able to compute membership probabilities for 1229 OCs present in catalogues previous to *Gaia* DR2 (where these catalogues included about 3000 objects Dias et al. 2002 and Kharchenko et al. 2013), and proved the non-existence of some of them. In parallel, Castro-Ginard et al. (2018) developed a machine learning (ML) methodology to search for unnoticed OCs in the *Gaia* data and was able to detect 23 new OCs distributed throughout the sky in the TGAS data set (Michalik et al. 2015; Lindegren et al. 2016) and 53 new OCs in a region near the Galactic anticentre (Castro-Ginard et al. 2019). Since then, there have been many efforts to complete the OC census: Cantat-Gaudin et al. (2019a) found 41 OCs in the direction of Perseus using Gaussian mixture models; Sim et al. (2019) found 207 OCs by visually inspecting proper motion diagrams; and Liu & Pang (2019) recently reported 2 443 OCs, of which 76 were unknown and considered of high quality, by dividing the sky into small 3D regions and employing a friends-of-friends algorithm to search for overdensities in the $(l, b, \varpi, \mu_{\alpha^*}, \mu_\delta)$ space.

⋆ Full Table 1 and Table 2 are only available at the CDS via anonymous ftp to cdsarc.u-strasbg.fr (130.79.128.5) or via http://cdsarc.u-strasbg.fr/viz-bin/cat/J/A+A/635/A45

All of these previous studies analysed either a particular region of the Galactic disc, or divided the entire Galactic disc into areas defined by the limiting number of stars that the algorithms are able to deal with due to the computational complexity and resources needed when dealing with Big Data catalogues such as *Gaia*. The implementation of such methodologies in a Big Data environment, where the division of the search region of the sky into small regions depends only on the targeted structures and not on any computational limitation, is a key step in blind all-sky searches.

In this paper, we adapt the methodology described in Castro-Ginard et al. (2018; 2019, CG18 and CG19 hereafter) to run in a Big Data environment. The methodology consists in the application of an unsupervised clustering algorithm, DBSCAN, to find overdensities in a five-dimensional parameter space $(l, b, \varpi, \mu_{\alpha^*}, \mu_\delta)$. The confirmation of these overdensities as plausible clusters is done by recognising an isochrone pattern in the colour–magnitude diagram (CMD) of the candidates using a deep learning artificial neural network (ANN).

This paper is organised as follows. In Sect. 2 we discuss the methodology used, and how we adapted it to a Big Data environment. Section 3 describes the data used. A review of the new OCs found is presented in Sect. 4, as well as some general properties of the new OCs and a comparison with other OC catalogues. This section also includes some specific comments on the capabilities of the methodology. Finally, conclusions are presented in Sect. 5.

2. Methodology

This section summarises the methodology used to systematically search for unknown OCs. The method is fully described in CG18, and was applied to *Gaia* DR2 data in CG19 to find new OCs in a region near the Galactic anticentre.

The method consists in three main steps: preparing the data, identifying clusters with DBSCAN, and confirming them with an ANN.

In the first part, where the data are prepared, the region to be searched is divided into rectangles of size $L \times L$ where the five parameters $(l, b, \varpi, \mu_{\alpha^*}, \mu_\delta)$ used to look for the overdensities are standardised. This division into small regions is necessary to compute an average density of the region, where the clusters located in that region represent local overdensities. Contrary to other papers, the size of these regions is defined by its homogeneity and not by the limitations of the hardware or algorithm.

Once the data are prepared, the overdensities are found using a density-based clustering algorithm, DBSCAN (Ester et al. 1996), which uses a statistical distance (computed as the Euclidean distance in our case) to define close-by stars in 5D as a cluster. This step has been improved with respect to CG18 and CG19 because of the larger volume of data to be analysed (see Sect. 2.1 for details). The choice of DBSCAN is convenient because it does not require an *a priori* number of clusters to be found, it is able to find arbitrarily shaped clusters, and it only requires two input parameters $(\epsilon, minPts)$. The ϵ parameter is the radius of the hypersphere in which to search for close neighbours (members of the same cluster); it is automatically computed in each $L \times L$ rectangle using the fact that the separation between stars in a cluster is smaller than between field stars (see Sect. 2.2 in CG18 for details on the computation of ϵ). The parameter *minPts* refers to the minimum number of stars within ϵ to consider them as a cluster. Once DBSCAN finds the statistical clusters in a grid defined by the $L \times L$ rectangles, the grid is shifted by $L/3$ and $2L/3$ where the algorithm is run again to account for clusters in the borders.

The value of *minPts* is optimised, together with L, using *Gaia*-like simulated data. We used a *Gaia* Universe Model Snapshot (GUMS) to simulate field stars (Robin et al. 2012) including errors at the time of *Gaia* DR2[1]. Open clusters simulated using the *Gaia* Object Generator (GOG Luri et al. 2014) were added to the GUMS simulation as the objects to be found by DBSCAN. A pair of $(L, minPts)$ is considered to be optimal if a balance is reached in terms of low contamination and high efficiency.

For true data, the whole process is run over the several $(L, minPts)$ optimal parameters to assess the reliability of the clusters found. The more times a statistical cluster has been found within the explored $(L, minPts)$ pairs, the more likely it is to be a real OC. The values of $(L, minPts)$ used are 35 combinations of $L \in [9', 15']$ and $minPts \in [8, 16]$.

As a last step, overdensities found with DBSCAN are classified into real OCs or just statistical clusters using an ANN (Hinton 1989), trained to recognise the characteristic isochrone pattern of OCs in the CMD. This step has also been improved with respect to CG18 and CG19, resulting in a more robust classification with the use of deep learning (see Sect. 2.2).

2.1. Distributed computation of DBSCAN

So far, the method has been applied to small-volume data sets (i.e. to TGAS in CG18, and to a region in the Galactic anticentre up to a magnitude of $G = 17$ in CG19) for design and validation purposes. Both previous studies used the DBSCAN implementation from scikit-learn (Pedregosa et al. 2011), an easy-to-use API that provides ML algorithms for Python. However, the higher stellar density to be analysed in other regions of the disc, such as towards the Galactic centre for example, requires a ML library able to be deployed in a distributed environment and to handle larger volumes of data.

Here, we used PyCOMPSs (Tejedor et al. 2017) to find overdensities in the whole Galactic disc ($0° \leq l \leq 360°$ and $-20° \leq b \leq 20°$) down to a magnitude of $G = 17$. PyCOMPSs is a task-based programming model that automatically manages the distribution of the computation depending on the available resources. Using PyCOMPSs, we build an application that uses DBSCAN from scikit-learn on different regions of the Galactic disc in parallel. This speeds up the computation time and allows us to process a volume of data that does not fit in the memory of a single machine.

The algorithm is deployed on the MareNostrum 4 supercomputer[2] installed at the Barcelona Supercomputing Center (BSC). The nodes used for the computation of DBSCAN have 96 *GB* of memory and 48 cores per node. For performance-comparison purposes, we ran DBSCAN with the same configuration that we used in CG18 on the TGAS data set. In that case, in CG18, the computation of DBSCAN for all the optimal parameters took 18 hours in a sequential execution on a single machine, whereas when using PyCOMPSs the whole computation takes ∼1.4 h in one node (48 cores) and less than 18 minutes in four nodes (192 cores, see Sect. 5 from Álvarez Cid-Fuentes et al. 2019, for a detailed comparison).

For this case, the analysis of the whole Galactic disc (defined as $-20° \leq b \leq 20°$) up to magnitude $G = 17$ using DBSCAN on four nodes (192 cores) takes an average of 8.27 hours per pair of parameters, ranging from 5.67 to 11.17 hours depending on the pairs of $(L, minPts)$.

[1] Errors computed with the prescription given in https://github.com/agabrown/PyGaia

[2] https://www.bsc.es/marenostrum

2.2. Open cluster validation with deep learning

The application of DBSCAN over a large volume of data with several optimal pairs of parameters ($L, minPts$) picks up a large number of statistical overdensities that correspond to real OCs, also including overdensities only in statistical terms. To automatically decide whether or not a given statistical cluster is a real OC we have trained an ANN to recognise the isochrone patterns that stars in OCs follow in a CMD. For both CG18 and CG19 we used a simple multi-layer perceptron with one hidden layer to make the classification. In this paper, due to the large number of statistical clusters found, a more complex model is needed for robust classification. We designed a "deep" ANN, with several convolutional layers to perform the classification.

This deep ANN is implemented in PyTorch[3] (Paszke et al. 2017), a popular and powerful deep learning library. It takes a 2D histogram in $G_{BP} - G_{RP}$ vs. G, as input, that is, a CMD, and is trained to decide whether it belongs to a real OC or not. The network is built in two blocks; a first block consisting in a set of convolutional layers which are able to learn the features and geometry of the isochrone pattern in the CMD, and a second block with two fully connected layers where the classification of the learned features is performed. After each layer, a ReLU activation function ($f(x) = \max(0, x)$) is added, which has been shown to give better results than other activation functions (LeCun et al. 2012).

2.2.1. Building the training set

One of the caveats of deep learning is that it requires a large amount of training samples to learn the possible configurations of the feature space. The CMDs of the approximately 1500 confirmed OCs are not sufficient to train the network. Moreover, some of these OCs do not have enough stars ($minPts$ at least) with magnitudes of $G \leq 17$ or the isochrone is very dispersed, and therefore we had to remove these clusters from the training set. To enlarge the training set we used data-augmentation techniques (see description in Sect. 2.3.2 in CG18) on the real known OCs. In addition to the known OCs, we used simulated isochrones from the PARSEC code (Bressan et al. 2012). To build the set of isochrones, we assume solar metallicity ($Z \simeq 0.0152$) and ages ranging from $\log(age) = 6.6$ dex to $\log(age) = 10.3$ dex in steps of 0.1 dex. For each age, the isochrone is filled with a population of a total mass of $10^4\,M_\odot$ following the IMF described in Kroupa (2001). We then select different subsamples of the whole population to create the simulated OCs, and we locate them at different distances (ranging from 0.4 to 4 kpc) to better represent the parameter space. For each subsample, the CMD is built in the $G_{BP} - G_{RP}$ versus G space using the photometric pass bands described in Maíz Apellániz & Weiler (2018). Finally, in order to mimic *Gaia* DR2 results, we add photometric errors (Evans et al. 2018) using an analytical prescription provided by Carrasco et al. (private communication) and a fraction of binaries. On the negative identification side, we use CMDs from random (field) stars located at different fields in the whole studied area.

Each CMD is converted to a 2D histogram, and as a preprocessing step, we normalise the data (each pixel of the histogram is limited between 0 and 1) before feeding the whole 2D histogram to the network. To reach better classification performance, a logarithmic normalisation was done in order to highlight the lower density regions so that the network takes into account the contamination from field stars when performing the classification.

[3] https://pytorch.org/

2.2.2. Performance of the classification

The performance of the classification is assessed in two steps. On the one hand, the whole training set is split into training and test with 80% and 20% of the whole set, respectively. This is useful when designing the network architecture because the true classification of each sample is known. The final architecture is chosen to be the one that minimises the test loss.

On the other hand, the model is applied to the anticentre area as in CG19, where we found 53 new OCs from 491 candidates. We do not know the true classification of each of those 491 samples, so the final parameters of the ANN here are tuned to keep 80% (at least) of the OCs confirmed in that region, minimising the manually discarded statistical clusters. When applying the final model to classify all the statistical clusters found in the Galactic disc, we can recover this 80% requirement (in terms of known OCs recovered) showing that the results are equivalent in both sets.

3. Data

The data used to perform the blind search for OCs are those of the *Gaia* DR2 (Gaia Collaboration 2018). In DR2, *Gaia* provides precise astrometry and kinematics ($l, b, \varpi, \mu_{\alpha^*}, \mu_\delta$) in addition to excellent photometry in three broad bands (G, G_{BP}, G_{RP}). The search is focused on the Galactic disc, defined as $0° \leq l \leq 360°$ and $-20° \leq b \leq 20°$, because the expectation of finding OCs in that region is maximum; i.e. 99% of the known OCs catalogued in Cantat-Gaudin et al. (2018) are in $|b| < 20°$, similarly for Dias et al. (2002) and Kharchenko et al. (2013) with 96% and 94% of the total reported objects in $|b| < 20°$, respectively.

The data set is also limited in magnitude up to $G = 17$, where the median astrometric uncertainties are 0.094 mas for the parallax, and 0.158 and 0.137 mas yr^{-1} for μ_{α^*} and μ_δ, respectively (Lindegren et al. 2018). On the photometric side, up to magnitude $G = 17$ the uncertainties are at the level of ∼0.001 mag for G, ∼0.006 mag for G_{BP}, and ∼0.01 mag for G_{RP} (Evans et al. 2018). We consider these uncertainty levels to be adequate limits with which to obtain satisfactory results with our method. This results in a sample containing 122 727 809 stars.

4. Results

The described methodology is applied to the whole Galactic disc. This results in a list of 2 213 possible OC candidates, including the already known OCs and newly discovered ones.

4.1. Comparison with existing catalogues

To report only newly discovered OCs, we cross-match our list of detections with other catalogues to see which groups are already known.

4.1.1. Cantat-Gaudin et al. (2018)

We consider a candidate to be matched with one OC in the Cantat-Gaudin et al. (2018) catalogue if their mean parameters are compatible within $2\sigma_i$, where σ_i is the standard deviation computed from the members of each OC in the 5D astrometric space, $i = \{l, b, \varpi, \mu_{\alpha^*}, \mu_\delta\}$. From our 2 213 OC candidates,

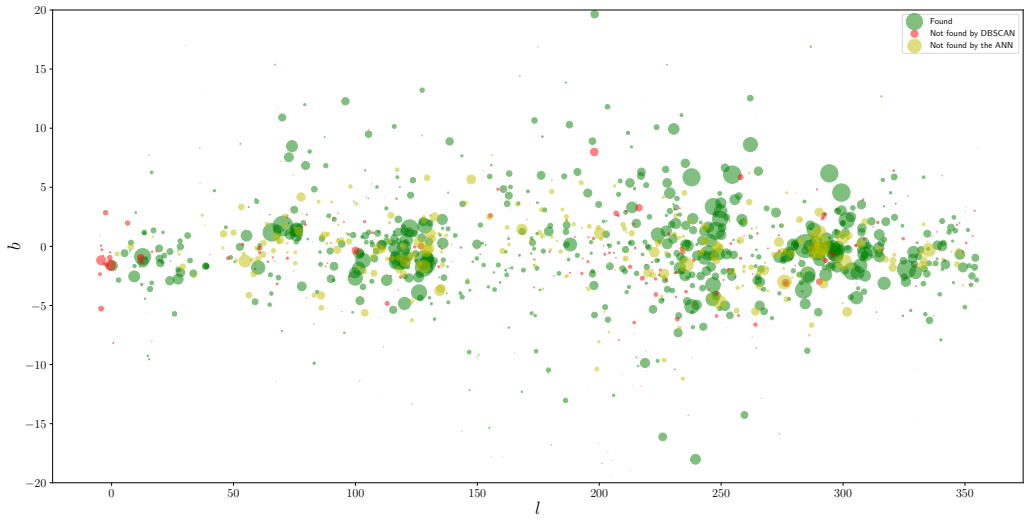


Fig. 1. Distribution in Galactic coordinates (l vs. b) of the OCs catalogued in Cantat-Gaudin et al. (2018). Green dots represent OCs that our method recovers, red dots are OCs not found by DBSCAN, and yellow dots are OCs which are found by DBSCAN but for which the CMD is not recognised by our ANN. The size of the dots is proportional to the star density of the cluster (see text, Eq. (1)).

688 are listed in Cantat-Gaudin et al. (2018) with our matching criteria. This represents ~81% of the OCs reported in Cantat-Gaudin et al. (2018) used in the training set for the ANN, where we removed OCs either with few members up to $G = 17$ or with poorly defined empirical isochrones in the CMDs that would confuse the ANN for the classification.

Our strategy to compute the DBSCAN parameters (L, $minPts$) relies on the higher star density of a cluster compared to field stars. Therefore, our detection is limited to the most compact objects in the field of search ($L \times L$). This is seen in Fig. 1, where a distribution of l versus b of the catalogued OCs is shown. The OCs found using our method are plotted in green, whereas those not found are plotted either in red (if not found by DBSCAN) or in yellow (if its sequence in the CMD is not well defined and is therefore not recognised by our ANN). The size of the dots is proportional to the density of the cluster in the 5D astrometric space, computed as 68% of the total number of stars of the cluster divided by the volume of a 5D hypersphere:

$$V_5 = \frac{\pi^{\frac{5}{2}}}{\Gamma(\frac{5}{2}+1)} r^5, \tag{1}$$

where $r = (\sigma_l^2 + \sigma_b^2 + \sigma_\varpi^2 + \sigma_{\mu_{\alpha^*}}^2 + \sigma_{\mu_\delta}^2)^{\frac{1}{2}}$ for each cluster. The OCs found using our method are mostly high-density groups, whilst those not found are low-density objects (which are near a higher density object) or their sequence in the CMD is not recognised as an isochrone by our ANN.

4.1.2. Castro-Ginard et al. (2018, 2019), and Cantat-Gaudin et al. (2019a)

The method discussed in this paper was presented in CG18, where a blind search was performed over the TGAS data (Lindegren et al. 2016). The 23 OCs found in CG18, mainly closer than 1 kpc (due to the bright limiting magnitude), are not likely to be found with *Gaia* DR2 due to the very different star density of the data set and the parameters (L, $minPts$) used in the search. However, we can find UBC 3, UBC 6, UBC 8, UBC 9, and UBC 27.

Castro-Ginard et al. (2019) and Cantat-Gaudin et al. (2019a) applied different methodologies to an area covering the Galactic anticentre. These latter authors found 53 and 41 previously unknown OCs, respectively, with 21 OCs in common. They found that the techniques are complementary, with none of the explored methods being able to detect all the objects.

These studies analysed a very particular region of the disc, where the star density is low compared to any other disc region. In the present work, we are able to find 42 out of the 53 (i.e 80%) OCs found in CG19 using the same methodology. The reason for not finding the 11 remaining OCs is that the parameters (L, $minPts$) used in the DBSCAN search (in the case of CG19) were optimised for that region of low stellar density. When optimising these parameters for a blind search of the whole Galactic disc, one has to account for regions with very different stellar densities. The optimal parameters chosen here are those that show the best performance in general terms, reaching a balance between low- and high-density regions. For the case of Cantat-Gaudin et al. (2019a), we were only able to find 24 out of the 41 reported OCs for similar reasons.

4.1.3. Dias et al. (2002) and Kharchenko et al. (2013)

These catalogues contain about 3 000 OCs each, compiled from heterogeneous data sources, which makes a cross-match with our candidates difficult. A candidate is considered to be tentatively matched with one object in those catalogues if its centres lie within a circle of 0.5° in radius. If two objects are tentatively matched by this positional criterium, we check if the mean values in $(\mu_{\alpha^*}, \mu_\delta)$ are compatible by performing a Welch t-test (Welch 1947), with a threshold p-value of 0.05 to reject the null hypothesis (i.e., to reject their compatibility). To perform the Welch t-test, we take the Kharchenko et al. (2013) most probable members for the cluster central part as the number of members for each OC in Kharchenko et al. (2013). These catalogues do not report the mean parallax for each OC but an estimation of the distance instead, with no associated uncertainty. Therefore, no comparison is made in this dimension.

Table 1. Some examples of the proposed OCs ordered by increasing l.

Name	α[deg]	δ[deg]	l[deg]	b[deg]	θ[deg]	ϖ[mas]	d[kpc]	μ_{α^*}[mas · yr^{-1}]	μ_δ[mas · yr^{-1}]	$V_{\rm rad}$[km · s^{-1}]	N ($N_{V_{\rm rad}}$)
					Class A						
UBC 91	267.42(0.07)	−28.76(0.07)	0.61(0.07)	−0.67(0.06)	0.09	0.42(0.03)	$2.37^{+0.18}_{-0.16}$	−0.59(0.09)	−1.12(0.11)	−(−)	83(0)
UBC 92	269.88(0.07)	−26.65(0.06)	3.53(0.07)	−1.49(0.06)	0.09	0.38(0.04)	$2.66^{+0.31}_{-0.25}$	2.13(0.09)	0.41(0.09)	−10.79(2.85)	105(2)
UBC 93	268.57(0.05)	−25.39(0.05)	4.03(0.04)	0.17(0.05)	0.07	0.34(0.03)	$2.95^{+0.25}_{-0.22}$	−0.93(0.11)	−1.88(0.09)	−(−)	52(0)
UBC 94	269.63(0.09)	−24.64(0.1)	5.17(0.1)	−0.29(0.08)	0.13	0.75(0.01)	$1.34^{+0.03}_{-0.02}$	−1.66(0.07)	−4.45(0.06)	−(−)	41(0)
UBC 95	268.25(0.06)	−22.17(0.09)	6.66(0.09)	2.06(0.07)	0.11	0.49(0.03)	$2.03^{+0.12}_{-0.1}$	−0.15(0.13)	−1.28(0.11)	−16.16(−)	84(1)
UBC 96	273.76(0.09)	−16.33(0.1)	14.31(0.11)	0.39(0.08)	0.14	0.62(0.02)	$1.62^{+0.07}_{-0.06}$	0.64(0.11)	0.93(0.08)	−(−)	41(0)
UBC 97	274.78(0.1)	−15.73(0.08)	15.3(0.1)	−0.18(0.08)	0.12	0.73(0.02)	$1.36^{+0.03}_{-0.03}$	−0.87(0.08)	−1.15(0.08)	−(−)	33(0)
UBC 98[a]	288.83(0.15)	−22.14(0.14)	15.38(0.13)	−14.93(0.16)	0.2	1.53(0.03)	$0.65^{+0.01}_{-0.01}$	0.56(0.11)	−6.66(0.17)	−(−)	23(0)
UBC 99[a]	282.02(0.09)	−18.3(0.09)	16.18(0.08)	−7.52(0.09)	0.13	1.06(0.03)	$0.94^{+0.03}_{-0.03}$	−1.16(0.1)	−4.1(0.13)	−(−)	52(0)
UBC 100	281.26(0.07)	−11.12(0.1)	22.3(0.1)	−3.65(0.07)	0.12	0.7(0.01)	$1.43^{+0.03}_{-0.03}$	−1.1(0.08)	−3.33(0.09)	−(−)	25(0)
UBC 101	279.5(0.09)	−7.14(0.07)	25.05(0.08)	−0.28(0.08)	0.11	0.42(0.02)	$2.41^{+0.15}_{-0.13}$	−0.31(0.09)	−3.03(0.08)	15.89(−)	54(1)
UBC 102	280.61(0.08)	−6.89(0.09)	25.77(0.09)	−1.15(0.08)	0.12	0.52(0.02)	$1.94^{+0.08}_{-0.08}$	−1.04(0.09)	−2.51(0.11)	9.97(−)	42(1)
UBC 103	280.63(0.05)	−6.6(0.08)	26.04(0.07)	−1.04(0.06)	0.09	0.28(0.03)	$3.54^{+0.35}_{-0.29}$	−0.4(0.09)	−2.27(0.09)	−3.99(−)	97(1)
UBC 104	280.69(0.05)	−6.26(0.07)	26.37(0.06)	−0.93(0.06)	0.08	0.29(0.03)	$3.45^{+0.44}_{-0.35}$	0.49(0.09)	−0.8(0.09)	−1.25(2.17)	61(2)
UBC 105	280.33(0.09)	−5.43(0.08)	26.94(0.08)	−0.23(0.08)	0.12	0.47(0.03)	$2.14^{+0.12}_{-0.11}$	0.46(0.11)	−0.99(0.09)	−(−)	75(0)
				⋮							
					Class B						
UBC 336	267.98(0.03)	−27.83(0.03)	1.66(0.03)	−0.62(0.03)	0.04	0.31(0.02)	$3.2^{+0.18}_{-0.16}$	0.75(0.08)	0.14(0.07)	−25.48(−)	22(1)
UBC 337	271.72(0.08)	−24.65(0.08)	6.09(0.07)	−1.94(0.08)	0.11	0.57(0.02)	$1.77^{+0.06}_{-0.06}$	0.47(0.08)	−0.72(0.07)	−(−)	40(0)
UBC 338	271.53(0.07)	−24.23(0.08)	6.37(0.08)	−1.59(0.06)	0.1	0.6(0.02)	$1.66^{+0.06}_{-0.06}$	0.01(0.08)	−1.77(0.09)	−15.86(−)	38(1)
UBC 339	271.31(0.04)	−23.31(0.05)	7.08(0.05)	−0.96(0.04)	0.06	0.39(0.02)	$2.59^{+0.12}_{-0.11}$	0.57(0.07)	−0.59(0.08)	−(−)	19(0)
UBC 340	270.77(0.09)	−22.66(0.07)	7.4(0.06)	−0.21(0.09)	0.11	0.7(0.02)	$1.42^{+0.03}_{-0.03}$	0.72(0.07)	−2.57(0.08)	−(−)	27(0)
UBC 341	276.45(0.1)	−17.06(0.09)	14.87(0.1)	−2.23(0.08)	0.13	0.48(0.03)	$2.1^{+0.13}_{-0.11}$	−0.21(0.12)	−1.49(0.1)	−3.75(−)	94(1)
UBC 342	273.91(0.17)	−14.92(0.17)	15.61(0.13)	0.94(0.2)	0.24	0.6(0.03)	$1.66^{+0.09}_{-0.08}$	−0.17(0.11)	−1.04(0.14)	−(−)	66(0)
				⋮							
					Class C						
UBC 572	280.42(0.07)	−21.95(0.06)	12.2(0.06)	−7.78(0.07)	0.09	0.65(0.02)	$1.54^{+0.05}_{-0.05}$	0.98(0.1)	−0.63(0.11)	−33.02(7.31)	23(2)
UBC 573	275.01(0.07)	−9.44(0.09)	20.95(0.09)	2.58(0.07)	0.11	0.53(0.02)	$1.88^{+0.09}_{-0.08}$	−0.18(0.1)	−4.48(0.1)	−(−)	17(0)
UBC 574[a]	282.32(0.08)	−4.36(0.09)	28.8(0.09)	−1.51(0.08)	0.12	0.58(0.0)	$1.73^{+0.01}_{-0.01}$	1.06(0.02)	0.21(0.04)	−10.15(−)	9(1)
UBC 575	291.01(0.08)	−5.13(0.11)	32.05(0.1)	−9.58(0.09)	0.13	0.91(0.02)	$1.09^{+0.02}_{-0.02}$	−0.3(0.07)	−5.18(0.08)	−(−)	9(0)
UBC 576	284.68(0.04)	0.42(0.06)	34.13(0.06)	−1.43(0.04)	0.07	0.74(0.02)	$1.34^{+0.03}_{-0.03}$	−0.74(0.08)	−3.56(0.09)	−(−)	17(0)
UBC 577	282.17(0.05)	22.12(0.09)	52.54(0.09)	10.47(0.05)	0.1	1.0(0.02)	$1.0^{+0.02}_{-0.02}$	−1.04(0.11)	3.07(0.07)	−0.59(16.75)	9(4)
				⋮							

Notes. The parameters shown are the mean (and standard deviation) for the (N) members found also including the apparent angular size (θ) and estimated distance (d) with one sigma confidence interval. Radial velocity is included when available and is computed with $N_{V_{\rm rad}}$ members. The name follows the numeration of CG19. The full list can be found online at the CDS. [(a)]coincidence with Sim et al. (2019) or Liu & Pang (2019), see Sect. 4.1.5. [(b)]tentative identification with Kharchenko et al. (2013), see Sect. 4.1.3.

Most of the coincidences with these catalogues have already been taken into account by the cross-match of our candidates with Cantat-Gaudin et al. (2018). However, we find five OCs that are compatible with the position and proper motion (with p-value > 0.05) criteria described above. These objects are flagged in our Table 1.

With our methodology we are also able to identify objects related with known star forming regions. Some of these are listed in the aforementioned catalogues. We find objects related with σ-Ori, Collinder 228, Bochum 10, NGC 1980, NGC 1981, NGC 6514, NGC 6530 and NGC 6604 (Reipurth 2008a,b). While σ-Ori is listed as a possible stellar association in Dias et al. (2002) it is considered a moving group in Kharchenko et al. (2013). Collinder 228 has variable extinction according to Dias et al. (2002) and has nebulosity according to Kharchenko et al. (2013). Bochum 10 and NGC 6604 are normal clusters in both catalogues. NGC 1980 and 1918 are considered in Dias et al. (2002) to be a normal OC and an embedded OC in a possible OB association, respectively, while they are considered as nebulosities in Kharchenko et al. (2013). Finally, NGC 6514 and NGC 6530 are listed as normal OCs in Dias et al. (2002) and as nebulosities in Kharchenko et al. (2013).

4.1.4. Bica et al. (2019)

Bica et al. (2019) compiled a catalogue with 10 978 stellar clusters, associations, and candidates reported previous to *Gaia* DR2, by combining catalogues from different studies on different surveys (Digital Sky Survey, 2MASS, WISE, VVV, Spitzer and Herschel). Among the groups listed by Bica et al. (2019), the OCs amount to 3000. Others are about 300 globular clusters, about 5 000 embedded clusters (which are hardly seen by *Gaia*) and about 1200 asterisms. The coincidences among OCs have been discussed above. We find 45 additional coincidences

with their catalogue. These matches correspond to globular clusters (GCs), which Bica et al. (2019) include, and which were not taken into account in the previous cross-matches.

The detection of GCs using our methodology is a good diagnostic test. On the one hand, DBSCAN is able to detect these GCs repeatedly among all the DBSCAN runs (for all optimal $L, minPts$ parameters). For example, ω-Cen, the most massive GC known with $4 \times 10^6\,M\odot$, is the cluster found the highest number of times by our algorithm. On the other hand, the ANN was trained with CMDs from real OCs and from simulated stellar populations at different ages. Since OCs are mostly young objects, the contribution to the recognition of such an old isochrone (>10 Gyr) comes from the simulated data (with the appropriate error model). Therefore, the use of simulated CMDs not only contributes by increasing the training set, but also allows the ANN to recognise cases in the real data that were trained using simulations.

4.1.5. Sim et al. (2019) and Liu & Pang (2019)

Recently, Sim et al. (2019) found 207 new OCs located within 1 kpc by visually inspecting *Gaia* DR2 proper-motion diagrams searching for overdensities. The criteria used to consider one of these objects as matched with one of our candidates are similar to those discussed in the previous section. We consider an identification as tentative if the centres of both objects lie within a circle of 0.5° in radius and then we compare the rest of the astrometric parameters. Firstly, we find that one of these objects, UPK 19, corresponds to UBC 32, already reported by CG18. In this case, UPK 19 and UBC 32 are separated by 0.18° in the sky and the rest of their mean astrometric parameters differ by $(2\sigma_{\varpi}, 0.14\sigma_{\mu_{\alpha^*}}, 0.15\sigma_{\mu_\delta})$. Secondly, eight of our OC candidates are identified with one UPK object. All the identifications are compatible within 1σ in proper motions. The mean parallaxes are compatible within 1.91σ (at most). This larger discrepancy is because Sim et al. (2019) do not report mean parallaxes but the estimated distance instead, and the transformation from parallax to distance may lead to big differences. However, we consider these objects as matched.

Similarly, Liu & Pang (2019) identified 2 443 star clusters in the Galactic disc using a clustering algorithm in the 5D astrometric space $(l, b, \varpi, \mu_{\alpha^*}, \mu_\delta)$. Most of these star clusters were previously reported. Of their high confidence candidates, 76 are reported as new objects. Among these 76, we find 4 coincidences with CG18 and CG19. These are the cases for their clusters with IDs 1973, 2143, 2230, and 2385 which are identified with UBC 74, UBC 72, UBC 56, and UBC 7 (from CG18 and CG19), respectively. All the identifications are within 0.5° and within 2σ in $(\varpi, \mu_{\alpha^*}, \mu_\delta)$. From our list of new OC candidates, we find 45 cases that are compatible with one of the 76 from Liu & Pang (2019), with the same matching criteria.

4.2. Newly found open clusters

We select as new OCs those candidates that are found more than three times among all the runs to which we applied the method (each time with a different set of optimal parameters $(L, minPts)$; see Sect. 2). This results in a list of 676 tentative new structures.

These structures are further divided into three categories: new OCs of class A, class B, and class C; plus other stellar structures that were discarded. We classify the new OCs into these categories by visually inspecting the CMD of the candidates, and the distribution of their member stars in the astrometric space (Fig. 2), including radial velocity when available. Table 1[4] lists the mean parameters of the candidates proposed as OCs $(\alpha, \delta, l, b, \varpi, \mu_{\alpha^*}, \mu_\delta, V_{\rm rad})$ as well as the apparent angular size computed as $\theta = \sqrt{\sigma_l^2 + \sigma_b^2}$. An estimation of the distance by the inversion the mean parallax is also included, with (asymmetric) confidence intervals. A list with the members for each OC, as computed by DBSCAN, is available in Table 2[5].

The number of OCs in these categories are 245 OCs in class A, 236 in class B, and 101 in class C. Table 3 shows the mean $(\theta, \varpi, \sigma_{\mu_{\alpha^*}}, \sigma_{\mu_\delta}, N, N_{\rm found})$ for each class. Figure 2 shows one OC from each category. Class A clusters typically show a high concentration of the member stars in all five astrometric parameters $(l, b, \varpi, \mu_{\alpha^*}, \mu_\delta)$, and a clean isochrone in a CMD. Clusters in class B show a more sparse distribution in the five astrometric parameters, and many include a low number of contaminant (field) stars which can be seen more clearly in the CMD. While clusters in class C are typically poorly populated and show an isochrone that could have a higher degree of contaminant stars. From the OCs classified as class A, 115(47%) have stars evolved beyond the main sequence; this represents the oldest population of this class.

From the OCs classified in class A, 139 have stars with radial velocity measurements, and 85 contain more than two stars with radial velocity measurements. For those, the mean dispersion of the radial velocities within cluster member stars is 5.47 km s^{-1}. For the OCs in class B, 93 from 236 have radial velocity measurements, and 42 have more than two stars with these measurements. The mean radial velocity dispersion for class B clusters is 6.59 km s^{-1}. Finally, for class C clusters, only 38 have stars with radial velocities, of which 20 have measurements for more than two stars. In this latter case, the mean dispersion is 11.81 km s^{-1}. A certain amount of this dispersion could be due to multiplicity. Since the clustering did not take into account the radial velocity in order to detect the OCs, this external check shows the frequency of contaminant stars that clusters in each class may have.

4.2.1. Comments on the new open clusters

The newly found clusters have mean parallaxes ranging from 0.09 to 2.58 mas. Estimating their distance as the inverse of their mean parallax yields distances from 387 pc to ∼11 kpc. Inverting parallaxes is however not a good approach for objects with large relative parallax uncertainties (Luri et al. 2018), and a more sophisticated method should be applied to estimate the distance to the most distant OCs. Figure 3 shows a comparison between the distribution of parallaxes of the known OCs with the new findings, with light orange representing previously known OCs and light blue representing OCs found in this study. The OCs found represent an increase in the OC census of 18% in clusters closer than 1 kpc, 54% in clusters at between 1 and 2 kpc and 49% in clusters further than 2 kpc.

The distribution of the new OCs in the Galactic plane is shown in Fig. 4 (projection in the $X - Y$ plane in Fig. 5). Of the new OCs, 83.5% are located at Galactic latitudes $|b| < 5°$, 8.2% are located within $5° < |b| < 10°$ and only 8.3% are found at $|b| > 10°$. The black dots represent the newly found OCs (their angular size is proportional to the number of members) while the red density contours represent the known ones. We see that the distribution of the new OCs follows a similar distribution to the previously reported ones. In these figures, we can see that the present study detected relatively few new objects between Galactic longitudes of 140° and 210°. This region has already

[4] Full version, with the 582 OCs, available online at the CDS.
[5] Table 2 is only available online at the CDS.

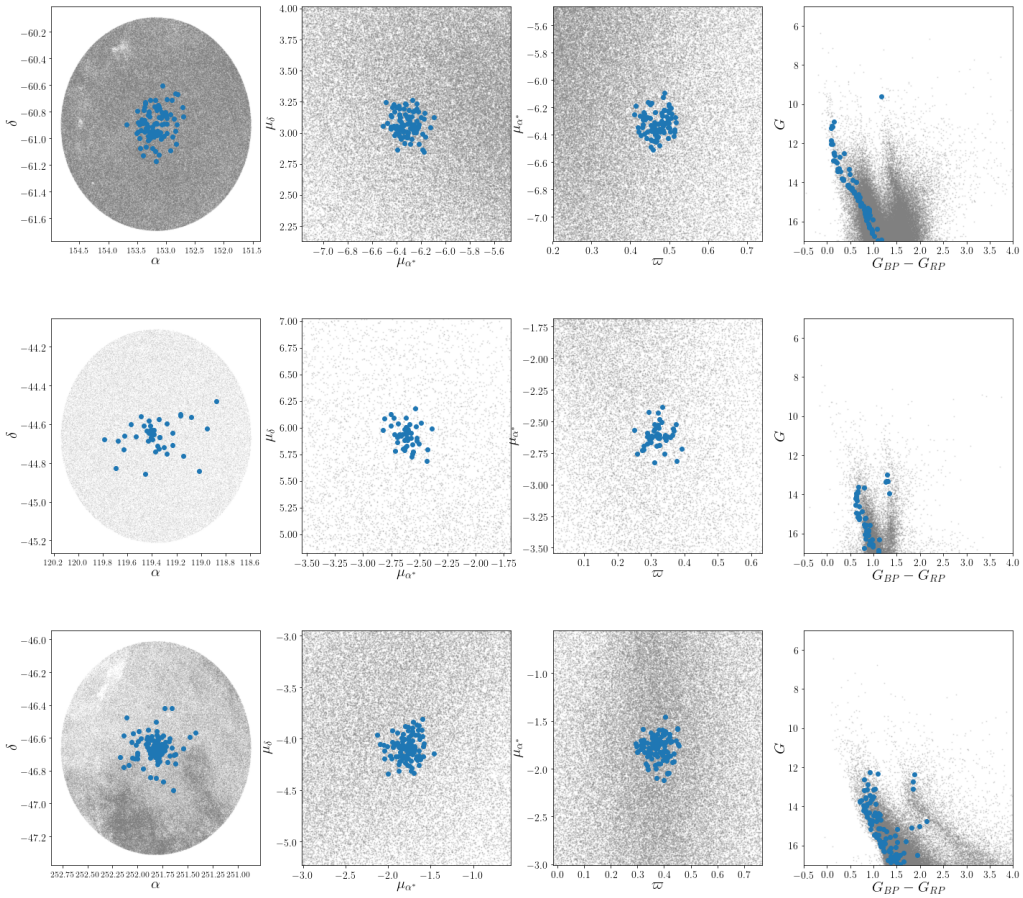


Fig. 2. Examples of class A (*top row*), class B (*middle row*), and class C (*bottom row*) clusters. The columns represent, *from left to right*, a distribution of the member stars (in blue) and field stars (grey) for: i) position in (α, δ), ii) proper motions in $(\mu_{\alpha^*}, \mu_\delta)$, iii) distribution in (ϖ, μ_{α^*}), and iv) a CMD in G vs. $G_{\rm BP} - G_{\rm RP}$. Rows correspond to OCs UBC 257, UBC 478, and UBC 669, respectively. Classes A, B, and C correspond to different levels of reliability (see Sect. 4.2).

Table 3. Mean parameters for each of the OC classes.

	θ	ϖ	$\sigma_{\mu_{\alpha^*}}$	σ_{μ_δ}	N	$N_{\rm found}$
Class A	0.14	0.58	0.11	0.11	78.3	25.3
Class B	0.12	0.44	0.10	0.10	51.1	16.3
Class C	0.11	0.36	0.11	0.11	26.3	10.2

Notes. The parameters shown are angular size, parallax, proper motions, number of members, and number of times found within all runs of the method.

been the target of two cluster searches using *Gaia* DR2 data (in CG19 and Cantat-Gaudin et al. 2019a), and fewer objects are left to be discovered here. Figure 6 shows the distribution of the known (red dots) and newly found OCs (black dots). We see that none of the new OCs are found at high $|Z_{\rm Gal}|$ in the inner disc ($R_{\rm Gal} < 7$ kpc) where real OCs are unlikely to be found (Cantat-Gaudin & Anders 2020).

4.2.2. Specific remarks on UBC 274

UBC 274 is a newly found OC at a relatively low Galactic latitude ($b \sim -12.8°$) and at a distance of $d \sim 2$ kpc. It is the

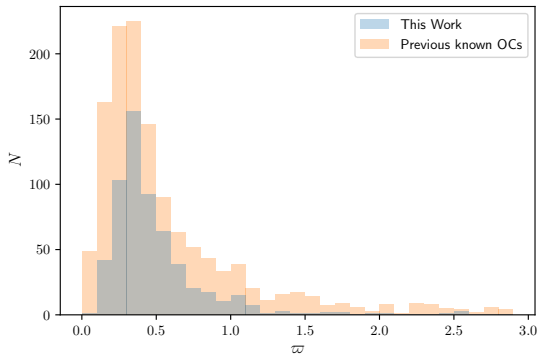


Fig. 3. Parallax histogram of the new OCs (light blue) and OCs known previous to this study (light orange), i.e. CG18, CG19, Cantat-Gaudin et al. (2018), and Cantat-Gaudin et al. (2019a).

clearest new detection made with our method, that is, the cluster found the highest number of times within the pairs of $(L, minPts)$ explored, one of the most massive OCs we can find (with 365 stars), and one of the biggest in size. There are 15 stars with radial velocity measurements, of which 13 are in agreement with

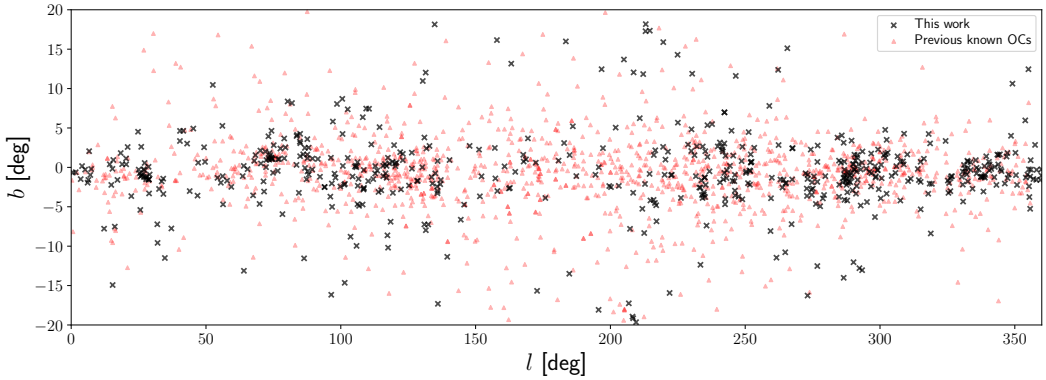


Fig. 4. Distribution of the OC census in l vs. b. Black crosses represent new OCs while red triangles represent OCs in CG18, CG19, Cantat-Gaudin et al. (2018), and Cantat-Gaudin et al. (2019a).

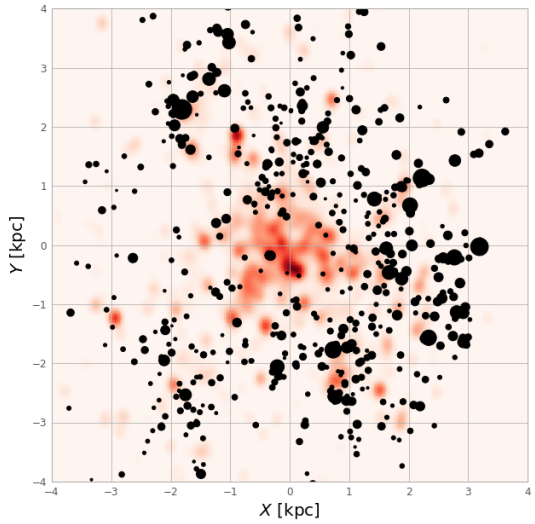


Fig. 5. Distribution of the OCs projected in the $X - Y$ plane. Previously known OCs (CG18,CG19, Cantat-Gaudin et al. 2018, 2019a) are shown as a density map in red. Newly found OCs reported here are shown as black dots, where the size is proportional to the number of members of each cluster.

a mean value of $-22.92\,\mathrm{km\,s^{-1}}$; they have a standard deviation of $1.26\,\mathrm{km\,s^{-1}}$, and so they are compatible with the membership. The non-compatible stars have a radial velocities of -10.68 and $-8.00\,\mathrm{km\,s^{-1}}$, at 9σ and 11σ difference, respectively; they may be field stars or multiple stars.

Figure 7 shows a distribution of the member stars of UBC 274 in the five astrometric dimensions, and in a CMD. These members show a concentrated clump in $(\varpi, \mu_{\alpha^*}, \mu_\delta)$, well distinguishable from the field stars. UBC 274 shows an elongated shape in the spatial distribution in the direction of the proper motion. The CMD shows a clean isochrone from which we can estimate an age of ∼3 Gyr. Fewer than 20% of the previously known clusters have ages greater than 1 Gyr, and only 5% have ages greater than 2 Gyr. We can also identify some blue straggler candidates.

Tidal tails in intermediate and old age OCs due to disruption by the gravitational field have been detected in well-known clusters like the Hyades, Praesepe, and Coma Berenices by Röser et al. (2019), Röser & Schilbach (2019), Tang et al. (2019) based on *Gaia* DR2. The elongation of UBC 274 (Fig. 8) suggests that it is another example of disruption taking place.

4.2.3. Substructure in star forming regions

It has been known for a long time that star forming regions are in groups and form structures and filaments (e.g. Bouy & Alves 2015). *Gaia* DR2 has allowed for the spatial and kinematic substructure of several star forming regions to be accurately determined (Zari et al. 2018; Lim et al. 2019; Galli et al. 2019; Cantat-Gaudin et al. 2019b) and has even allowed the internal dynamics of these groups to be studied. We identified several objects possibly related to known star forming regions. For instance, in the Carina Nebula, we are able to find seven groups which are related to the nebula. Figure 9 shows the spatial distribution of those groups. The points in different colours represent the stars found for each of the new UBC clusters, and dashed circles represent known clusters related to the nebula. We see that even in a blind search, we are able to detect several subgroups which could be related to the same structure. For instance, Collinder 228 and UBC 505 share sky coordinates but they are found as two different objects due to the difference in parallax, which is 0.42 and 0.29 mas, respectively.

5. Conclusions

We devised a methodology to blindly search for open clusters in the Galactic disc using the *Gaia* DR2 astrometric and photometric data. The method is based on two ML algorithms, first an unsupervised learning algorithm (DBSCAN) detects overdensities in the astrometric space $(l, b, \varpi, \mu_{\alpha^*}, \mu_\delta)$ and then a supervised ANN recognises the isochrone pattern that some of these statistical overdensities (the ones that correspond to real OCs) show in a CMD, identifying them as actual OCs.

In order to scan the whole Galactic disc using a strategy driven by the targeted OCs and not the computational limitations, the method has to be adapted to a Big Data environment. We use the PyCOMPSs parallelisation scheme to deploy the clustering algorithm to the MareNostrum Supercomputer at the BSC. This enables us to search for overdensities independently of the density of the region, for example higher density regions such as the direction of the Galactic centre. Once the statistical densities are detected, and because of the large number of them, a more reliable photometric confirmation of the

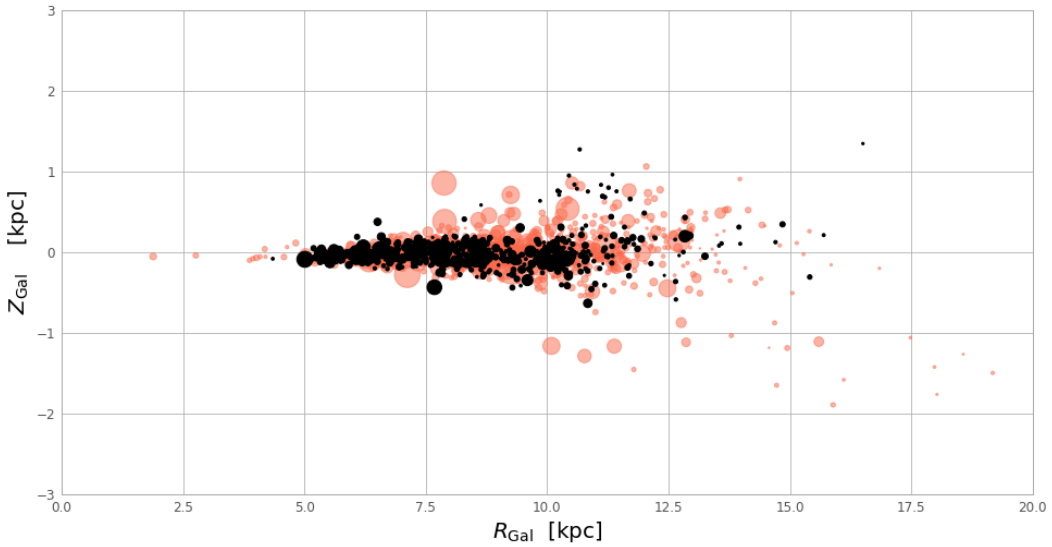


Fig. 6. Distribution of the OCs in $R - Z$ in Galacto-centric coordinates. Previously known OCs (CG18,CG19, Cantat-Gaudin et al. 2018, 2019a) are shown as red dots while newly found OCs are shown in black dots; the sizes of the dots are proportional to the number of members of each cluster.

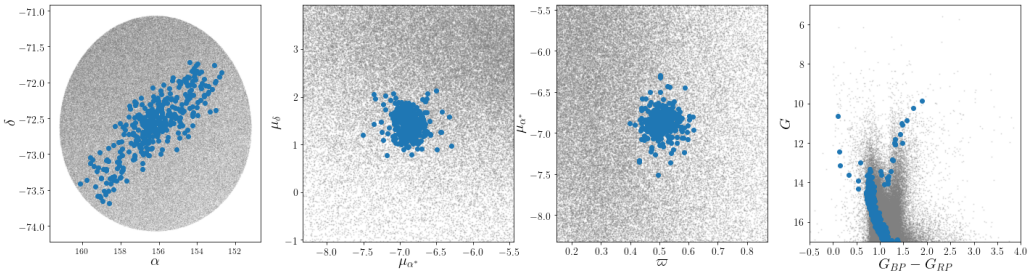


Fig. 7. Distribution of the member stars of UBC 274 (blue points) in comparison with field stars (grey points). The leftmost plot is a distribution in position (α, δ). The inner left plot shows the proper motion vector diagram while the inner right plot includes the parallax (ϖ, μ_{α^*}). The rightmost plot is a CMD.

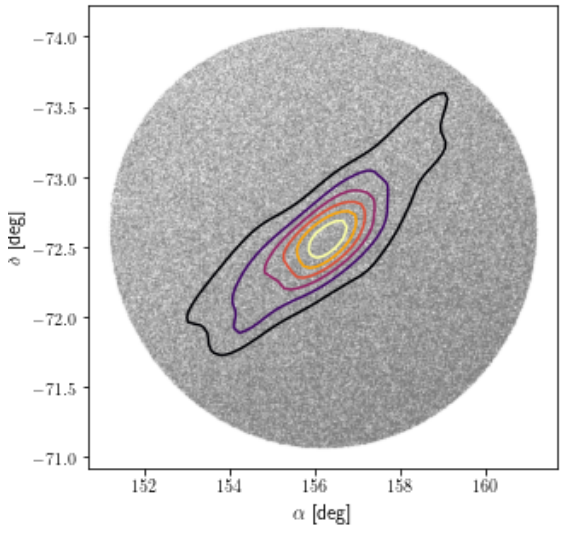


Fig. 8. Density contours for the members in cluster UBC 274, and field stars (grey points). UBC 274 shows an elongated shape in its outskirts.

candidate is needed. This is achieved by applying deep learning methods to an ANN, which outperform the simple multi-layer perceptron when 2D correlations are present (a CMD in G vs. $G_{\rm BP}$–$G_{\rm RP}$).

The methodology is able, even in a blind search where the parameters are tuned to find the largest number of OCs, to find substructures in richer regions or even features of individual objects such as their tidal tails. This suggests that with a fine tuning of the parameters, the methodology can be adapted to study single objects in more detail.

The method was first devised using TGAS data in CG18, and successfully applied to a low-density disc region (the Galactic anticentre) using *Gaia* DR2 in CG19, finding a total of 76 new OCs. In this paper, the method is applied to the whole Galactic disc ($|b| < 20°$) up to a magnitude of $G = 17$, finding a total of 582 previously unknown OCs, which represents a 45% increase in the detection of this class of objects.

The OCs found represent an increase of 18% up to 1 kpc, 54% between 1 and 2 kpc, and 49% further than 2 kpc. The mean angular size of the clusters found is 0.13° and the mean number of members is 58.3. One of the most interesting clusters found is UBC 274, which is about 3 Gyr old at $b = -12.8°$, and shows an elongated shape due to disruption by tidal tails.

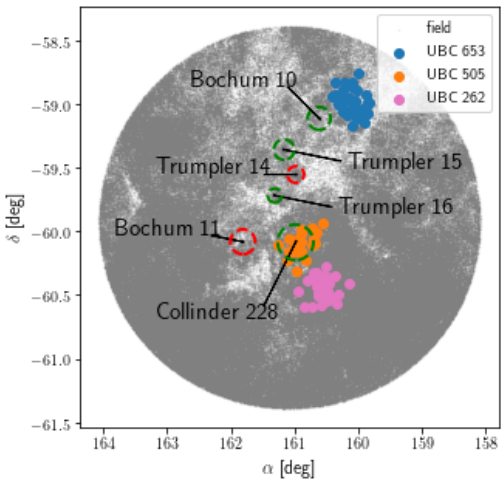


Fig. 9. Region around the Carina Nebula. Grey points represent field stars, while points in blue, orange, and pink represent UBC 653, UBC 505, and UBC 262 respectively. The dashed circle represents the locations of the OCs Cantat-Gaudin et al. (2018), which are related to the Carina Nebula. Dashed green circles are objects found by our method and dashed red circles are objects not found.

Acknowledgements. ACG thanks Dr. T. Antoja for her comments on the writing; ACG also thanks Dr. Jordi Vitrià and Dr. Santi Seguí for their useful comments on the ANN implementation and training. This work has made use of results from the European Space Agency (ESA) space mission *Gaia*, the data from which were processed by the *Gaia Data Processing and Analysis Consortium* (DPAC). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement. The *Gaia* mission website is `http://www.cosmos.esa.int/gaia`. The authors are current or past members of the ESA *Gaia* mission team and of the *Gaia* DPAC. This work was partially supported by the MINECO (Spanish Ministry of Economy) through grant ESP2016-80079-C2-1-R and RTI2018-095076-B-C21 (MINECO/FEDER, UE), and MDM-2014-0369 of ICCUB (Unidad de Excelencia "María de Maeztu"). This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement H2020-MSCA-COFUND-2016-754433. This work has been partially supported by the Spanish Government (SEV2015-0493), by the Spanish Ministry of Science and Innovation (contract TIN2015-65316-P), by Generalitat de Catalunya (contract 2014-SGR-1051). The research leading to these results has also received funding from the collaboration between Fujitsu and BSC (Script Language Platform). L.C. acknowledges support from "programme national de physique stellaire" (PNPS) and from the "programme national cosmologie et galaxies". This research has made use of the TOPCAT (Taylor et al. 2005). This research has made use of the VizieR catalogue access tool, CDS, Strasbourg, France. The original description of the VizieR service was published in A&AS 143, 23.

References

Álvarez Cid-Fuentes, J., Solà, S., Álvarez, P., Castro-Ginard, A., & Badia, R. 2019, Proceedings of the 15th International Conference of Science , 96
Bica, E., Pavani, D. B., Bonatto, C. J., & Lima, E. F. 2019, AJ, 157, 12
Bouy, H., & Alves, J. 2015, A&A, 584, A26
Bressan, A., Marigo, P., Girardi, L., et al. 2012, MNRAS, 427, 127
Cantat-Gaudin, T., & Anders, F. 2020, A&A, 633, A99
Cantat-Gaudin, T., Jordi, C., Vallenari, A., et al. 2018, A&A, 618, A93
Cantat-Gaudin, T., Krone-Martins, A., Sedaghat, N., et al. 2019a, A&A, 624, A126
Cantat-Gaudin, T., Jordi, C., Wright, N. J., et al. 2019b, A&A, 626, A17
Castro-Ginard, A., Jordi, C., Luri, X., et al. 2018, A&A, 618, A59
Castro-Ginard, A., Jordi, C., Luri, X., Cantat-Gaudin, T., & Balaguer-Núñez, L. 2019, A&A, 627, A35
Dias, W. S., Alessi, B. S., Moitinho, A., & Lépine, J. R. D. 2002, A&A, 389, 871
Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. 1996, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96 (AAAI Press), 226
Evans, D. W., Riello, M., De Angeli, F., et al. 2018, A&A, 616, A4
Gaia Collaboration (Prusti, T., et al.) 2016, A&A, 595, A1
Gaia Collaboration (Brown, A. G. A., et al.) 2018, A&A, 616, A1
Galli, P. A. B., Loinard, L., Bouy, H., et al. 2019, A&A, 630, A137
Hinton, G. 1989, Artif. Intell., 40, 185
Kharchenko, N. V., Piskunov, A. E., Schilbach, E., Röser, S., & Scholz, R.-D. 2013, A&A, 558, A53
Kroupa, P. 2001, MNRAS, 322, 231
LeCun, Y. A., Bottou, L., Orr, G. B., & Müller, K.-R. 2012, Efficient BackProp (Berlin, Heidelberg: Springer, Berlin Heidelberg), 9
Lim, B., Nazé, Y., Gosset, E., & Rauw, G. 2019, MNRAS, 490, 440
Lindegren, L., Lammers, U., Bastian, U., et al. 2016, A&A, 595, A4
Lindegren, L., Hernández, J., Bombrun, A., et al. 2018, A&A, 616, A2
Liu, L., & Pang, X. 2019, ApJS, 245, 32
Luri, X., Palmer, M., Arenou, F., et al. 2014, A&A, 566, A119
Luri, X., Brown, A. G. A., Sarro, L. M., et al. 2018, A&A, 616, A9
Maíz Apellániz, J., & Weiler, M. 2018, A&A, 619, A180
Michalik, D., Lindegren, L., & Hobbs, D. 2015, A&A, 574, A115
Paszke, A., Gross, S., Chintala, S., et al. 2017, NIPS-W
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, J. Mach. Learn. Res., 12, 2825
Reipurth, B. 2008a, Handbook of Star Forming Regions, Volume I: The Northern Sky, 4
Reipurth, B. 2008b, Handbook of Star Forming Regions, Volume II: The Southern Sky, 5
Robin, A. C., Luri, X., Reylé, C., et al. 2012, A&A, 543, A100
Röser, S., & Schilbach, E. 2019, A&A, 627, A4
Röser, S., Schilbach, E., & Goldman, B. 2019, A&A, 621, L2
Sim, G., Lee, S. H., Ann, H. B., & Kim, S. 2019, J. Korean Astron. Soc., 52, 145
Tang, S.-Y., Pang, X., Yuan, Z., et al. 2019, ApJ, 877, 12
Taylor, M. B. 2005, in Astronomical Data Analysis Software and Systems XIV, eds. P. Shopbell, M. Britton, & R. Ebert , ASP Conf. Ser., 347, 29
Tejedor, E., Becerra, Y., Alomar, G., et al. 2017, Int. J. High Perform. Comput. Appl., 31, 66
Welch, B. L. 1947, Biometrika, 34, 28
Zari, E., Hashemi, H., Brown, A. G. A., Jardine, K., & de Zeeuw, P. T. 2018, A&A, 620, A172

Part II

OPEN CLUSTERS IN A GALACTIC CONTEXT

5

MILKY WAY SPIRAL STRUCTURE AND EVOLUTION TRACED BY OPEN CLUSTERS

This chapter contains the submitted version of our last paper, which uses open clusters as tracers of the Milky Way spiral arms.

Our knowledge of the OC population has dramatically improved thanks to the *Gaia* data, as seen in previous Chapters. A number of named structures, thought to be OCs, were found to be not physical groups whilst other new OCs have been detected, increasing the number of objects characterised with *Gaia* from $\sim$ 1200 to more than 2000 (Cantat-Gaudin et al. 2020a). It is known that star formation happen preferentially in spiral arms. Therefore the youngest clusters (tens of Myr), which are not far from their birthplace, trace the spiral arm structure in the disc and its evolution. Given the improvements in the OC census, it is timely to re-examine the spiral structure and evolution of our Milky Way, and this is the goal of this Chapter.

We have used the OC population as the main tracers for the dynamics of the Milky Way spiral arms over the last $\sim$ 80 Myr. We were able to disfavour density waves as the main drivers for the formation spiral arms in our Galaxy, finding a more transient behaviour with the arms co-rotating with stars at all Galactocentric radius. We also used the youngest OCs ($<$ 30 Myr), in addition to known high-mass star-forming regions, to define the present spiral arm segments in the Solar neighbourhood.

Astronomy & Astrophysics manuscript no. manuscript
February 23, 2021
©ESO 2021

On the Milky Way spiral arms from open clusters in *Gaia* EDR3

A. Castro-Ginard[1], P.J. McMillan[2], X. Luri[1], C. Jordi[1], M. Romero-Gómez[1], T. Cantat-Gaudin[1], L. Casamiquela[3], Y. Tarricq[3], C. Soubiran[3], and F. Anders[1]

[1] Dept. Física Quàntica i Astrofísica, Institut de Ciències del Cosmos (ICCUB), Universitat de Barcelona (IEEC-UB), Martí i Franquès 1, E08028 Barcelona, Spain
e-mail: acastro@fqa.ub.edu
[2] Lund Observatory, Department of Astronomy and Theoretical Physics, Lund University, Box 43, SE-22100, Lund, Sweden
[3] Laboratoire d'Astrophysique de Bordeaux, Univ. Bordeaux, CNRS, B18N, allée Geoffroy Saint-Hilaire, 33615 Pessac, France

Received date / Accepted date

ABSTRACT

Context. The physical processes driving the formation of Galactic spiral arms are still under debate. Studies using open clusters favour the description of the Milky Way spiral arms as long-lived structures following the classical density wave theory. Current studies comparing the *Gaia* DR2 field stars kinematic information of the Solar neighbourhood to simulations, find a better agreement with short-lived arms with a transient behaviour.
Aims. Our aim is to provide an observational, data-driven view of the Milky Way spiral structure and its dynamics using open clusters as the main tracers, and to contrast it with simulation-based approaches. We use the most complete catalogue of Milky Way open clusters, with astrometric *Gaia* EDR3 updated parameters, estimated astrophysical information and radial velocities, to re-visit the nature of the spiral pattern of the Galaxy.
Methods. We use a Gaussian mixture model to detect overdensities of open clusters younger than 30 Myr that correspond to the Perseus, Local, Sagittarius and Scutum spiral arms, respectively. We use the birthplaces of the open cluster population younger than 80 Myr to trace the evolution of the different spiral arms and compute their pattern speed. We analyse the age-distribution of the open clusters across the spiral arms to explore the differences in the rotational velocity of stars and spiral arms.
Results. We are able to increase the range in Galactic azimuth where present-day spiral arms are described, better estimating its parameters by adding 264 young open clusters to the 84 high-mass star forming regions used so far, thus increasing by a 314% the number of tracers. We use the evolution of the open clusters from their birth positions to find that spiral arms nearly co-rotate with field stars at any given radius, discarding a common spiral pattern speed for the spiral arms explored.
Conclusions. The derivation of different spiral pattern speeds for the different spiral arms disfavours classical density waves as the main drivers for the formation of the Milky Way spiral structure, and is in better agreement with simulation-based approaches that tend to favour transient spirals. The increase in the number of known open clusters, as well as in their derived properties allows us to use them as effective spiral structure tracers, and homogenise the view from open clusters and field stars on the nature of the Galactic spiral arms.

Key words. Galaxy: disc — open clusters and associations: general — astrometry — Methods: data analysis

1. Introduction

The location of our Solar system within the Milky Way disc makes it challenging to obtain a detailed picture of its structure. This is particularly true for the spiral structure. The number and location of the spiral arms still remains unclear, as well as their nature. Lin & Shu (1964) proposed a theoretical mechanism for the formation of spiral arms, widely known as the density wave theory, where the spiral arms rotate like a rigid solid at a constant angular velocity (*i.e.* pattern speed) in spite of the differential rotation of the stars and interstellar medium, causing the spiral arms to be long-lived (see Shu 2016, for a review of the classic theory). Alternatively, Toomre (1964) proposed that spiral arms could be reforming short-lived structures composed by individual arms, each of them behaving as a wave at a constant pattern speed, which overlap causing transient spiral arms with no global spiral pattern speed (Quillen et al. 2011; Sellwood & Carlberg 2014). This latter short-lived arms can be also explained with material arms that co-rotate with disc stars (Wada et al. 2011; Grand et al. 2012), causing the spiral pattern to grow from local gravitational instabilities and then to disappear, with continuous instabilities regenerating the pattern again.

Since the first attempts to explain the nature of the arms almost 60 years ago, no clear conclusion has been reached. Dobbs & Baba (2014) proposed types of observational evidence to shed light on the nature of the spiral structure, based on the different rotation velocities for the spiral pattern and disc stars. These strategies include the direct derivation of the spiral pattern speed for each arm, which will help in favouring either a density wave theory, where the arms share a global constant spiral pattern speed, or a transient behaviour which shows a spiral pattern speed decreasing with Galactocentric radius (Shabani et al. 2018). The distribution of ages of the stellar clusters across any spiral arm also indicates a difference in the velocities of both structures, the distribution of stars and the arm. Given the improvements on the open cluster catalogue made in light of *Gaia* second data release (*Gaia* DR2, Gaia Collaboration et al. 2018a), and using them as main tracers of Galactic spiral structure, both observational evidences can be pursued for the first

time providing a new view of the nature of the Milky Way spiral arms.

Open clusters (OCs) are excellent tracers of the spatial structure of the young stellar population in the Galactic disc. They are groups of stars, gravitationally bound, which were born from the same molecular cloud and, therefore, have very similar positions, velocities, ages and chemical composition (Lada & Lada 2003). For an OC, the estimation of its properties such as the parallax, proper motion, radial velocity, age or extinction is more reliable than for individual field stars because it relies on a (large) number of members, whose parameters are averaged.

Using OCs as spiral arms main tracers, Naoz & Shaviv (2007) found that spiral pattern speeds for the Perseus, Local and Sagittarius spiral arms decrease with Galactocentric radius, finding evidence for multiple spiral sets. Also using young OCs as spiral arms main tracers and *Gaia* DR2, Dias et al. (2019) obtained a common pattern speed for the Perseus, Local and Sagittarius spiral arms of 28.2 ± 2.1 km s^{-1} kpc^{-1}, supporting the idea of the density wave nature of the spiral structure. This results in a co-rotation radius, *i.e.* Galactocentric radius at which the spiral pattern speed coincides with the velocity from the Galactic rotation curve, of $R_c = 8.51 \pm 0.64$ kpc. Junqueira et al. (2015) used a sample of giant stars from OCs observed by APOGEE DR10 (Anders et al. 2014) to find a pattern speed of 23.0 ± 0.5 km s^{-1} kpc^{-1}, with a corresponding co-rotation radius of $R_c = 8.74$ kpc, compatible to the previous result within uncertainties. However, even though R_c is a fundamental parameter in the density wave scenario, there is not a consensus on its value yet. Different studies, using different tracers, place it from 6.7 kpc to beyond the Perseus arm, located at ~ 10 kpc (Drimmel & Spergel 2001; Monguió et al. 2015; Michtchenko et al. 2018).

From a complementary point of view, by comparing the kinematic substructure of field stars in the Solar neighbourhood to simulated data, Hunt et al. (2018) showed that a simulated Galaxy with transient spiral arms reproduces the arches and ridges seen in the velocity distribution of *Gaia* DR2 (Gaia Collaboration et al. 2018b; Antoja et al. 2018; Ramos et al. 2018). In this transient scenario R_c is not an important parameter since the spiral arms would co-rotate with stars at their Galactocentric radius, causing short-lived arms. A number of authors looking at the field population, some using simulation-based approaches, tended to favour a transient nature for the spirals (Quillen et al. 2018; Hunt et al. 2020; Kamdar et al. 2020).

Since the publication of the *Gaia* DR2, hundreds of new OCs have been detected (Castro-Ginard et al. 2018, 2019, 2020; Liu & Pang 2019; Sim et al. 2019) and have been added to the OCs known before (and confirmed by) *Gaia* DR2 (Cantat-Gaudin et al. 2018). For this compendium of OCs, information about age, distance and line-of-sight extinction was computed (Cantat-Gaudin et al. 2020) and radial velocities were compiled from different ground-based spectroscopic surveys (Tarricq et al. 2020). Altogether results in a robust OCs catalogue that offers the chance to trace the spiral structure of the Galactic disc (in the Solar neighbourhood) and its evolution over the past few hundred Myr.

Our aim for this paper is to use this recent and homogeneous OC sample with information on astrometric mean parameters, radial velocities and astrophysical parameters available to discriminate as far as possible among different theories for the nature of the spiral structure of the Milky Way, supporting either classical density waves or transient spiral arms.

The paper is organised as follows. In Sect. 2 we describe the OC sample that we use throughout the analysis. In Sect. 3 we study the spatial distribution of the reported OCs, particularly the youngest ones, and derive the present-day spiral arms structure. In Sect. 4 we use the astrophysical information of OCs (*i.e.* phase-space coordinates and ages) to test the density wave nature of the spiral arms, by computing the spiral pattern speed for the Perseus, Local, Sagittarius and Scutum spiral arm segments. In Sect. 5 we explore the imprints left in the age-distribution of the open clusters across the spiral arms, by the diferences in the rotational velocity of the stars and the arms. The discussion on the results obtained is done in Sect. 6, and the conclusions are found in Sect. 7.

2. The open cluster sample

The data used throughout the paper are those from the OCs identified in the *Gaia* DR2 data (Gaia Collaboration et al. 2018a), with their mean astrometric values updated with *Gaia* EDR3 measurements (Gaia Collaboration et al. 2020). The use of OCs allows us to have better constrained parameters than using field stars. The parameters needed for our methodology are the mean astrometric parameters, *i.e.* $(l, b, \mu_{\alpha^*}, \mu_\delta)$; the astrophysical parameters derived from *Gaia* astrometry and photometry, *i.e.* OC age, distance and line-of-sight extinction (Cantat-Gaudin et al. 2020); and radial velocity measurements for each OC (Tarricq et al. 2020).

2.1. Gaia EDR3 astrometry

The sample of OCs used in this work includes those known previous to (and confirmed by) *Gaia* DR2 (Cantat-Gaudin et al. 2018). Additionally, we have included the large number of clusters which have been found in *Gaia* DR2 data (*e.g.* Castro-Ginard et al. 2018, 2019, 2020; Sim et al. 2019; Liu & Pang 2019). We updated the OC mean astrometric parameters with the *Gaia* EDR3 astrometric information, which has a greater precision in its measurements given the longer time baseline for the observations. In total there are 2017 OCs in these catalogues. For those clusters we use the sky coordinates and proper motions for the centre of the OC $(l, b, \mu_{\alpha^*}, \mu_\delta)$, which are computed from its member stars. The uncertainties in the position (l, b) can be neglected. The uncertainties in the mean proper motions are below 0.2 mas yr^{-1}, generally around 0.1 mas yr^{-1}.

2.2. Age, distance and line-of-sight extinction

Cantat-Gaudin et al. (2020) published a catalogue of ages, distances and line-of-sight extinctions for the OCs known to date. They used an artificial neural network to infer these parameters for each OC from its color-magnitude diagram in the *Gaia* passbands (G, G_{BP}, G_{RP}) and parallax information (ϖ). The authors were able to compute these astrophysical parameters for 1878 of the 2017 OCs reported. For the 139 others, the OC had too few members or its CMD was too red, and no reliable estimation could be obtained using *Gaia* data alone.

The uncertainties in those quantities depend on the number of cluster members (see Sect. 3.4 of Cantat-Gaudin et al. 2020, for details). We took the one sigma uncertainty for the age as $\sigma_{\log t} \in [0.15, 0.25]$ dex, which are the values for the young OCs uncertainties recommended by Cantat-Gaudin et al. (2020). The uncertainty on the distance modulus is within 0.1 to 0.2 mag, corresponding to a 5%-10% distance uncertainty. Cantat-Gaudin et al. (2020) reported that they found no systematics effects on the determination of the parameters with respect to the literature.

2.3. Radial velocities

Radial velocities used in this work are those compiled by Tarricq et al. (2020). The authors crossmatched the OC members with several radial velocity catalogs. In addition to *Gaia*-RVS (Katz et al. 2019), they used data from ground-based large spectroscopic surveys: the latest public version of the *Gaia*-ESO survey (Randich et al. 2013), APOGEE DR16 (Ahumada et al. 2020), RAVE DR6 (Steinmetz et al. 2020), GALAH DR2 (Buder et al. 2018; Zwitter et al. 2018). They also included data from other radial velocity catalogues: Nordström et al. (2004), Mermilliod et al. (2008, 2009), Worley et al. (2012), the OCCASO survey (Casamiquela et al. 2016) and Soubiran et al. (2018).

This radial velocity catalogue consists of 1382 clusters, 1315 of them with astrophysical parameters estimated by Cantat-Gaudin et al. (2020), with a median uncertainty on the weighted mean radial velocity of 1.13 km s^{-1}, based on more than 10 stars for 18% of the sample and on at least 3 stars for the 50%. We use the derived mean radial velocity per cluster.

2.4. Final OC sample

Figure 1 shows the distribution of ages of the two subsamples of OCs, the one with age, distance and line-of-sight extinction determination is represented as a solid black line; whilst the grey bars represent the clusters with radial velocity measurements. The red dash-dotted lines correspond to ages equal to 10 and 80 Myr, and they show the subset of OCs that we use to compute the spiral pattern speed in Sect. 4.

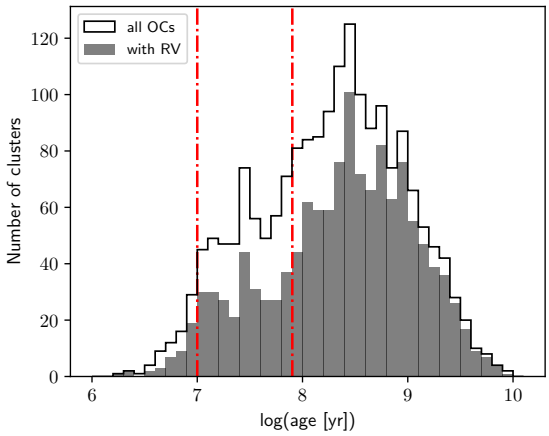


Fig. 1: Histogram of OC ages. Solid line shows OCs with age estimation available whilst solid bars show the subset of OCs with radial velocity measurements. Red dash-dotted vertical lines correspond to 10 and 80 Myr.

3. Present-day OC spatial distribution

The spiral pattern is clearly seen in a heliocentric X-Y projection of the OCs with ages younger than 150 Myr (see Fig. 8 and Fig. 1 of Cantat-Gaudin et al. 2020; Kounkel et al. 2020, respectively) while for older age bins this pattern disappears. Further dividing this 0-150 Myr range in four bins, we are able to spot the OCs overdensities corresponding to the spiral arms in a range of ages. This traces the evolution of the spiral pattern during the time interval where the overdensities are seen (see Sect. 4.2). Figure 2 shows how spiral arm segments in the Solar neighbourhood are clearly seen in OC overdensities for the youngest age interval explored (0-30 Myr), and how these overdensities show an increasing dispersion with time, so a slow dilution. The black shaded regions represent the spiral arms as modeled by Reid et al. (2014).

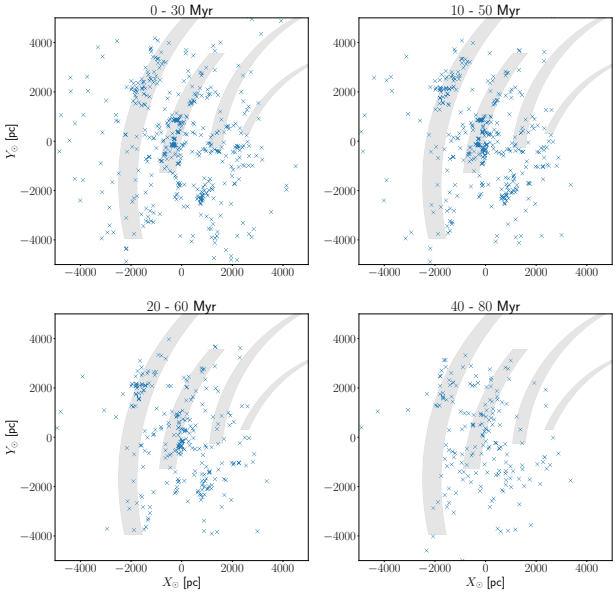


Fig. 2: Distribution of OCs in heliocentric X - Y, in different age bins. The Galactic centre is towards positive X values, and the direction of the Galactic rotation is towards positive Y values. The OCs correspond from left to right and from top to bottom to: ages less than 30 Myr, from 10 to 50 Myr, from 20 to 60 Myr and from 40 to 80 Myr. The spiral arms defined by Reid et al. (2014) are overplotted.

3.1. Re-determination of current spiral arms

We re-determine the parameters of the present-day spiral arms by using the hypothesis that OCs are born in spiral arms (Roberts 1969), and that the youngest OCs ($\lesssim$ 30Myr) have not moved far from their birth places (Dias & Lépine 2005). Thus, considering the usual log-periodic spiral arms, each arm should be detected as an overdensity following the relation used by Reid et al. (2014)

$$\ln \frac{R_G}{R_{G,ref}} = -(\theta_G - \theta_{G,ref}) \tan \psi, \tag{1}$$

where R_G and θ_G are Galactocentric radius and azimuth along the arm, respectively. And $R_{G,ref}$, $\theta_{G,ref}$ (taken to be near the median value of θ_G) and ψ are a reference Galactocentric radius and azimuth, and the pitch angle for a given arm. The Galactocentric azimuth is taken to be $\theta_G = 0$ on the Sun-Galactic centre line, and growing towards the Galactic rotation direction.

We detect the overdensities using a Gaussian mixture model (GMM) in the $(\ln R_G, \theta_G)$ space. A GMM is able to describe all the points in the parameter space as a weighted sum of Gaussians. This representation of our sample allows us to describe each arm, expected to follow Eq. 1 with some dispersion, as

a Gaussian along that direction (straight line in the $(\ln R_G, \theta_G)$ space). The number of Gaussians to fit is automatically selected using the Bayesian information criterion (BIC).

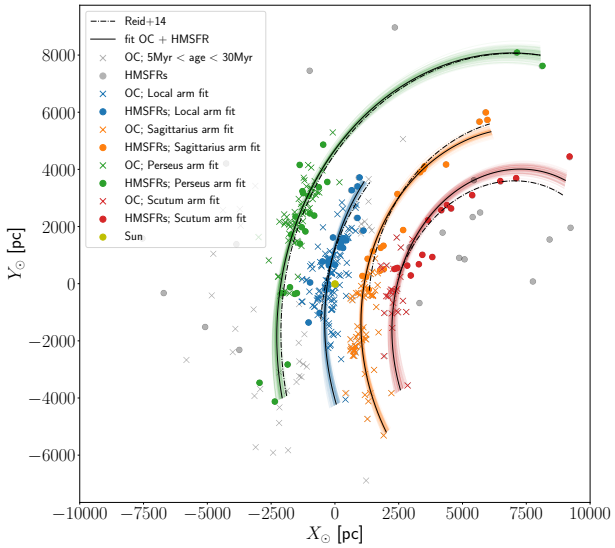


Fig. 3: Heliocentric X - Y distribution of OCs (crosses) younger than 30 Myr and HMSFRs (dots) from Reid et al. (2014), used to fit the spiral arms. Different colors correspond to different arms. The assignments to each arm is computed using a Gaussian Mixture Model. Solid black lines are the fitted spiral arms with the parameters in Table 1, while shaded regions account for 1σ uncertainties. Dash-dotted lines correspond to the spiral arms defined by HMSFRs only. The Galactic centre is towards positive X and the Galactic rotation direction is towards positive Y.

Once the Gaussian field is obtained, we select the Gaussian components with the four highest weights, corresponding to the four arm segments. We find 56, 121, 61 and 26 OCs younger than 30 Myr assigned to the Perseus, Local, Sagittarius and Scutum arms, respectively. To increase the number of spiral arm tracers, and be able to trace the arm in a wider range of Galactic azimuth, we include the data from Reid et al. (2014) used to fit the spiral arms. These data correspond to 103 high-mass star forming regions (HMSFRs) with parallax and proper motion measurements obtained using Very Long Base Interferometry (VLBI) techniques, 84 of which are assigned to one of the four explored spiral arms. In order to obtain the parameters for each arm, we fit Eq. 1 to the OCs and HMSFRs assigned to each arm by the minimum least squares method (348 tracers in total, 264 OCs and 84 HMSFRs). The parameters obtained for each arm are listed in Table 1.

Figure 3 shows the representation of the spiral arms defined by the OCs and HMSFRs. The black solid lines correspond to the best fit value for each arm, and the black shaded regions correspond to the 1σ uncertainty taken into account the correlations among the estimated parameters. Our all-sky OC sample provides a good complement to the ground-based observations used by Reid et al. (2014), who do not cover the fourth quadrant. By increasing the number of total tracers by factor of 4 we can better constrain the estimation of the mean Galactocentric radius (R_G) and pitch angle (ψ) finding lower uncertainties in these values, as well as increasing the θ_G range where the arms are defined. The spiral arms defined by Reid et al. (2014) using HMSFRs are shown in dash-dotted lines to compare with our definition using both young OCs and HMSFRs.

Table 1: Fitted parameters, including statistical errors, for present-day spiral arms.

Arm	$N_{\rm tracers}$ OC+HMSFR	$\theta_{G,\rm ref}$ [deg]	$\theta_{G,\rm range}$ [deg]	$R_{G,\rm ref}$ [kpc]	ψ [deg]
Perseus	56 + 24	−13.0	(−20.9, 88.2)	10.88 ± 0.04	9.8 ± 0.9
Local	121 + 25	−2.3	(−26.9, 26.6)	8.69 ± 0.01	8.9 ± 1.3
Sagittarius	61 + 18	3.5	(−39.3, 67.7)	7.10 ± 0.01	10.6 ± 0.8
Scutum	26 + 17	−4.8	(−32.7, 100.9)	6.02 ± 0.02	14.9 ± 1.6

4. Spiral Pattern Speed

The spiral pattern speed is indicative of the nature of the spiral arms. As described in Gerhard (2011), the most direct way to estimate the pattern speed of the spiral arms is through the OC population due to the robustness with which their parameters can be estimated, by averaging over their members. With the assumption that the OCs are born in spiral arms (Roberts 1969), and integrating backwards the present OCs position to their birthplaces, it is possible to compute the rotation rate at which a spiral arm has moved to reach its present-day position.

We compute the birthplace of the OCs by integrating backwards in time following each OC orbit. The orbits are integrated following a gravitational potential composed by a spherical nucleus and bulge, a Navarro-Frenk-White dark matter halo and a Miyamoto-Nagai disc, where its parameters have been adapted to follow the observed rotation curve of the Milky Way (Bovy 2015). The numerical processing is done using the Python package `GALA` (Price-Whelan 2017), which uses a Leapfrog integration scheme to trace back the orbits in time steps of 0.1 Myr. The determination of the uncertainties on the birth position is done via Monte Carlo sampling from the uncertainties on the age of each OC, which is the biggest source of error in our case (see description of the OC sample in Sect. 2).

It is important to note that the birth position of each OC reveals the location of a spiral arm at a time equals to the birth time of the OC. Similar to the method described in Dias & Lépine (2005), the arm at this previous epoch is rotated forward with a rotational velocity equal to the pattern speed of that arm, Ω_p, during a time equal to the age of the OC. We consider each arm to have a unique pattern speed, which is constant during the whole time interval considered, and free to be different from other arms pattern speed.

The procedure to compute the pattern speed Ω_p that best describes the data for each arm is as follows:

- Detect overdensities that correspond to the Perseus, Local, Sagittarius and Scutum spiral arms (see Sect. 3.1). Study each arm separately.
- Integrate backwards each OC orbit to find their birthplaces. These OC birthplaces represent the location of the spiral arm at the time the OC was born.
- Rotate past location of the arm with a circular motion at a given pattern speed (Ω_p) during the age of the cluster (t) to find its expected present-day location, i.e. $\theta_{G,\rm now} = \theta_{G,\rm birth} + \Omega_p * t$.
- Iterate over Ω_p to find the optimal value by minimising the distance of the recovered present-day locations of the spiral arms to their analytical present-day description.
- Repeat procedure for 1 000 Monte Carlo realisations to account for the uncertainties in the birthplace of the OC.

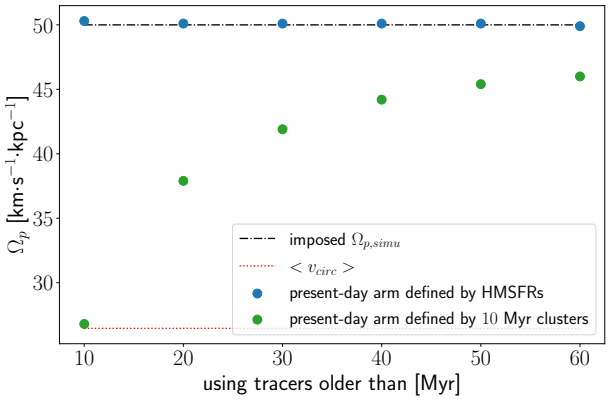
– For each arm, report the best value for Ω_p as the mean value of all the pattern speeds obtained, with the standard deviation as its dispersion.

4.1. Test simulation

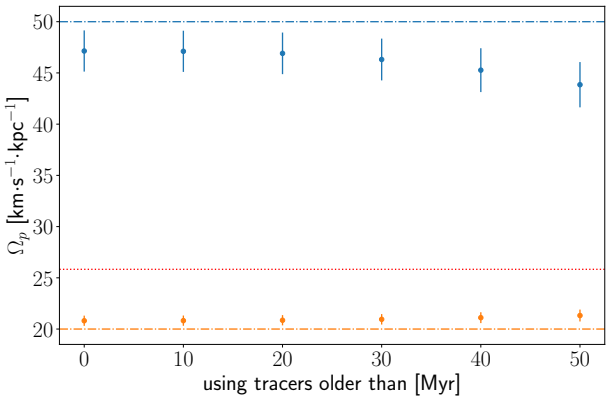
To test our ability to recover the pattern speed, we set up a basic simulation following the evolution of both a density wave spiral pattern and the objects born in it. We generate a log-periodic spiral arm with the parameters taken from Reid et al. (2014, Table 2, Local arm), which we take to be the present-day position. In this case, the actual spiral arms are described using the parameters found by Reid et al. (2014), and not our own estimation (Table 1); this is because we want to keep the position of the spiral arm and its velocity to be defined by independent tracers (by HMSFR and OCs, respectively). We rotate backwards the arm keeping its shape parameters unchanged, at a constant pattern speed which we assume. At times $T = 10, 20, 30, 40, 50$ and 60 Myr, a set of simulated clusters (equivalent to OCs) is generated, which are used as tracers of the spiral pattern speed.

Firstly, we test the effect of the definition of the present-day spiral arm on the determination of its pattern speed, in the ideal case of the OCs moving with circular orbits (Test 1). We let the simulated clusters with different ages evolve to their present-day position with circular orbits using the Milky Way rotation curve from Bovy (2015). In this case, the spiral pattern speed is fixed at 50 km s^{-1}kpc^{-1}, while the mean circular velocity of the simulated clusters is 26.37 ± 1.91 km s^{-1}kpc^{-1}. Figure 4a shows the capabilities of the method to compute the spiral pattern speed when the present-day spiral arm is defined by i) independent means (HMSFR), or ii) the youngest simulated clusters. In the first case (blue dots), the imposed value for the spiral pattern speed (black dash-dotted line) is always recovered. In the second case (green dots), we recover the value for the circular velocity (red dotted line) when the exact same tracers are used to define the present-day spiral arm and its pattern speed, but we can asymptotically approach the true value when older tracers are considered for the spiral pattern speed computation. From Test 1, we learn that the tracers to define the present-day spiral arms should be independent of the tracers used to compute their pattern speed. This is achieved by defining the present-day spiral arms using the HMSFRs reported in Reid et al. (2014), which are younger than 10 Myr, and using OCs older than 10 Myr to compute the spiral pattern speed.

Secondly, we test if the methodology is able to distinguish two different spiral pattern speeds in a more realistic situation (Test 2). The simulated clusters are now evolved using circular velocities and non-zero peculiar velocities, which are drawn from a Gaussian distribution $\mathcal{N}(0, 5)$ km s^{-1}. We also add errors to the age of the simulated clusters, consistent with those in our catalogues (see. Sect 2.2). The method is run for spiral pattern speeds of 20 and 50 km s^{-1}kpc^{-1}, and assuming that the present-day spiral arm is known from independent tracers, *i.e.* HMSFRs. In Fig. 4b we show the obtained values for the two different experiments. We can recover the imposed pattern speed with a systematic error that ranges from 0.8 to 6 km s^{-1}kpc^{-1} for these different cases of Ω_p. From Test 2, we see that even though the accuracy is not enough to recover the exact individual pattern speeds, the methodology is accurate enough to differentiate between the two scenarios, the 20 and the 50 km s^{-1}kpc^{-1} cases.



(a) Test 1. Imposed $\Omega_p = 50$ km s^{-1}kpc^{-1}. Blue dots show the recovered Ω_p value when considering the present-day spiral arm defined by the HMSFRs, while green dots represent the recovered value when the present-day arm is defined with 10 Myr clusters.



(b) Test 2. Pattern speed obtained for the cases of $\Omega_p = 20$ and 50 km s^{-1}kpc^{-1}, orange and blue, respectively. Dots show the recovered value, including errorbars.

Fig. 4: Recovered Ω_p for two tests. The y-axis represents the value of the pattern speed, and the x-axis represents the minimum age of the stellar objects considered for the computation, *i.e.* for $x = 10$ we consider objects with age ≥ 10 Myr. The dash-dotted lines represent the imposed value and dots represent the recovered value. Dotted red line shows the mean circular velocity of the stellar objects.

4.2. Estimation of Ω_p

As seen in the test simulations, the success of the methodology relies on the ability to define the present-day spiral structure by tracers other than OCs, which are used to trace the spiral pattern speed. As already mentioned, we consider that this is achieved by describing the present-day spiral structure with the HMSFRs reported in Reid et al. (2014), and therefore we use this definition of the present-day spiral arms to compute Ω_p. The authors provide the parameters from a fitting using 84 HMSFRs, younger than 10 Myr, with parallax and proper motion measurements from VLBI, as said in Sect. 3.1. This information is available for the Perseus, Local, Sagittarius and Scutum arms.

Once the present-day spiral arms are defined, we have to find the present position of the OCs born in each of these arms. In the density wave theory, the OCs may have evolved differently from

Table 2: Pattern speed (in $\mathrm{km\,s^{-1}kpc^{-1}}$) obtained for different age bins for each spiral arm analysed.

Arm	Ω_p (10 - 50 Myr)	Ω_p (20 - 60 Myr)	Ω_p (40 - 80 Myr)	Ω_p (10 - 80 Myr)
Perseus	20.08 ± 3.35	21.69 ± 2.89	25.56 ± 2.63	24.41 ± 2.05
Local	35.97 ± 0.93	37.37 ± 1.12	34.42 ± 1.25	34.79 ± 1.13
Sagittarius	32.39 ± 4.24	36.78 ± 3.40	30.25 ± 1.82	30.81 ± 1.72
Scutum	49.93 ± 2.17	48.37 ± 2.63	46.93 ± 2.45	47.26 ± 1.99

their mother spiral arms, *i.e* OCs move at a velocity approximately given by the Milky Way rotation curve while the spiral pattern moves at Ω_p. Even though the evolution follows different paths, an overdensity of very young OCs in (R_G, θ_G) will come from the same arm (see Fig. 2). As described in Sect. 3.1, the OCs belonging to each of the arms are seleted using a GMM in the $(\ln R_G, \theta_G)$ space.

We apply the described methodology to the OCs younger than 80 Myr, for different age ranges to account for the effects seen in Test 1 of Sect. 4.1. Table 2 shows the computed spiral pattern speed for the four spiral arms explored, and they show a similar trend as in the test simulation scenario. Following the same argument, we can say that the methodology is good enough to distinguish among different true spiral pattern speeds.

The recovered values for Ω_p, shown in Fig. 5, are decreasing as the Galactocentric reference radius ($R_{G,ref}$) of the spiral arm increases. The Ω_p for the explored arms follow the Galactic rotation curve which is represented by the dotted line, except for the case of the Local arm. This can be related to the fact that the Local arm is not considered a long arm but a small armlet instead, however this deserves further study. Our results are in agreement with the findings of Quillen et al. (2018), who estimated the spiral pattern speeds for different spiral features to explain the arcs and ridges seen in the velocity distribution of the Solar neighbourhood in *Gaia* DR2. The authors studied how the orbits of known moving groups could be perturbed by the presence of a spiral arm, and found that a spiral arm segment in the outer disc located at $\sim$ 2 kpc from the Sun, with a pattern speed of 20 ± 3 $\mathrm{km\,s^{-1}kpc^{-1}}$ could be responsible for the outer boundary of the Sirius/UMa moving group. This finding is in perfect agreement with the spiral speed of the Perseus arm segment we computed using OCs as the main tracers, with a $\Omega_p = 20.11 \pm 3.35$ $\mathrm{km\,s^{-1}kpc^{-1}}$ at a Galactocentric radius of 10.88 kpc. Our results for the rest of the spiral arm segments explored are in a similar agreement (see Table 1 from Quillen et al. 2018), also for the case of the Local arm where the authors found a pattern speed higher than the angular velocity from the rotation curve. The decreasing spiral pattern speed with Galactocentric radius, with spiral arms nearly co-rotating with Galactic rotation, as expected if the spiral arms are short-lived transient structures (Grand et al. 2012; Kawata et al. 2014).

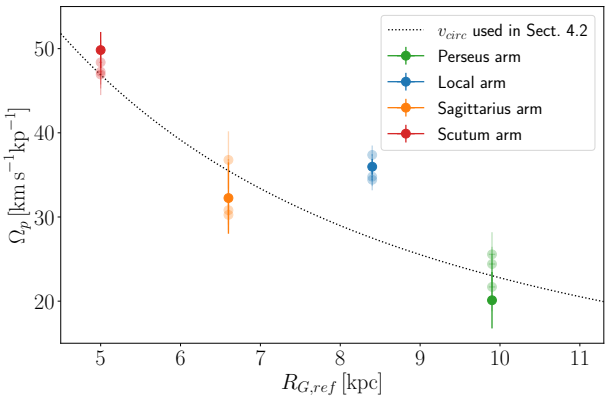


Fig. 5: Computed spiral pattern speed for the Scutum, Sagittarius, Local and Perseus arms (from left to right). Solid dots show the first column in Table 2, corresponding to 10 - 50 Myr interval. The transparent dots are for the rest of the columns. The dotted line shows the circular velocity from the Milky Way rotation curve. The estimated Ω_p values show a decreasing trend with Galactocentric radius.

5. Cluster ages across the spiral arms

The analysis of the distribution of the OCs as a function of age across a given arm can provide clues on the nature of the spiral arms (Dobbs & Baba 2014), and therefore it offers an independent approach to support the findings of the previous section. As studied by Dobbs & Pringle (2010), the differences in the rotational velocity of the stellar distribution and the spiral arms, lead to different distributions of the OCs across the present-day spiral arms. Such distributions depend on the spiral arm formation mechanisms. The authors considered a set of four simulations where the spiral structure has been excited by different possible mechanisms, i) a global density wave, ii) a central rotating bar, iii) flocculent spiral or iv) tidally induced arms; and discussed how would be the age distribution of the clusters across a given spiral arm in each of the explored cases. A density wave and/or bar induced spiral arms yield a trend in age across the arms. Flocculent or tidally induced mechanisms yield several individual peaks across the arm, with no age-gradient. This age gradient is due to the difference in velocity with which the different structures (spiral pattern and clusters) are moving, while in the density wave or bar induced spirals scenario the spiral pattern moves with a fixed, constant pattern speed, the clusters move following the galactic rotation curve. That results in older (younger) clusters leading the spiral arm, if the clusters are inside (outside) the co-rotation radius. In the opposite case, for the flocculent and tidally induced arms where the spiral pattern and the stars move at roughly the same speed, the section of the arm contains clusters of different ages with no clear gradient across the arm.

In Fig. 6, we show a plot reproducing Fig. 4 of Dobbs & Pringle (2010) but using our OCs sample. The different panels show histograms of the number of OCs, for different cluster ages, across a circular section (500 pc wide) located at distances of 10 kpc, 8.3 kpc and 7 kpc from the Galactic centre, *i.e.* approximately along the Perseus, Local and Sagittarius arm, respectively. None of the arms shows the aforementioned age gradient, which should be clearer as we move away from a hypothetical co-rotation radius. This indicates that the velocity of the clusters, *i.e.* the stellar distribution velocity, is very similar that the rotation velocity of the spiral arms, therefore co-rotating with them. We have explored different sizes of the age bins reaching the same conclusion in all cases. The non-presence of the age gradi-

ent favours the flocculent spirals or the external tidal interaction, where spiral arms tend to be transient, as the mechanisms for the excitation of the spiral structure.

The completeness of the OC population may play a role in the interpretation of Fig. 6. Castro-Ginard et al. (2020) tested how many known (prior to *Gaia*) OCs could be recovered using their detection algorithm. This recovery fraction was then used by Anders et al. (2020) to estimate the completeness of the OC population as a function of age, finding that the recovered fraction of OCs is $\lesssim 60\%$ for the very young OCs (in the range of 1-10 Myr), and reaching $\gtrsim 90\%$ for older OCs. Therefore, even if the fraction of youngest population in Fig. 6 may be underestimated, the older populations (defining the age gradient) are nearly complete, reaching the same conclusions of no age gradient at all.

6. Discussion

The nature of the spiral arms has been studied in other galaxies taking advantage of our external point of view. Shabani et al. (2018) studied the distribution of stellar clusters across the spiral arms in NGC 1566, M 51 and NGC 628. They find an age gradient across the arm only in NGC 1566 (a grand design spiral galaxy with a strong bar), which is compatible with the density wave scenario (Dobbs & Pringle 2010). For the case of M 51, the spiral structure is excited by the tidal interactions with its companion, and for NGC 628 the spiral arms are consistent with a pattern speed decreasing with radius, both leading to a transient spiral nature. Also in external galaxies, the measurements of spiral pattern speeds that vary as a function of radius (Meidt et al. 2008; Speights & Westpfahl 2012) or the evolution of the spiral arms pitch angle (Pringle & Dobbs 2019), support a transient nature for their spiral structure. This transient behaviour of the spiral structure, with the arms co-rotating with disc stars, is also expected from N-body simulations for unbarred galaxies or galaxies with a weak bar (Roca-Fàbrega et al. 2013); while galaxies with a strong bar quickly develop a spiral pattern whose pattern speed is constant with radius, behaving as a global density wave as for the case of NGC 1566.

For the case of the Milky Way, the lack of a homogeneous OC catalogue (before *Gaia*) prevented from reaching a firm conclusion (Monguió et al. 2017). Thanks to the *Gaia* mission, the study of OCs has reached a maturity, in terms of purity and homogeneity of the catalogue, and robustness of its estimated parameters, that allows us to apply different approaches to revisite the spiral nature of the Milky Way.

We have explored the nature of the spiral structure of the Milky Way by comparing the angular velocities in which the stellar distribution and the spiral pattern move. The spiral arms should move with a global constant pattern speed in the density wave scenario, regardless of the Galactocentric reference radius of the arm. This is not what we deduce from our sample of young OCs as main tracers. We see that different spiral arm segments move with a different angular velocity, which tend to decrease with their Galactocentric reference radius. This behaviour is related to a short-lived transient spiral structure.

The procedure applied in this work to compute the spiral pattern speed uses the hypothesis of spiral arms with a constant shape during the time interval explored. Our tests using simulations show that the methodology is accurate enough to discard a unique pattern speed for all the spiral arms studied. This is in contrast with the work done by Dias et al. (2019) who, using the same methodology, reported a spiral pattern speed of $\Omega_p = 28.2 \pm 2.1$ km s^{-1}kpc^{-1}, common for all the explored spiral arms. Here, the inclusion of hundreds of newly discovered OCs (Castro-Ginard et al. 2020), with an updated estimation of ages, distances and line-of-sight extinctions for the whole OC sample (Cantat-Gaudin et al. 2020) and the addition of radial velocities for a large fraction of them (Tarricq et al. 2020), together with a robust statistical treatment, allows us to distiguish among different true pattern speeds for different spiral arms (Fig. 5, Table 2).

The effects that may change the shape of the spiral arms (*e.g.* the shear of the Galactic disc or the evolution of the pitch angle) are not included in the assumptions of this work, nor in the works using similar procedures (Dias & Lépine 2005; Junqueira et al. 2015; Dias et al. 2019). However, if we consider that these effects are small over the course of $\sim$ 50 Myr, the values obtained for the spiral pattern speeds suggest that spiral arms are structures that co-rotate with stars at any radii, revealing a transient nature of these arms (Grand et al. 2012). Therefore, our results with OCs agree with other works dealing with the kinematic substructure in the solar neighborhood. These works, some including simulations, tend to explain the kinematics of moving groups, or features in the action-angle space, with a transient behaviour of the Galactic spiral arms (Quillen et al. 2018, 2020; Hunt et al. 2018; Sellwood et al. 2019).

In addition, we have explored the imprint in the age-distribution of the OCs across the spiral arms, and we do not see the predicted age gradient of density wave or bar-driven spiral arms (Dobbs & Pringle 2010), even when the effects due to the incompleteness of our OC sample are taken into account. The combination of both results allow us to favour a flocculent Milky Way with transient spiral arms, disfavouring the density wave scenario with a grand design morphology. This idea of a flocculent Milky Way was already studied by Quillen (2002), who found multiple spiral features, each with a different pattern speed which is decreasing with Galactocentric radius. From the morphology of the spiral arms, a flocculent Milky Way was favoured by Xu et al. (2016) due to a long Local arm located between the Perseus and Sagittarius arms that would not be explained by a density wave theory with a pure grand design morphology. N-body simulations are also in agreement with density waves not explaining the spiral structure in our Galaxy (Honig & Reid 2015).

7. Conclusions

We have analysed the OC population with *Gaia* EDR3 astrometric parameters, radial velocities compiled from different surveys, and astrophysical parameters computed from *Gaia* DR2 astrometry and photometry, in order to derive the structure of the spiral arms in the Solar neighbourhood and to discriminate among several hypothesis about their nature.

We show that each of the four investigated arms exhibits a different pattern speed. Using a combination of statistical and data mining techniques, we find that each spiral arm has a spiral pattern speed which tend to decrease with Galactocentric radius, following the Galactic rotation curve, favouring a transient behaviour for these arms.

We analyse the age-ditribution of the OC population across the spiral arms to see the imprint of the different angular velocities of the stellar distribution and the spiral arm segments, if any. We see no indication of the age gradient predicted by Dobbs & Pringle (2010) to be a sign of a density wave-like footprint, thus favouring a flocculent Milky Way.

These two independent experiments, based on the most complete OC sample to date, allow us to disfavour the density wave theory of spiral structure and point towards a transient nature of

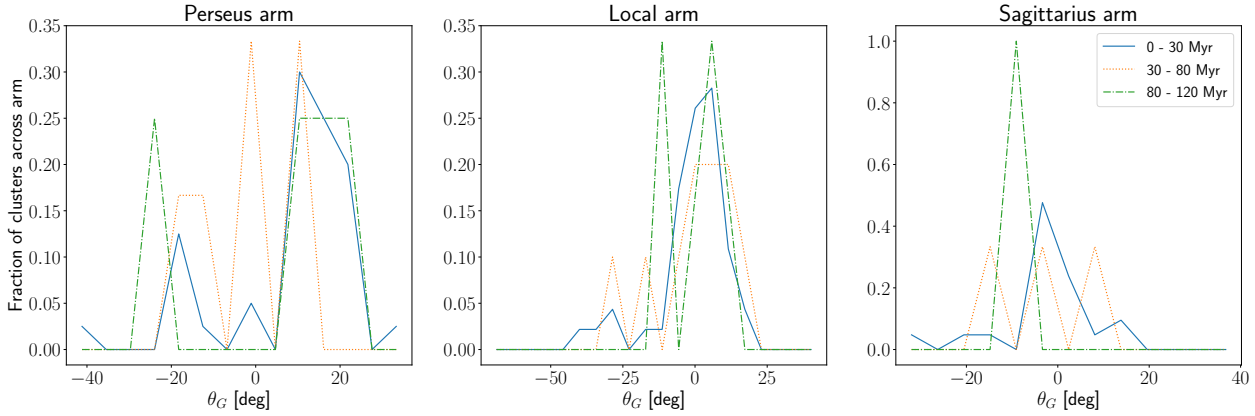


Fig. 6: Fraction of OCs for different age bins across the Perseus (left), Local (middle) and Sagittarius (right) arms. The x-axis represents the Galactic azimuth (θ_G) explored for a circular section centred at the Galactic centre, at a distance of 10 (left), 8.3 (middle) and 7 (right) kpc. The y-axis gives the number of OCs in each θ_G bin over the total number of OCs in the whole circular section. Solid blue lines, dotted orange lines and dash-dotted green lines correspond to clusters in 0-30 Myr, 30-80 Myr and 80-120 Myr age bins.

the spiral arms. This behaviour is seen here for the first time using OCs data, due to the increase in the OC sample with radial velocities available and better estimation of ages and distances. This points towards the same direction as the conclusions obtained by other authors by comparing *Gaia* DR2 kinematic information in the solar neighbourhood with simulations including different kinds of spiral arms, representing an agreement on the results using these two different (complementary) datasets.

Given the transient nature of the spiral arms proposed here, where the stellar objects co-rotate with the arm at any radius, we can increase the number of tracers of these spiral arms by adding the youngest OCs ($\leq$ 30 Myr) to the HMSFRs used to define the present-day arms. As a result, we increase by 314% the number of tracers (adding 264 OCs to the 84 HMSFRs used in Reid et al. 2014), and report updated parameters for the Perseus, Local, Sagittarius and Scutum spiral arms, spanning a wider range in Galactic azimuth (Table 1).

Acknowledgements. ACG thanks Dr. Lennart Lindegren, Dr. Teresa Antoja, Dr. Francesca Figueras and Dr. Maria Monguió for their useful suggestions. ACG also thanks Dr. Louise Howes for her comments on the writing. This work has made use of results from the European Space Agency (ESA) space mission *Gaia*, the data from which were processed by the *Gaia Data Processing and Analysis Consortium* (DPAC). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement. The *Gaia* mission website is http://www.cosmos.esa.int/gaia. The authors are current or past members of the ESA *Gaia* mission team and of the *Gaia* DPAC. This work was (partially) supported by the Spanish Ministry of Science, Innovation and University (MICIU/FEDER, UE) through grants RTI2018-095076-B-C21, ESP2016-80079-C2-1-R, and the Institute of Cosmos Sciences University of Barcelona (ICCUB, Unidad de Excelencia 'María de Maeztu') through grants MDM-2014-0369 and CEX2019-000918-M. ACG acknowledges Spanish Ministry FPI fellowship n. BES-2016-078499. PM gratefully acknowledges support from a research project grant from the Swedish Research Council (Vetenskaprådet). FA is grateful for funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 800502 H2020-MSCA-IF-EF-2017. This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement H2020-MSCA-COFUND-2016-754433. This work has been supported by the Spanish Government (SEV2015-0493), by the Spanish Ministry of Science and Innovation (contract TIN2015-65316-P), by Generalitat de Catalunya (contract 2014-SGR-1051). This research has made use of the VizieR catalogue access tool, CDS, Strasbourg, France. The original description of the VizieR service was published in A&AS 143, 23.

References

Ahumada, R., Allende Prieto, C., Almeida, A., et al. 2020, ApJS, 249, 3
Anders, F., Cantat-Gaudin, T., Quadrino-Lodoso, I., et al. 2020, arXiv e-prints, arXiv:2006.01690
Anders, F., Chiappini, C., Santiago, B. X., et al. 2014, A&A, 564, A115
Antoja, T., Helmi, A., Romero-Gómez, M., et al. 2018, Nature, 561, 360
Bovy, J. 2015, ApJS, 216, 29
Buder, S., Asplund, M., Duong, L., et al. 2018, MNRAS, 478, 4513
Cantat-Gaudin, T., Anders, F., Castro-Ginard, A., et al. 2020, A&A, 640, A1
Cantat-Gaudin, T., Jordi, C., Vallenari, A., et al. 2018, A&A, 618, A93
Casamiquela, L., Carrera, R., Jordi, C., et al. 2016, MNRAS, 458, 3150
Castro-Ginard, A., Jordi, C., Luri, X., et al. 2020, A&A, 635, A45
Castro-Ginard, A., Jordi, C., Luri, X., Cantat-Gaudin, T., & Balaguer-Núñez, L. 2019, A&A, 627, A35
Castro-Ginard, A., Jordi, C., Luri, X., et al. 2018, A&A, 618, A59
Dias, W. S. & Lépine, J. R. D. 2005, ApJ, 629, 825
Dias, W. S., Monteiro, H., Lépine, J. R. D., & Barros, D. A. 2019, MNRAS, 486, 5726
Dobbs, C. & Baba, J. 2014, PASA, 31, e035
Dobbs, C. L. & Pringle, J. E. 2010, MNRAS, 409, 396
Drimmel, R. & Spergel, D. N. 2001, ApJ, 556, 181
Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2018a, A&A, 616, A1
Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2020, arXiv e-prints, arXiv:2012.01533
Gaia Collaboration, Katz, D., Antoja, T., et al. 2018b, A&A, 616, A11
Gerhard, O. 2011, Memorie della Societa Astronomica Italiana Supplementi, 18, 185
Grand, R. J. J., Kawata, D., & Cropper, M. 2012, MNRAS, 421, 1529
Honig, Z. N. & Reid, M. J. 2015, ApJ, 800, 53
Hunt, J. A. S., Hong, J., Bovy, J., Kawata, D., & Grand, R. J. J. 2018, MNRAS, 481, 3794
Hunt, J. A. S., Johnston, K. V., Pettitt, A. R., et al. 2020, MNRAS, 497, 818
Junqueira, T. C., Chiappini, C., Lépine, J. R. D., Minchev, I., & Santiago, B. X. 2015, MNRAS, 449, 2336
Kamdar, H., Conroy, C., Ting, Y.-S., & El-Badry, K. 2020, arXiv e-prints, arXiv:2007.10990
Katz, D., Sartoretti, P., Cropper, M., et al. 2019, A&A, 622, A205
Kawata, D., Hunt, J. A. S., Grand, R. J. J., Pasetto, S., & Cropper, M. 2014, MNRAS, 443, 2757
Kounkel, M., Covey, K., & Stassun, K. G. 2020, arXiv e-prints, arXiv:2004.07261
Lada, C. J. & Lada, E. A. 2003, ARA&A, 41, 57
Lin, C. C. & Shu, F. H. 1964, ApJ, 140, 646
Liu, L. & Pang, X. 2019, ApJS, 245, 32
Meidt, S. E., Rand, R. J., Merrifield, M. R., Shetty, R., & Vogel, S. N. 2008, ApJ, 688, 224
Mermilliod, J. C., Mayor, M., & Udry, S. 2008, A&A, 485, 303
Mermilliod, J.-C., Mayor, M., & Udry, S. 2009, A&A, 498, 949
Michtchenko, T. A., Lépine, J. R. D., Pérez-Villegas, A., Vieira, R. S. S., & Barros, D. A. 2018, ApJ, 863, L37

Monguió, M., Grosbøl, P., & Figueras, F. 2015, A&A, 577, A142
Monguió, M., Negueruela, I., Marco, A., et al. 2017, MNRAS, 466, 3636
Naoz, S. & Shaviv, N. J. 2007, New A, 12, 410
Nordström, B., Mayor, M., Andersen, J., et al. 2004, A&A, 418, 989
Price-Whelan, A. M. 2017, The Journal of Open Source Software, 2
Pringle, J. E. & Dobbs, C. L. 2019, MNRAS, 490, 1470
Quillen, A. C. 2002, AJ, 124, 924
Quillen, A. C., Carrillo, I., Anders, F., et al. 2018, MNRAS, 480, 3132
Quillen, A. C., Dougherty, J., Bagley, M. B., Minchev, I., & Comparetta, J. 2011, MNRAS, 417, 762
Quillen, A. C., Pettitt, A. R., Chakrabarti, S., et al. 2020, arXiv e-prints, arXiv:2006.01723
Ramos, P., Antoja, T., & Figueras, F. 2018, A&A, 619, A72
Randich, S., Gilmore, G., & Gaia-ESO Consortium. 2013, The Messenger, 154, 47
Reid, M. J., Menten, K. M., Brunthaler, A., et al. 2014, ApJ, 783, 130
Roberts, W. W. 1969, ApJ, 158, 123
Roca-Fàbrega, S., Valenzuela, O., Figueras, F., et al. 2013, MNRAS, 432, 2878
Sellwood, J. A. & Carlberg, R. G. 2014, ApJ, 785, 137
Sellwood, J. A., Trick, W. H., Carlberg, R. G., Coronado, J., & Rix, H.-W. 2019, MNRAS, 484, 3154
Shabani, F., Grebel, E. K., Pasquali, A., et al. 2018, MNRAS, 478, 3590
Shu, F. H. 2016, ARA&A, 54, 667
Sim, G., Lee, S. H., Ann, H. B., & Kim, S. 2019, Journal of Korean Astronomical Society, 52, 145
Soubiran, C., Jasniewicz, G., Chemin, L., et al. 2018, A&A, 616, A7
Speights, J. C. & Westpfahl, D. J. 2012, ApJ, 752, 52
Steinmetz, M., Matijevič, G., Enke, H., et al. 2020, AJ, 160, 82
Tarricq, Y., Soubiran, C., Casamiquela, L., et al. 2020, arXiv e-prints, arXiv:2012.04017
Toomre, A. 1964, ApJ, 139, 1217
Wada, K., Baba, J., & Saitoh, T. R. 2011, ApJ, 735, 1
Worley, C. C., de Laverny, P., Recio-Blanco, A., et al. 2012, A&A, 542, A48
Xu, Y., Reid, M., Dame, T., et al. 2016, Science Advances, 2, e1600878
Zwitter, T., Kos, J., Chiavassa, A., et al. 2018, MNRAS, 481, 645

Part III

SUMMARY OF RESULTS, DISCUSSION AND CONCLUSIONS

6

SUMMARY OF RESULTS, DISCUSSIONS AND CONCLUSIONS

In this chapter, I present a summary of results and discussions. I also summarize the conclusions we reached and I outline how this work can be continued and extended in the future.

This thesis contributes to the understanding of the open cluster (OC) population in the Milky Way disc and the improvement of the catalogue of OCs. It also contains a strong methodological component, which sets the path for future searches of stellar substructure in Big Data catalogues such as *Gaia*. The work developed in this thesis can be separated in several parts summarised below.

First, I describe the development of a fully automated data mining methodology to blindly search for unknown OCs in the *Gaia* data. The method represents a complete machine learning pipeline, from the data preparation to the combination of unsupervised and supervised learning techniques that results in newly detected OCs. In a first step, the region of interest is scanned to find statistical overdensities in a five-dimensional astrometric parameter space, $(l, b, \varpi, \mu_{\alpha^*}, \mu_\delta)$. This is done using the `DBSCAN` method (Ester et al. 1996), a density based clustering algorithm which separates the data into different clusters based on the proximity of its member stars in the n-dimensional space. After this, these overdensities are classified into mere random statistical overdensities or real physical OCs using an ANN (Hinton 1989). The network is trained to recognise the isochrone pattern that stars which are members of an OC follow in a CMD.

Second, the newly found OCs, together with the already known ones, are used to shed light into the structure of the Galactic disc and how it has formed, particularly focusing in its spiral arms. Since spiral arms are known to be sites of star formation, I have related the youngest population of these OCs to the Milky Way spiral arms and provided an OC point of view of their dynamics. I have used OC radial velocities (Tarricq et al. 2020) and astrophysical parameters such as age, distance and line-of-sight extinction (Cantat-Gaudin et al. 2020b, see Appendix a), to discriminate as far as possible among different theories for spiral arms formation using two independent approaches: i) OC age distribution across spiral arms, and ii) direct computation of the spiral pattern speeds using OCs as their tracers.

THE OPEN CLUSTER CENSUS

The blind search for OCs in the Galactic disc, described as $|b| < 20^\circ$ and $G \leq 17$, from *Gaia* DR2 (Part i) has resulted in the discovery of more than 650 UBC clusters so far (UBC, named after the *University of Barcelona* Castro-Ginard et al. 2018, 2019, 2020). This represents the major single contribution to the OC catalogue based on *Gaia* data.

Previous to *Gaia*, the OC census counted with $\sim$ 3000 objects compiled from heterogeneous data sources (Dias et al. 2002; Kharchenko et al. 2013), which makes the exploitation of this catalogue a difficult task. With the release of *Gaia* DR2, this census was reduced to $\sim$ 1200 objects, where the rest were discarded since they were considered as asterisms or too distant or too reddened to be seen by *Gaia* (Cantat-Gaudin et al. 2018). From contributions such as the one presented in this thesis, the current OC census amounts $\sim$ 2000 objects, challenging some assumptions based on the previous catalogue.

The OC population known previous to *Gaia* was claimed to be complete up to 1.8 kpc. However, the 100% of the OCs found in Castro-Ginard et al. (2018, see Chapter 2) are closer than 1.8 kpc, with $\sim$ 87% of them even closer than 1 kpc (mainly due to the TGAS limiting magnitude of $G = 12$ Lindegren et al. 2016), demonstrating that such completeness was not the real case. This detection pattern is repeated in posterior searches, now using *Gaia* DR2 as the main data source, increasing the census in 18% up to 1 kpc and in 54% between 1 and 2 kpc (Castro-Ginard et al. 2020, see Chapter 4). This shows that our knowledge of the OC population is far to be complete, even for the closest objects. Recently, in Anders et al. (2021) we have attempted a first study of the completeness of the OC census based on the results in Castro-Ginard et al. (2020), and we have found that the recovery fraction of known OCs as a function of age is of 60% for OCs in the range of $1 - 10$ Myr and up to 90% for the older ones, indicating that blind detection algorithms are still necessary to reach a better completeness, perhaps using improved data from future *Gaia* releases.

OCs were detected at any distance, however, we found a void region where very few objects are present (Castro-Ginard et al. 2019, see Chapter 3). This region, named the Gulf of Camelopardalis, is located near the Galactic anticentre at $l \sim 140^\circ$, and reveals a complex structure at the second Galactic quadrant which became visible only after the release of *Gaia* DR2. The lack of new detections in this region is not due to shortcomings of our methodology, but a real physical feature that has been confirmed by independent studies (Cantat-Gaudin et al. 2019b) and deserves further analysis.

INCLUSION OF BIG DATA

One of the strong points of this thesis is the inclusion of a framework to enable the Big Data analysis of the *Gaia* data. This is achieved by adapting the methodology to work on a supercomputer facility, which in our case is the MareNostrum supercomputer at the Barcelona Supercomputing Center. To deploy the clustering algorithm, we used a parallelisation scheme based on graphs, PyCOMPSs (Tejedor et al. 2017), which distributes the computation based on the data dependencies of the code. Furthermore, in order to get full advantage of the MareNostrum supercomputer, we used *dislib* (Álvarez Cid-Fuentes et al. 2019), a Python-based machine learning library to analyse large scale datasets in a high-performance computing environment, to further parallelise the computation of `DBSCAN` if needed.

Using this Big Data framework, we were able to carry out a search for OCs driven by the physical properties of these objects and not the computational limitations. The tessellation scheme used to search for local overdensities allowed us to scan regions as large as $16^\circ \times 16^\circ$, which can contain 10^7 sources in *Gaia* DR2 up to magnitude $G = 17$. In an ongoing analysis of *Gaia* EDR3, where the uncertainties in parallax and proper motions have substantially decreased, the magnitude limit has been increased to $G = 18$ to reach the previous uncertainty level and thus the number of sources to be analysed in each region has been doubled. Adapting the methodology to run in a Big Data environment in MareNostrum, has decreased the execution time of the clustering algorithm in Castro-Ginard et al. (2018, which run on TGAS) from 18 hours to less than 1.5 hours in a single node (48 cores, Álvarez Cid-Fuentes et al. 2019, see Sect. V). But most importantly, it has enabled the processing of such a large amount of data in *Gaia* DR2 (122 727 809 stars in the Galactic disc), which would be impossible in regular approaches.

The second step which has beneficted from Big Data is the ANN confirmation of statistical overdensities as OCs. Given the impossibility of visually inspecting the huge amount of statistical clusters found by `DBSCAN`, the reliability of the automatic confirmation had to be increased. This was done by adopting a Deep Learning approach for the ANN, with the inclusion of several convolutional layers to automatically extract the meaningful features of an isochrone pattern in a CMD. The Deep-ANN used in Castro-Ginard et al. (2020) outperformed the single layer ANN in Castro-Ginard et al. (2018, 2019), and resulted in 676 tentative OC candidates of which 582 (86%) were later confirmed as OC candidates.

SOME NEW OPEN CLUSTERS OF INTEREST

The goal of Part i was to develop a methodology to detect as much clusters as possible, therefore the parameters of the methodology are tuned to find the largest number of OCs, and not targeted to OCs with some particular features. However, we were able to detect clusters with particular features which are interesting targets for specialised studies. I list some interesting examples below, all of them with ongoing or planned follow-ups.

UBC 7 and Collinder 135, a physical pair

In Castro-Ginard et al. (2018) we found two significant overdensities in the location of Collinder 135. One of them corresponds to the reported values of Collinder 135, while the mean parameters for the second overdensity were distinct enough to consider it as a separate OC. When inspecting the photometry of both structures, two sequences overlapped showing a very similar age, which revealed that both structures most probably formed in the same process. Therefore, in Castro-Ginard et al. (2018) we concluded that the two groups are two distinct OCs, perhaps with a common origin. This has been independently confirmed by Kovaleva et al. (2020), who investigated the origin of both groups, and concluded that the two clusters have formed as a physical pair and the actual separation of its centres is 24 pc.

This same pair has been studied by Cantat-Gaudin et al. (2019a), as part of a bigger complex, to study the three-dimensional structure and kinematics of the Vela-Puppis region. The authors found seven distinct groups with different ages, which are found to be in expansion respect one to another.

Substructure in the Carina Nebula

In Castro-Ginard et al. (2020), we found seven groups in an area of $\sim$ 14 deg^2 around the Carina Nebula. These groups correspond to the known Bochum 10, Trumpler 15, Trumpler 16 and Collinder 228 in addition to the newly discovered UBC 262, UBC 505 and UBC 653 clusters. This ability to find new substructure in known star-forming regions shows that our methodology can be fine tuned for such purpose, with the goal of accurately determine the structure and kinematics of those regions, as done in Zari et al. (2018), Lim et al. (2019), Galli et al. (2019), and Cantat-Gaudin et al. (2019a) already using *Gaia* DR2 data.

UBC 274, a disrupting old OC

Our methodology is also able to detect OCs with distinct features. This is the case of UBC 274, an old ($\sim$ 3 Gyr) cluster located at a relatively low latitude ($b \sim -12°$) at a distance of about 2 kpc from the Sun. UBC 274 represents a clear proof of the success of our methodology since it is the clearest new detection (detected the highest number of times in our Monte Carlo scheme, see Chapter 4), which is explained because UBC 274 is one of the most populated (365 member stars) and one of the biggest clusters in size among our new findings. It also shows the success of the photometric confirmation via our Deep-ANN: regardless of the presence of numerous blue stragglers the Deep-ANN is able to characterise the two-dimensional correlations in the CMD and recognise the presence of a populated main sequence and a red giant branch.

What makes UBC 274 interesting is its age ($\sim$ 3 Gyr), less than 20% of the catalogued clusters are older than 1 Gyr (and less than 5% older than 2 Gyr). It also shows signs of being disrupted by the gravitational field, forming tidal tails. This disruption events have already been characterised in other intermediate and old age OCs, such as in the Hyades, Praesepe, Blanco1, Ruprecht 147, and Coma Berenices (Röser et al. 2019a,b; Tang et al. 2019; Yeh et al. 2019; Zhang et al. 2020; Gaia Collaboration et al. 2020b). However, all these clusters are closer than 300 pc, and detecting tidal tails in farther clusters is challenging mainly because the uncertainties in the parallax rapidly increase (Luri et al. 2018). Therefore, the detection of such features in UBC 274 ($d \sim 2$ kpc) makes this cluster a perfect candidate for follow-up observations.

Soon after the reporting of UBC 274, Piatti (2020) studied the binary sequence of the cluster finding that the spatial and kinematical patterns of the binary population are very similar that those of the main sequence population, confirming the intense process of disruption that is taking place. Beyond this study, there are scheduled nights for observing UBC 274 with the S-PLUS filters in the T80-S telescope in Las Campanas. Moreover, also in Las Campanas, spectroscopic follow-up has already started in the MIKE spectrograph at the Magellan 2 Telescope. These photometric and high-resolution spectroscopic studies (P.I.: P. Jofré) will help in the better understanding of the disruption process.

ON THE NATURE OF MILKY WAY SPIRAL ARMS

Astrometric and astrophysical parameters are better estimated for an OC, averaging over all the member stars, than for single field stars. We have used this fact to estimate mean radial velocities (Tarricq et al. 2020), and astrophysical parameters such as age, distance and line-

of-sight extinction (Cantat-Gaudin et al. 2020b, see Appendix a) for 2017 known OCs, which is the total number of clusters discovered or confirmed by *Gaia* DR2. The increase in the number of known OCs, together with a better knowlegde of their mean parameters has enabled the dynamical study of the structures they trace. Particularly, and since the young OC population trace the spiral arms of our Galaxy disc (Kounkel et al. 2020; Cantat-Gaudin et al. 2020b), the study of the present age Galactic spiral arms and their evolution for the last $\sim$ 100 Myr has been tackled in Chapter 5.

We used the young population of OCs to discriminate as far as possible among different theories for the formation of the Milky Way spiral arms. On the one hand, we studied the distribution of the OCs as a function of age across a spiral arm. The features of such distribution are indicative of different rotation speeds for the stellar objects (OCs in our case), and spiral arms (birth places of OCs) (Dobbs et al. 2010). For spiral arms and OCs moving at different rotation velocities, an age gradient will appear in the OCs age distribution across the arm, with a stronger gradient for a bigger velocity difference; whereas no gradient will appear if both structures (arms and OCs) co-rotate. We find no OC age gradient across any spiral arm, which disfavours classical density waves as the main formation mechanism for spiral arms, favouring a flocculent Galaxy with transient spiral arms that co-rotate with OCs at any radius. On the other hand, we use the evolution of OCs from their birth positions to directly compute the different spiral pattern speeds for the Perseus, Local, Sagittarius and Scutum arms. We find no common spiral pattern speed for these arms, which are nearly co-rotating with their tracers (OCs in our case). This also discards the density wave scenario for the spiral nature of the Milky Way (Castro-Ginard et al., submitted).

We also tackled the present age structure of the spiral arms. We used 264 young OCs (younger than 30 Myr) in addition to the 84 high-mass star-formning regions used as classical tracers for spiral arms (Reid et al. 2014, 2019), to re-compute the structural parameters of the Perseus, Local, Sagittarius and Scutum spiral arms. By increasing the total number of tracers by a factor of 4, and increase the Galactocentric azimuthal range where these spirals are characterised, we are able to better contrain the arms parameters such as mean Galactocentric radius and pitch angle.

SCOPE OF MACHINE LEARNING IN A BIG DATA FUTURE FOR ASTRONOMY

The scope of machine learning in Astronomy goes beyond *Gaia*. With a trend to increase the volume of data delivered by astronomical surveys and the impossibility to manually inspect all this information, machine learning tasks such as classification, regression, clustering, etc., are

becoming more and more popular. In Kuhn et al. (2020, Appendix b), we developed an strategy to detect and characterise young stellar objects (YSOs) in near-infrared and infrared data from the Spitzer Space Telescope (Werner et al. 2004), in combination with 2MASS (Skrutskie et al. 2006), UKIDSS (Lawrence et al. 2007) and VVV (Smith et al. 2018). Kuhn et al. (2020) represents a complete contribution to the field of Astrostatistics, from the data preparation and treatment of missing data, to a classification to recognise YSOs in near-infrared photometry, and a clustering methodology to characterise YSO environments.

With the advent of the data-driven age of Astronomy, represented in missions such as SDSS, 2MASS or *Gaia*, and in the near future surveys such as WEAVE, 4MOST, *Euclid* or the Vera C. Rubin Observatory (previously known as LSST), the adaptation of machine learning and data mining techniques, together with a Big Data treatment of the algorithms, will become an essential and indispensable task to overcome the data analysis in such different data domains.

FUTURE WORK

A natural next step is the application of the methodology to the next *Gaia* releases. Particularly, for *Gaia* EDR3 (already public at the moment of the thesis deposit) and *Gaia* DR3, the improvement in the precision of parallaxes and proper motions in EDR3 will allow to search for OCs in a larger volume, while in the inclusion of millions of radial velocities and mean BP, RP and RVS spectra will allow for the use of this information, together with the astrometric measurements, to increase the parameter space where the overdensities are found and target the search to objects other than OCs across the whole sky.

- The devised methodology has shown to be a powerful tool to analyse large amounts of data (122 727 809 stars for *Gaia* DR2 in the Galactic disc up to magnitude $G = 17$). I am currently exploring the possibility of expanding the search in the Milky Way disc by analysing *Gaia* EDR3 up to magnitude $G = 18, 19$. These deeper magnitude limits increase the number of stars to be analysed by a factor of 2 and 4 approximately. The hyperparameters of the methodology will have to be adjusted to work with *Gaia* EDR3 new uncertainty levels. Also, I will have to build a new training data set to improve the performance of the OC identification based on Deep Learning neural networks. This search on *Gaia* EDR3 will allow to increase the detection volume of the OCs due to the better uncertainty for parallaxes and proper motions, and the deeper magnitude limits. For the innermost volume of the Galactic disc (w.r.t. the Sun), where the search region overlaps with the region used in *Gaia* DR2, the new search in *Gaia* EDR3 using an homogeneous metholology

will allow for completeness studies on the population of OCs, and to find a higher level of substructure in the known clusters.

- I also plan to extend the search outside the Galactic disc. This will benefit from *Gaia* DR3 radial velocities and the parameters derived from the *BP*, *RP* and *RVS* spectra, but also from the addition of extra information from APOGEE, WEAVE or the future 4MOST spectroscopic surveys and Pan-STARRS, 2MASS or the Vera C. Rubin photometric surveys. The challenge of this approach is to adapt the search to a data domain where not all the dimensions have complete information (for instance, only a subset of *Gaia* DR3 will have radial velocities due to the different magnitude limits in G_{RVS} and G). The goal of this search is two fold. First, to increase the dimensionality of the parameter space where to search for astrophysical objects, to detect OCs at high Galactic latitudes where they are less likely to be found. Second, to expand the search to objects different than OCs.

The previous points are a generalisation of Part i. Regarding the future work related to Part ii and Appendix a, I summarise some ideas below.

- By using OCs as main tracers of the spiral structure of our Galaxy, we already have robust tracers since the mean parameters such as sky position, distance, velocity or age are better estimated for a stellar population such an OC, in comparison to field stars. The inclusion of the more precise astrometric data in *Gaia* EDR3 will improve the study of the membership probabilities for these OCs from scratch. New membership determinations will better constrain the mean parameters for the OCs. Also, the inclusion of photometry at longer wavelengths (2MASS, for instance) in the ANN-based estimation of astrophysical parameters (Appendix a) will result in a better estimation of these quantities from CMDs.

- In Chapter 5, we showed that OCs are valid tracers of the Galactic spiral arms and their evolution. However, for the case of the Milky Way we only have one example, and the results obtained may be subject to different interpretations. I would like to analyse, using simulations of galaxies where their spiral structure has been excited by different mechanisms, the signature left by spiral arm formation processes in an OC-like population in that simulated galaxy.

Part IV

APPENDICES

a

ESTIMATION OF AGES, DISTANCES AND LINE-OF-SIGHT EXTINCTIONS FOR THE OPEN CLUSTER POPULATION

This Appendix contains the published version of Cantat-Gaudin et al. (2020b).

Astrophysical parameteres such age, line-of-sight extinction or distance modulus for all stars in an OC were traditionally estimated by manually finding the isochrone which represents the best fit on a CMD. ANN have shown to be efficient in the automatisation of such a task. In the paper of this Appendix, we describe the application of an ANN to homogeneously estimate ages, line-of-sight extinction and distances for the whole OC population. The ANN is trained on a set of reference clusters, with well estimated parameters, complemented by variations of these reference clusters created using data augmentation techniques to cover a wide range of the parameters to be estimated. In addition to the CMD, the ANN takes the median parallax of the OC and two additional quantities estimated from the CMD to improve the results.

This paper results in a catalogue of reliable parameters for 1867 OCs, which represents the largest and most homogeneous catalogue of OC parameters.

A&A 640, A1 (2020)
https://doi.org/10.1051/0004-6361/202038192
© ESO 2020

Astronomy & Astrophysics

Painting a portrait of the Galactic disc with its stellar clusters⋆

T. Cantat-Gaudin[1], F. Anders[1], A. Castro-Ginard[1], C. Jordi[1], M. Romero-Gómez[1], C. Soubiran[2], L. Casamiquela[2], Y. Tarricq[2], A. Moitinho[3], A. Vallenari[4], A. Bragaglia[5], A. Krone-Martins[3,6], and M. Kounkel[7]

[1] Institut de Ciències del Cosmos, Universitat de Barcelona (IEEC-UB), Martí i Franquès 1, 08028 Barcelona, Spain
e-mail: tcantat@fqa.ub.edu
[2] Laboratoire d'Astrophysique de Bordeaux, Univ. Bordeaux, CNRS, UMR 5804, 33615 Pessac, France
[3] CENTRA, Faculdade de Ciências, Universidade de Lisboa, Ed. C8, Campo Grande, 1749-016 Lisboa, Portugal
[4] INAF-Osservatorio Astronomico di Padova, Vicolo Osservatorio 5, 35122 Padova, Italy
[5] INAF-Osservatorio di Astrofisica e Scienza dello Spazio, Via Gobetti 93/3, 40129 Bologna, Italy
[6] Donald Bren School of Information and Computer Sciences, University of California, Irvine, CA 92697, USA
[7] Department of Physics and Astronomy, Western Washington University, 516 High St, Bellingham, WA 98225, USA

Received 17 April 2020 / Accepted 6 May 2020

ABSTRACT

Context. The large astrometric and photometric survey performed by the *Gaia* mission allows for a panoptic view of the Galactic disc and its stellar cluster population. Hundreds of stellar clusters were only discovered after the latest *Gaia* data release (DR2) and have yet to be characterised.
Aims. Here we make use of the deep and homogeneous *Gaia* photometry down to $G = 18$ to estimate the distance, age, and interstellar reddening for about 2000 stellar clusters identified with *Gaia* DR2 astrometry. We use these objects to study the structure and evolution of the Galactic disc.
Methods. We relied on a set of objects with well-determined parameters in the literature to train an artificial neural network to estimate parameters from the *Gaia* photometry of cluster members and their mean parallax.
Results. We obtain reliable parameters for 1867 clusters. Our catalogue confirms the relative lack of old stellar clusters in the inner disc (with a few notable exceptions). We also quantify and discuss the variation of scale height with cluster age, and we detect the Galactic warp in the distribution of old clusters.
Conclusions. This work results in a large and homogeneous cluster catalogue, allowing one to trace the structure of the disc out to distances of ∼4 kpc. However, the present sample is still unable to trace the outer spiral arm of the Milky Way, which indicates that the outer disc cluster census might still be incomplete.

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

1. Introduction

The shape and dimension of our galaxy, which we commonly refer to as the Milky Way, is difficult to appreciate from our vantage point. From the pioneering work of early modern astronomers (Herschel 1785; Shapley 1918; Trumpler 1930) to recent studies (Reid et al. 2019; Gravity Collaboration 2019; Anders et al. 2019), the distance to individual objects is one of the most valuable pieces of information we rely on to reconstruct the overall structure of the Milky Way.

Among the variety of astronomical objects to which we can derive distances, stellar clusters present the advantage of spanning a wide range of ages, from a few million years (tracing episodes of recent star formation) to several gigayears (as old as the Galactic disc itself), which can be estimated with a greater precision than for individual stars. Samples of clusters with known ages have long been used to trace various properties of the Galactic disc, such as the path of its spiral arms (Becker & Fenkart 1970) or the evolution of its scale height (van den Bergh 1958). Although the precision and accuracy of age estimates are tied to the quality of the observational data and the correctness of theoretical models, distinguishing a young cluster from an old one is often relatively straightforward in a colour-magnitude diagram[1]. While the first catalogues of cluster parameters only reported sky coordinates (e.g. Melotte 1915) and sometimes distances (Trumpler 1930; Collinder 1931), modern catalogues also provide associated ages. The widely-used catalogue of Dias et al. (2002) is a curated compilation of parameters from a large number of studies, which was obtained with a variety of methods and photometric systems. Another widely-cited study by Kharchenko et al. (2013) presents an automated characterisation of the cluster population (known at the time), which was performed with all-sky 2MASS photometry (Skrutskie et al. 2006). It represents a homogeneous set of parameters, but to a lesser precision than dedicated studies of individual objects.

The second data release of the European Space Agency (ESA) *Gaia* mission (DR2: Gaia Collaboration 2018a) represents the deepest all-sky astrometric and photometric survey ever conducted. The *Gaia* astrometry (proper motions and parallaxes)

⋆ List of cluster parameters and complete list of their members are only available at the CDS via anonymous ftp to cdsarc.u-strasbg.fr (130.79.128.5) or via http://cdsarc.u-strasbg.fr/viz-bin/cat/J/A+A/640/A1

[1] Trumpler (1925) was the first to group clusters by age according to their magnitude-spectral class diagrams, but his evolutionary sequence was wrong. It was then believed that stars formed as giants and contracted into main-sequence dwarfs (see Sandage 1988, for a discussion).

allows us to identify the members of clusters, and it has enabled the discovery of several hundreds of new objects. Combining parallaxes with the deep *Gaia* photometry allows us to estimate cluster distances, ages, and extinctions on a large scale with unprecedented precision. Thus far, the largest study on this particular topic was conducted by Bossini et al. (2019), who derived parameters for 269 clusters (mostly nearby and well-populated). Despite the high precision of their results, this sample only constitutes less than 15% of the clusters for which members can be identified with *Gaia*.

The aim of the present work is to study the structure of the Galactic disc revealed by clusters of various ages. To this effect, we derived cluster parameters in a homogeneous and automatic fashion for ∼2000 Galactic clusters with members identified in the *Gaia* data. In Sect. 2 we present the input data and our list of reference clusters. Section 3 describes the artificial neural network that we built and trained in order to estimate parameters. Section 4 introduces our catalogue of cluster parameters. In Sect. 5 we use this cluster sample to trace the structure of the Galactic disc. Section 6 contains a discussion of the results, and Sect. 7 closes with concluding remarks.

2. Data

2.1. Cluster members from Gaia DR2

We retained the probable members (probability >70%) of 1481 clusters whose membership list was published by Cantat-Gaudin & Anders (2020), who estimated the membership probabilities for stars brighter than $G = 18$ using the unsupervised classification scheme UPMASK (Krone-Martins & Moitinho 2014; Cantat-Gaudin et al. 2018a). We collected the list of members provided by the authors for the recently discovered UBC clusters (Castro-Ginard et al. 2018, 2019, 2020).

We also applied UPMASK to the 56 cluster candidates proposed by Liu & Pang (2019). We were able to find secure members for 35 of them. These objects are listed in our catalogue as "LP", followed by the entry number given in Liu & Pang (2019). Since UPMASK is not suited for very extended clusters, we took the list of members for the nearby clusters Melotte 25 (the Hyades) and Melotte 111 (Coma Berenices), which were derived from *Gaia* DR2 astrometry by Gaia Collaboration (2018b). In total, this compiled sample comprises ∼230 000 stars that are brighter than $G = 18$, which belong to 2017 clusters.

2.2. Reference clusters

We compiled a list of 347 clusters with parameters (age, reddening, and distance modulus) that are known to a sufficient precision to be used as points of reference. Their ages and distances are shown in Fig. 1. We strove to use a small number of reference studies to maximise homogeneity, while also covering the entire parameter space and privileging studies that employed *Gaia* data for their membership selection.

The 269 clusters of Bossini et al. (2019) represent the bulk of this reference set, and they constitute the largest homogeneous sample of cluster ages obtained from *Gaia* data. Their parameters were determined by fitting PARSEC isochrones (Bressan et al. 2012) with the Bayesian code BASE9 (von Hippel et al. 2006) to *Gaia* DR2 photometry of the cluster members identified in Cantat-Gaudin et al. (2018b).

This sample contains few clusters that are older than 1 Gyr and few clusters that are more distant than 4 kpc. We therefore supplemented the sample with 36 clusters from the BOCCE survey, which focuses mainly on old clusters, of which many are distant and characterised with a combination of deep photometry and high-resolution spectroscopy (Bragaglia & Tosi 2006; Bragaglia et al. 2006; Tosi et al. 2007; Andreuzzi et al. 2011; Cignoni et al. 2011; Donati et al. 2012, 2014a, 2015; Ahumada et al. 2013).

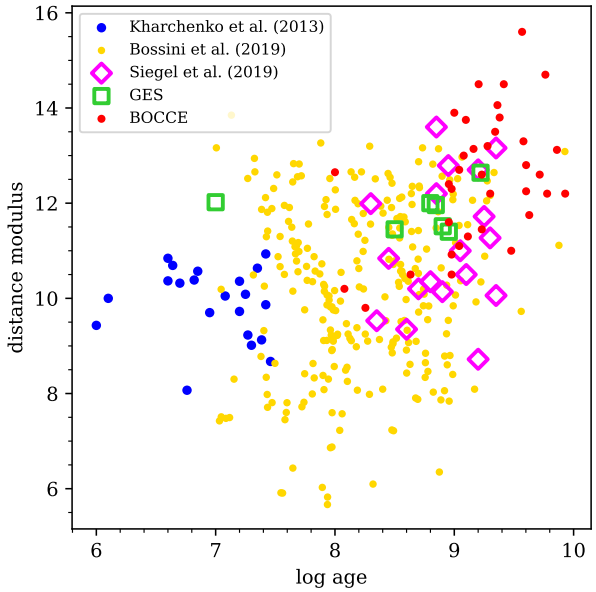


Fig. 1. Age and distance modulus of our reference clusters (described in Sect. 2.2).

Since these two samples contain very few clusters that are younger than $\log t \sim 7.5$, we supplemented the training set with 21 young clusters with distances smaller than 1.5 kpc and parameters that were taken from the catalogue of Kharchenko et al. (2013), which have visually well-defined colour-magnitude diagrams. We also included seven clusters that have been the subject of dedicated papers by the *Gaia*-ESO Survey: NGC 3293 (Delgado et al. 2016); NGC 4815 (Friel et al. 2014); NGC 6705 (Cantat-Gaudin et al. 2014); NGC 6802 (Tang et al. 2017); Pismis 18 (Hatzidimitriou et al. 2019); Trumpler 20 (Donati et al. 2014b); and Trumpler 23 (Overbeek et al. 2017). We consider their parameters to be especially reliable due to the large number of radial velocities collected for these studies (allowing for good membership selections) and precise metallicities.

The *Swift* UVOT Stars Survey provides cluster parameters for 49 clusters, which were studied with *Gaia* DR2 astrometry and isochrone fitting to near-ultraviolet photometry (Siegel et al. 2019). Eighteen of them are not present in the previously mentioned references, so we added them to our reference sample.

3. Cluster parameters and machine learning

Estimating the main parameters (age, distance modulus, extinction, and sometimes metallicity) of a star cluster is often done via isochrone fitting: A theoretical model of the sequence traced by a coeval group of stars in a two-dimensional colour-magnitude diagram (CMD) is compared to the observed distribution of stars. Perhaps surprisingly, designing a robust and efficient automatic procedure for isochrone fitting is far from trivial. Observed CMDs of clusters do not simply follow a single sequence, but they feature unresolved binaries (a problem addressed by the

τ^2 statistics of Naylor & Jeffries 2006), blue stragglers, broadened turnoffs (Marino et al. 2018; Bastian et al. 2018; Sun et al. 2019; Li et al. 2019; de Juan Ovelar et al. 2020), and almost always contamination by field stars, which can also be taken into account with ad-hoc statistics (as in e.g. Monteiro et al. 2010). The stellar phases that provide the most clues about the age and distance of a cluster (its turnoff, red clump, and red giants) also happen to be the least populated parts of a CMD[2], and they must be given a higher weight subjectively. The Bayesian code BASE9 (von Hippel et al. 2006; Jeffery et al. 2016) relies on robust statistical principles and it allows for the use of prior knowledge (most importantly, a distance constraint provided by e.g. *Gaia* parallaxes). However, its runs can be very time-consuming, it generally requires a large number of cluster members (it was in fact originally designed for globular clusters), and it is currently unable to deal with CMDs affected by differential extinction. The `ASteCa` package (Perren et al. 2015) uses a sophisticated approach with a modelling of a synthetic cluster from theoretical isochrones, but it is also relatively time-consuming and unable to deal with differential extinction and blue stragglers at present.

Isochrone fitting is therefore often performed by hand, which when done properly provides satisfactory results, but it is impractical to perform it on the samples of hundreds to thousands of clusters available from modern sky surveys. To address this problem and avoid direct comparisons with theoretical isochrones, we built a data-driven procedure to estimate the parameters of an unknown cluster based on its similarities with objects of known parameters. Although the age accuracy is ultimately tied to the reference values, which are derived from stellar evolution models, our approach has the advantage of putting all clusters on the same age scale and providing reliable relative ages. Learning from labelled CMDs can be thought of as a generalisation of the empirically calibrated morphological age index, which allows for a quick estimate of a cluster age by measuring the magnitude difference between its turn off and red clump (used by e.g. Lynga 1982; Janes & Adler 1982; Janes et al. 1988; Phelps et al. 1994; Carraro & Chiosi 1994; Janes & Phelps 1994; Friel 1995; Salaris et al. 2004), or the morphological age ratio (Anthony-Twarog & Twarog 1985; Twarog & Anthony-Twarog 1989).

3.1. Artificial neural network

The increasing size and dimensionality of astronomical datasets have made machine learning increasingly popular in the field (see e.g. the reviews by Fluke & Jacobs 2020; Baron 2019). Artificial neural networks (ANNs) are particularly popular due to their flexibility and performance at both classification (e.g. Ting et al. 2018; Castro-Ginard et al. 2018) and regression tasks (e.g. Leung & Bovy 2019; Kounkel & Covey 2019; Boucaud et al. 2020).

An ANN is a system that maps the input observables to the target output quantities through a series of nodes. Here, the three targets are the cluster age, extinction, and distance modulus. Nodes are organised in layers, where every node receives input from the previous layer and output from a non-linear function of the input to the successive layer. For this work, we use a rectified linear unit (ReLU). Formally, ANNs are universal approximators, which means that any continuous function can be approximated by an ANN with at least one hidden layer. Approximating a complex function might require a large number of nodes in the hidden layer, making the network slower to train and more prone to overfitting. An equivalent or better approximation can often be obtained with a smaller number of nodes if they are organised into several hidden layers in which each one contains an increasingly abstract representation of the data structure. For this study we experimented with various architectures and settled on an ANN with three hidden layers, as is shown in Fig. 2.

The main input observable that we provided to our ANN was a 2D histogram of the *Gaia* colour-magnitude diagram of each cluster, with a bin width of 0.2 mag in colour and 0.5 mag in magnitude. The histogram was pre-processed before being fed to the ANN. We took the logarithm of the counts and scaled it so the most populated bin always had a value of 1. The entire histogram contains 700 bins. Applying a principal component analysis to the flattened histograms of our training set (described in Sect. 3.2) shows that 99.9% of the variance can be expressed with only 410 components. We therefore applied the transformation computed on the training set, which reduced the number of input quantities by nearly half, with a negligible loss of information.

We also provided the median parallax $\langle\varpi\rangle$ to the ANN, which is a strong predictor of distance, especially for the most nearby clusters. For each cluster, we provided two additional quantities estimated from the CMD (illustrated in Fig. 2). The quantity $s_{\rm bright}$ is the slope in the relation between colour and magnitude for the stars whose distance-corrected magnitude[3] is brighter than 4. This quantity strongly correlates with the cluster age. Finally, we denote $\mathrm{MS}_{4,5}$ as the mean colour of stars whose distance-corrected magnitude is between 4 and 5. In this magnitude range, stars are always expected to be on the main sequence even in the oldest clusters, and their colour is a strong predictor of reddening.

If fewer than ten stars were available to estimate $s_{\rm bright}$, we set it to an edge value of -10. If no stars were available for $\mathrm{MS}_{4,5}$, which happens for distant and reddened clusters, we also set its value to -10. Thanks to their hidden layers, ANNs are able to approximate logical functions, which implicitly allows them to handle missing values.

3.2. Training set

Our first attempts to estimate cluster parameters involved ANNs, which were trained with mock CMDs. Such systems were extremely good at recovering the input parameters of other mock CMDs, but overall they returned disappointing results when applied to real, observed, *Gaia* CMDs. We therefore chose a data-driven approach that would not require us to generate mock clusters from theoretical models. Training machine learning procedures on labelled observed data is an increasingly common practice in various sub-fields of astronomy. For instance Ting et al. (2018) trained an ANN to distinguish red giant branch stars from red clump stars, Leung & Bovy (2019) determined elemental abundances with an ANN trained on high signal-to-noise ratio spectra, and Arnason et al. (2020) identified new X-ray binary candidates in M 31.

The basis of our training set are the clusters presented in Sect. 2.2. A good training set must not only cover a wide range of parameters, but also be dense enough so that the ANN cannot memorise it and it must learn how the relevant features relate to

[2] The pre-main sequence of young clusters is also a good age indicator, but these low-mass stars are too faint to be observed in most objects.

[3] The distance-corrected magnitude of a star is based on the cluster mean parallax $G = 5 \times \log_{10}(\langle\varpi\rangle/1000) + 5$ and does not include correction for reddening.

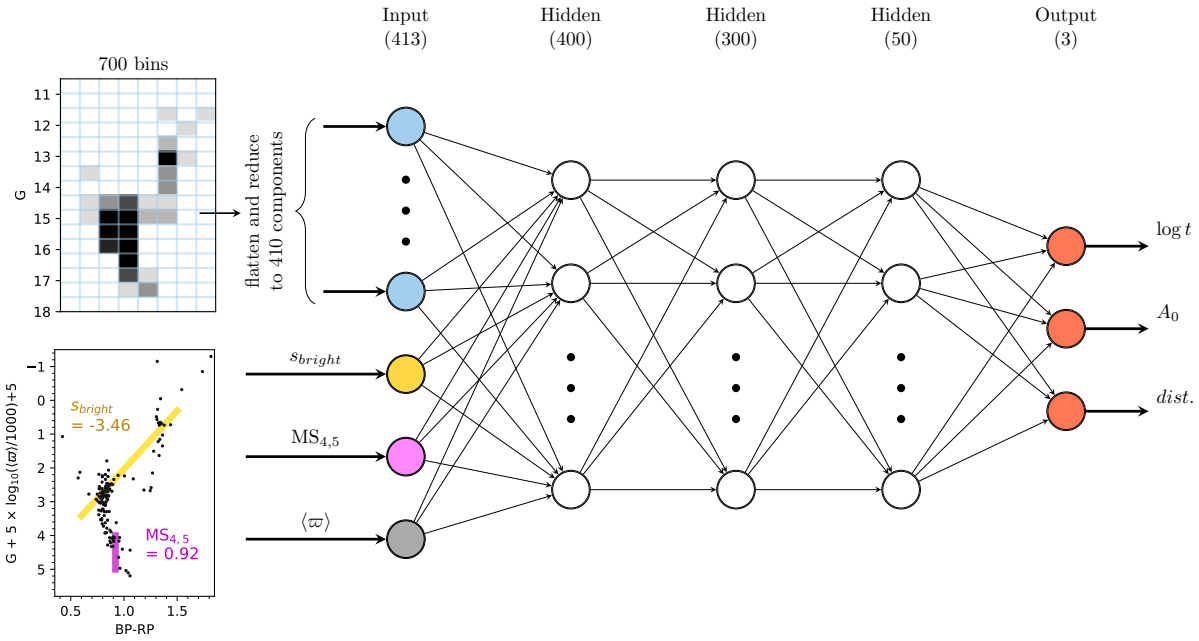


Fig. 2. Architecture of our artificial neural network, indicating the width (number of nodes) of each layer. The example cluster is Haffner 22. The input quantities are described in Sect. 3.1.

the output. We performed data augmentation by creating variations of the reference clusters by artificially increasing their distance modulus and their extinction, by sub-sampling them, and by adding differential extinction.

The simulated distance modulus was randomly picked between 0.5 mag smaller than the reference value and 16 mag (~15.85 kpc). We adjusted the simulated parallax accordingly and removed stars whose simulated G magnitude was fainter than 18. To account for the uncertainties in the mean parallax, the local parallax zero-point variation, and to simulate the known zero-point offset in parallaxes (Lindegren et al. 2018; Arenou et al. 2018), we then subtracted 0.029 mas and added a random offset that was uniformly picked between −0.05 and +0.05 mas. Adding noise to the simulated parallaxes is important so the ANN learns that for distant clusters, the distance modulus is mostly constrained by the CMD morphology and not by the parallax.

In order to cover a wide range in extinction, additional extinction was added up to $A_0 = 5$, using the polynomial relation presented by Danielski et al. (2018) and Gaia Collaboration (2018b). Differential reddening was added to half of the variations by first picking a random value between 0 and 1, setting the intensity of the differential reddening for this variation, then adding a random extinction picked between 0 and this maximum intensity.

Finally, we sub-sampled every reference cluster by picking a random number of stars, which went as low as ten, for every variation. In total we created 1500 versions of each reference cluster and 3000 for the clusters with $\log t < 7.4$ and $\log t > 9.4$ since there are few of them in our reference sample.

Since the cluster members were selected based on their astrometry only, many clusters (especially the distant ones) include a fraction of field star contaminants. They were not removed from the training set, which means some training examples contain field contamination. The trained ANN is therefore able to deal with contamination in the non-reference clusters that we characterise in this study.

3.3. Implementation and training

We implemented the ANN on a desktop computer as a multilayer perceptron regressor from the `scikit-learn` Python library (Pedregosa et al. 2011). The training was performed with the built-in ADAM solver (Kingma & Ba 2014). The `scikit-learn` implementation optimises the R-squared score defined as $R^2 = 1 - \frac{u}{v}$ where $u = \sum(y_{\rm true} - y_{\rm pred})^2$ is the residual sum of squares and $v = \sum(y_{\rm true} - \overline{y_{\rm true}})^2$ is the total sum of squares. A score of 1 would indicate a perfect prediction for all of the labels.

To make each training iteration faster and to alleviate the risk of the optimisation staying stuck in a local optimum, each iteration only used a random 20% of the training set. We built a validation set, which was created exactly like the training set, but containing other random variations of the reference clusters. We trained the ANN for 1000 iterations. At each iteration, we also verified the prediction of the ANN on the validation sample. We show in Fig. 3 that although the training score steadily increases, the validation score reaches a maximum of around 200 iterations then it slowly decreases, which is a sign that the ANN starts overfitting. For the rest of this study, the ANN that we use is the one that was trained for 200 iterations.

3.4. Performance on the validation set

To assess the ability of the ANN to recover ages, extinctions, and distances, we investigated its performance on the validation set. Figure 4 shows the difference between the age estimated by the ANN and the reference value as a function of the number of stars. We see from this figure that young clusters with very few stars tend to have their ages slightly overestimated because

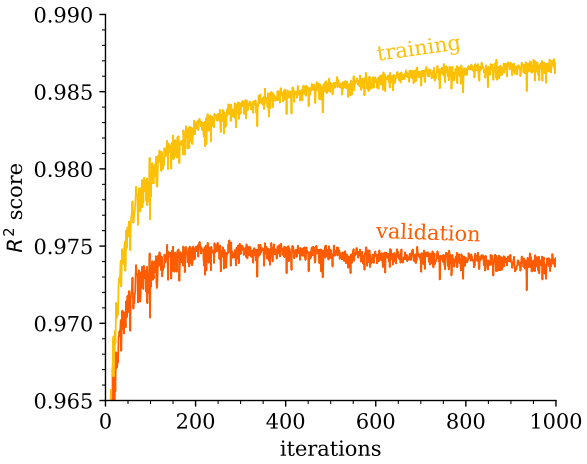


Fig. 3. Evolution of the training and validation scores with training iterations. The network used in this study is the result of 200 iterations.

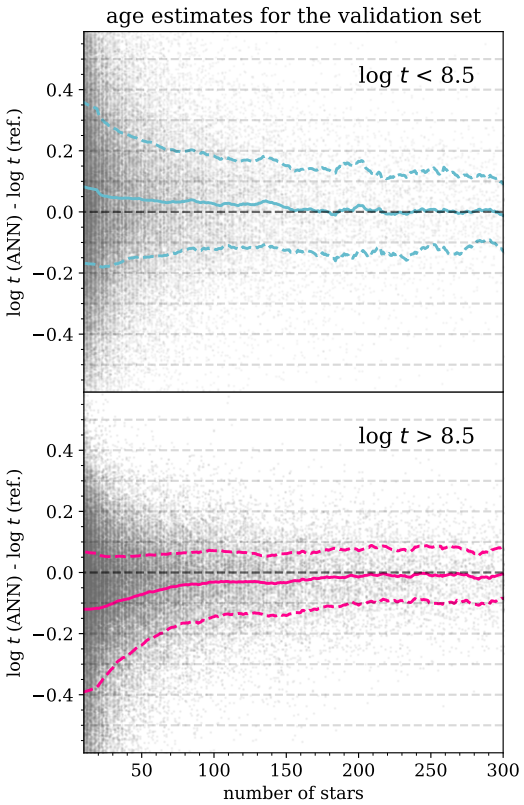


Fig. 4. Difference between the age estimate and the reference value for ~120 000 validation samples, split in two age groups. The full line is a running mean. The dashed lines represent the upper and lower standard deviation.

the sparsely populated turn off appears fainter. Whereas for old clusters, the absence of red giants makes them appear younger. This is not specific to our machine learning approach, but rather a general limitation of using CMDs to estimate cluster ages. In practice, less than 10% of our observed clusters have fewer than 20 members. In a successive step (Sect. 4), we also flag the clusters whose CMDs are too sparse and/or too blurry to show a meaningful pattern.

Overall, the uncertainty on the determination of $\log t$ ranges from 0.15 to 0.25 for young clusters and from 0.1 to 0.2 for old clusters. For the extinction and distance modulus, the precision of the ANN also depends on the number of stars, but only marginally on the age of the cluster. The typical uncertainty of A_0 ranges from 0.1 to 0.2 mag, and the typical distance modulus uncertainty ranges from 0.1 to 0.2 (~5% to 10% distance uncertainty).

If we assume that the reference values represent the ground truth, then these mean differences indicate the precision of our procedure. However the scatter encompasses both the uncertainties due to our methodology and the uncertainties of the reference parameters.

At the beginning of training, the weights of the ANN are initialised to random values. Every training run therefore converges to a slightly different final state. We have verified that the difference between several networks trained for 200 iterations with the same training set is negligible.

4. The catalogue of cluster parameters

We applied the trained ANN to estimate the parameters of all 2017 clusters mentioned in Sect. 2.1. We visually inspected the CMD of every cluster, with theoretical isochrones corresponding to the estimated parameters. For the large majority of them, the result looked satisfactory and closely matched the result that would have been obtained by a human expert. In 61 cases, the parameters had to be adjusted manually in order to better match the aspect of the CMD with a PARSEC isochrone (Bressan et al. 2012) of solar metallicity. The reason why the ANN performed poorly on these objects is not clear – they do not correspond to a specific age or distance range – and might be due to field contaminants. The parameters proposed by the ANN were still close enough to make this manual correction faster than having to pick an isochrone without a suggested starting point. We flagged these 61 objects in our catalogue.

We also flagged 81 clusters whose CMD is too blurred and reddened. They mostly distribute close to the Galactic plane in the direction of the Galactic centre, and most of them are known embedded clusters. Some of these objects include NGC 1579, which is associated with the Northern Trifid HII region, or the young massive clusters Westerlund 1 and Westerlund 2.

We further flagged 69 objects for which the CMD is too sparse to estimate meaningful parameters from photometry. Finally, we used literature values for three objects with a clear enough CMD but where the ANN failed to recover good parameters. Two of them are the very nearby Hyades (Melotte 25) and Coma Ber (Melotte 111), whose distance modulus is out of the range covered by our training set. We set their parameters to the values quoted by Gaia Collaboration (2018b). The third cluster is *Gaia* 2 for which our only members are red giant branch stars. We took its parameters from Koposov et al. (2017).

We end up with 1867 clusters with reliable parameters. We provide the list of all investigated clusters with their mean parameters and corresponding flags as an electronic table available at the CDS.

4.1. Comparisons with the literature

In the top row of Fig. 5, we show comparisons between our recovered parameters and the values listed by Kharchenko et al. (2013; hereafter K13), which were obtained by isochrone fitting

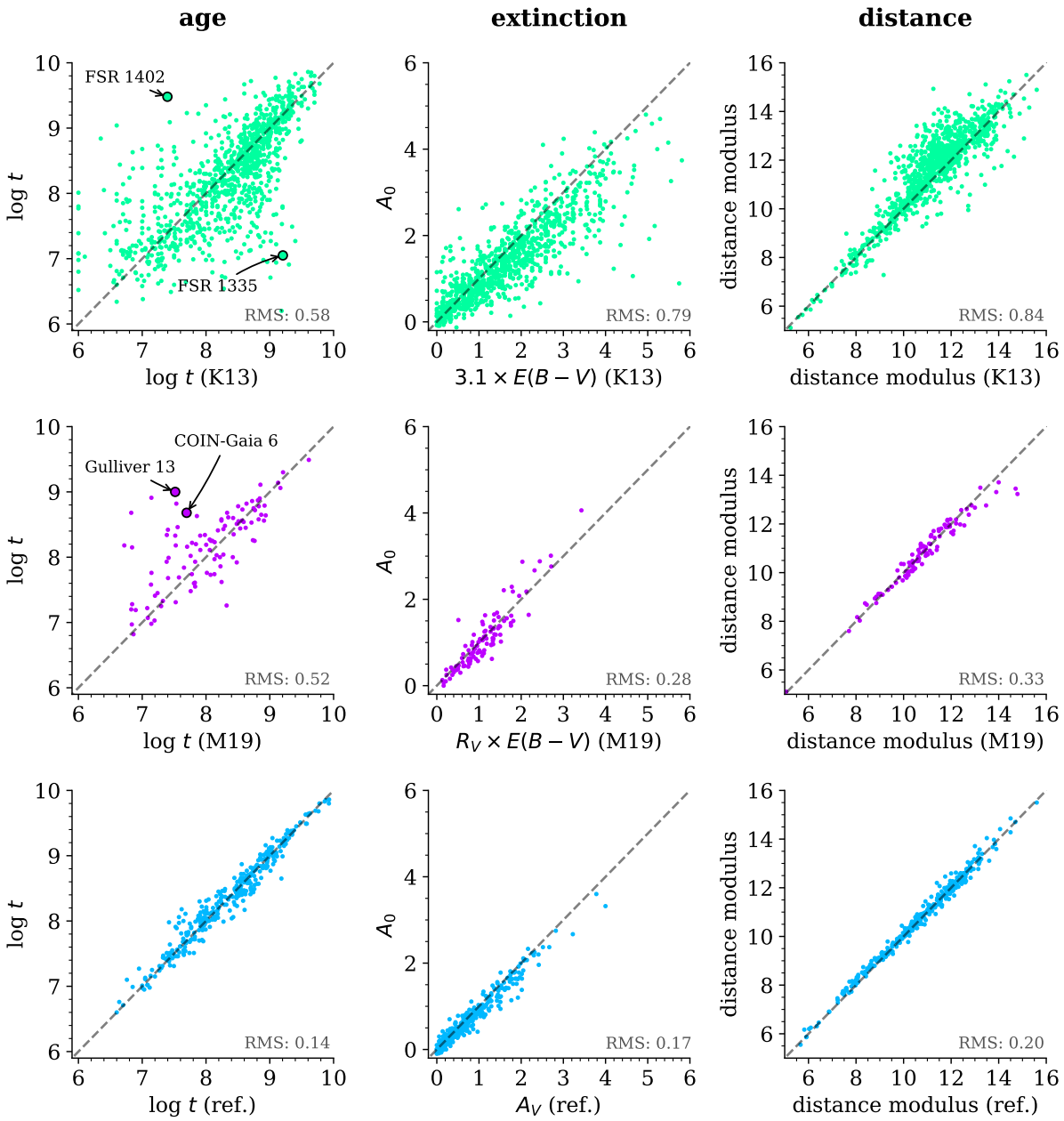


Fig. 5. *Top row*: comparison of the parameters for the clusters in common with Kharchenko et al. (2013). *Middle row*: comparison of the parameters for the clusters in common with Monteiro & Dias (2019). The CMDs and isochrones for the labelled clusters are shown in Fig. 6. *Bottom row*: comparison between our ANN parameters and the literature references presented in Sect. 2.2. All panels display the root mean square (rms) difference.

to 2MASS photometry (Skrutskie et al. 2006). Many of the clusters for which K13 lists old ages while we find young ages are very reddened objects, where the bright turnoff stars have been mistaken by K13 for a red branch (e.g. FSR 1335, whose CMD is shown in Fig. 6). Conversely, the cleaner membership and the distance constraint provided by the *Gaia* astrometry show that objects such as FSR 1402 (also shown in Fig. 6) are evolved clusters. Since FSR 1335 is young, sparse, and distant, any estimate of its age from just *Gaia* photometry of its brightest members is affected by large uncertainties. It is, however, evident that it is not an old cluster. Our procedure generally returns lower extinctions than K13. This could be due to our choice of defining A_0 as the extinction corresponding to the blue edge of the sequence in a CMD, before the effect of differential reddening, rather than determining the value for which the isochrone passes through the middle of the sequence.

A comparison with the parameters recently published by Monteiro & Dias (2019) is shown in the middle panel of Fig. 5. The authors relied on *Gaia* DR2 to select cluster members and constrain their distance, thus explaining the better agreement to

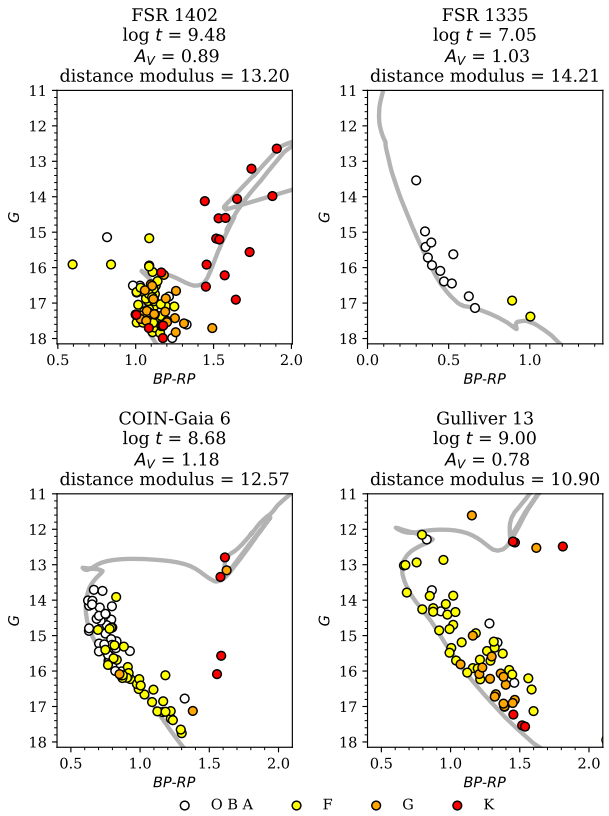


Fig. 6. Colour-magnitude diagram, colour-coded by spectral type from the effective temperatures of `StarHorse` (Queiroz et al. 2018; Anders et al. 2019) for the four clusters labelled in Fig. 5. The lines are PARSEC isochrones of solar metallicity.

our results. Several clusters still have discrepant age estimates, almost all of them are due to the presence of red stars that we consider to be cluster members. Two of them are labelled in Fig. 5, and their CMDs are shown in Fig. 6.

The bottom row of Fig. 5 shows comparisons with the reference values for the clusters we used to build the training set (presented in Sect. 2.2). The fact that we do not exactly recover the reference parameters is a good sign because it shows that the ANN did build an approximation of the relation between observables and cluster parameters, rather than memorising the aspect of reference clusters. The largest age discrepancies affect a handful of clusters for which Bossini et al. (2019) list ages $\log t \sim 7.6$, while our ANN estimates $\log t \sim 7.9$. These objects are too distant for their pre-main sequence stars to be visible, so the main age constraint is the ill-defined location of their turn off.

4.2. Composite Hertzprung-Russell diagram

Having an estimate of A_0 for each cluster, we corrected the observed colours and magnitudes for interstellar extinction by inverting the relations given in Danielski et al. (2018) and Gaia Collaboration (2018b). We then corrected G for distance modulus. The comprehensive Hertzprung-Russell diagram (HRD), which is made up of 1867 clusters, is shown in Fig. 7.

Since a single value of extinction was used for each cluster, this HRD is still affected by differential extinction, which is especially apparent in the elongation of the red clump. A few white dwarfs can be seen. They belong to the very nearby Hyades (Melotte 25), Coma Ber (Melotte 111), and Praesepe (NGC 2632). In the lower right part of the diagram, the presence of pre-main sequence stars is clearly visible in clusters younger than $\log t \sim 8$.

All of the cluster members used in this study have an apparent G magnitude that is brighter than 18. Since most of the old and very populated clusters are distant objects (e.g. Berkeley 32 or Collinder 261), few old stars with $M_G > 5$ are visible in the HRD.

4.3. Limitations and potential improvements

Although age, distance, and extinction are the parameters that contribute most to the aspect of a cluster in a CMD, metallicity also plays a role, especially for the coolest stars. Some studies leave it as a free parameter when performing isochrone fitting, but it is common to keep it fixed to an assumed value, as a wrong value mostly affects the reddening and only has a small impact on ages[4]. In this study we did not train the ANN to estimate metallicities, but the training set spans a large range in metallicity. Given that we fed the ANN a coarsely binned representation of the CMD, and given the strong degeneracy between metallicity and extinction, it is unlikely that our ANN could be used to make meaningful estimations of this quantity. An additional issue is that only a relatively small fraction of clusters have homogeneous and precise abundance determinations from high-resolution spectroscopy, which are and often from inhomogeneous sources (a problem discussed by Heiter et al. 2014), meaning that such a machine learning procedure would have to rely on a training set built with mock data.

Since we binned the millimag-precision *Gaia* DR2 photometry (Evans et al. 2018) into a grid with a resolution of 0.2 mag in colour and 0.5 mag in G magnitude, our approach is obviously not able to take advantage of the finest features observed in some *Gaia* CMDs. For the best-defined clusters, isochrone fitting procedures (e.g. Naylor & Jeffries 2006; von Hippel et al. 2006; Monteiro et al. 2010) are able to extract more information from the CMDs. We experimented with a finer binning of the CMD, but the size of the training set and the exponential increase in training time made this impractical. In the future, procedures employing an adaptive kernel density estimation might help to overcome this issue.

The use of ground-based photometry, especially at non-optical wavelengths, and value-added catalogues containing astrophysical parameters for individual stars (Andrae et al. 2018; Anders et al. 2019) could help to provide better constraints on the cluster parameters. Colour-magnitude diagrams are not an optimal approach for young clusters, especially when their pre-main sequence stars are not visible and the only age constraint is the colour of the bluest, most massive, identified member. They can also be affected by significant inhomogeneous extinction or feature small age spreads. When spectroscopic measurements are available, the lithium depletion boundary method (LDB) can provide a better constraint than photometry (e.g. Barrado y Navascués et al. 2004; Jeffries & Oliveira 2005), but it can return older ages than CMD fitting (e.g. 21 Myr versus 7.5 Myr in Jeffries et al. 2017). Lyra et al. (2006) have reported and

[4] The morphological age index of Salaris et al. (2004) includes a $\log t$ correction of 0.07 per dex of metallicity.

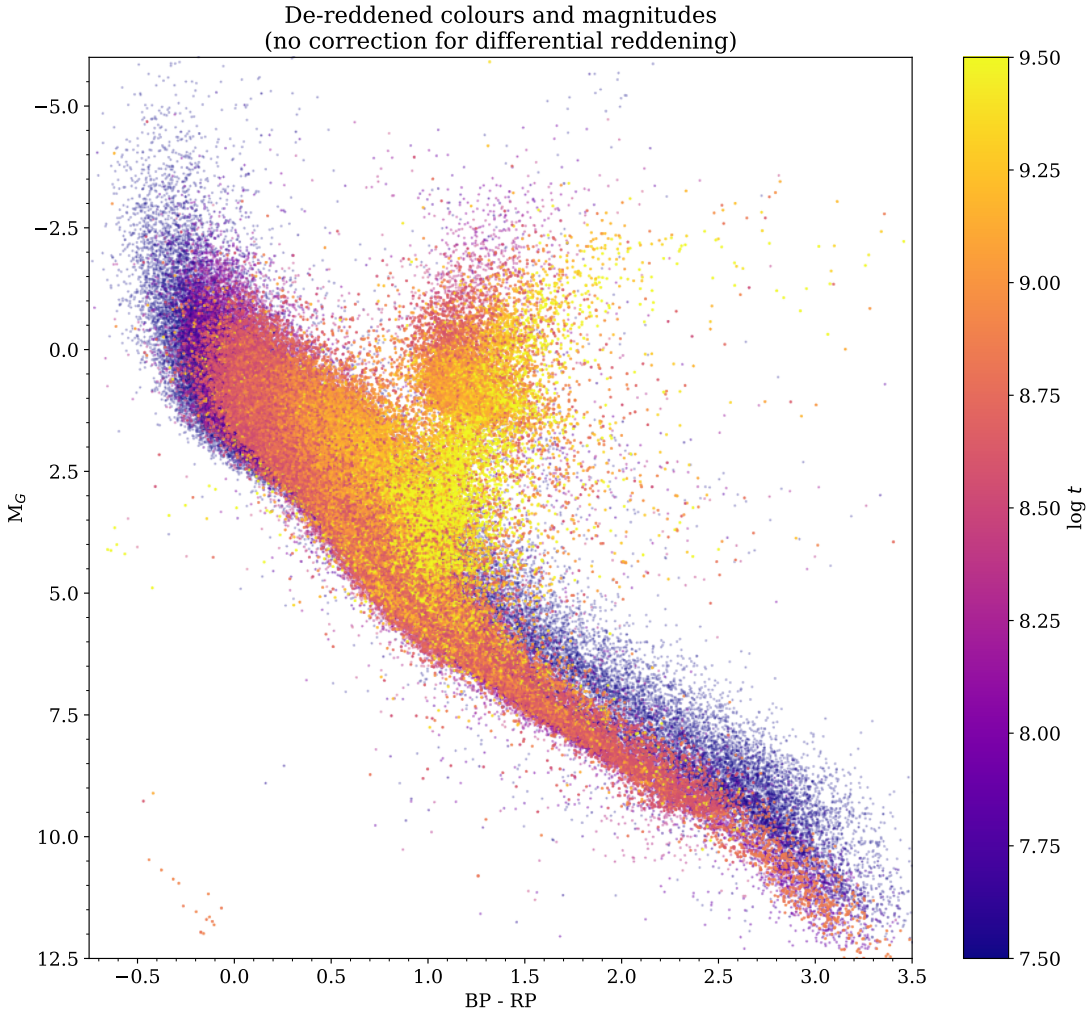


Fig. 7. Comprehensive Hertzprung-Russell diagram including 1867 clusters, colour-coded by cluster age.

discussed systematical differences between the nuclear ages, for main sequence stars, and contraction ages for pre-main sequence stars. Randich et al. (2018) performed a homogeneous analysis of seven clusters younger than ∼100 Myr, making use of three different sets stellar evolution models of (J, H, K_s, V) photometry and LDB models. They confirm that much of the scatter found in the literature for the age of these objects can be attributed to the use of different models or the choice of photometric passbands included in the isochrone fitting. An additionnal issue affecting young and embedded clusters is that star-forming regions are sometimes known to present anomalous reddening laws that differ from the general interstellar medium (e.g. Feinstein et al. 1973; Vazquez et al. 1996; Hur et al. 2012; Kumar et al. 2014), while the present study employs the same fixed reddening law for all clusters. However, Jordi et al. (2010) remark that varying the extinction law within the range reported by Fitzpatrick & Massa (2007) has a negligible effect on the *Gaia* photometry.

Another promising approach to deriving cluster ages is the analysis of stellar rotation (so-called gyrochronology, Barnes 2007), which presents the advantage of allowing age estimates for main sequence stars, and up to several billions of years (e.g. Meibom et al. 2015; Douglas et al. 2019). A spectacular application of this method is the characterisation of the recently discovered Pisces-Eridanus stream (Meingast et al. 2019). While it had been previously claimed (based on a single red giant with an uncertain membership status) that the structure could be 1 Gyr old, Curtis et al. (2019) show that 154 main sequence stars with available rotation periods exhibit a similar rotation pattern to the Pleiades (∼120 Myr). Curtis et al. (2019) also point out that although theoretical models have so far been unable to perfectly fit the observed loss of stellar angular momentum with age, empirical comparisons with benchmark clusters of a known age can already provide robust constraints. The Transiting Exoplanet Survey Satellite (TESS, Ricker et al. 2015) provides an all-sky survey from which light curves can be obtained, and many of its targets are cluster members (Bouma et al. 2019). In our sample, several clusters[5] are located at high Galactic latitudes and only contain late-type stars, but their ill-defined turnoff and the

[5] The "Class C" clusters UBC 605, 610, 625, 632, 642, and 649 from Castro-Ginard et al. (2020) are compact in astrometric space but their CMDs are sparse and blurry.

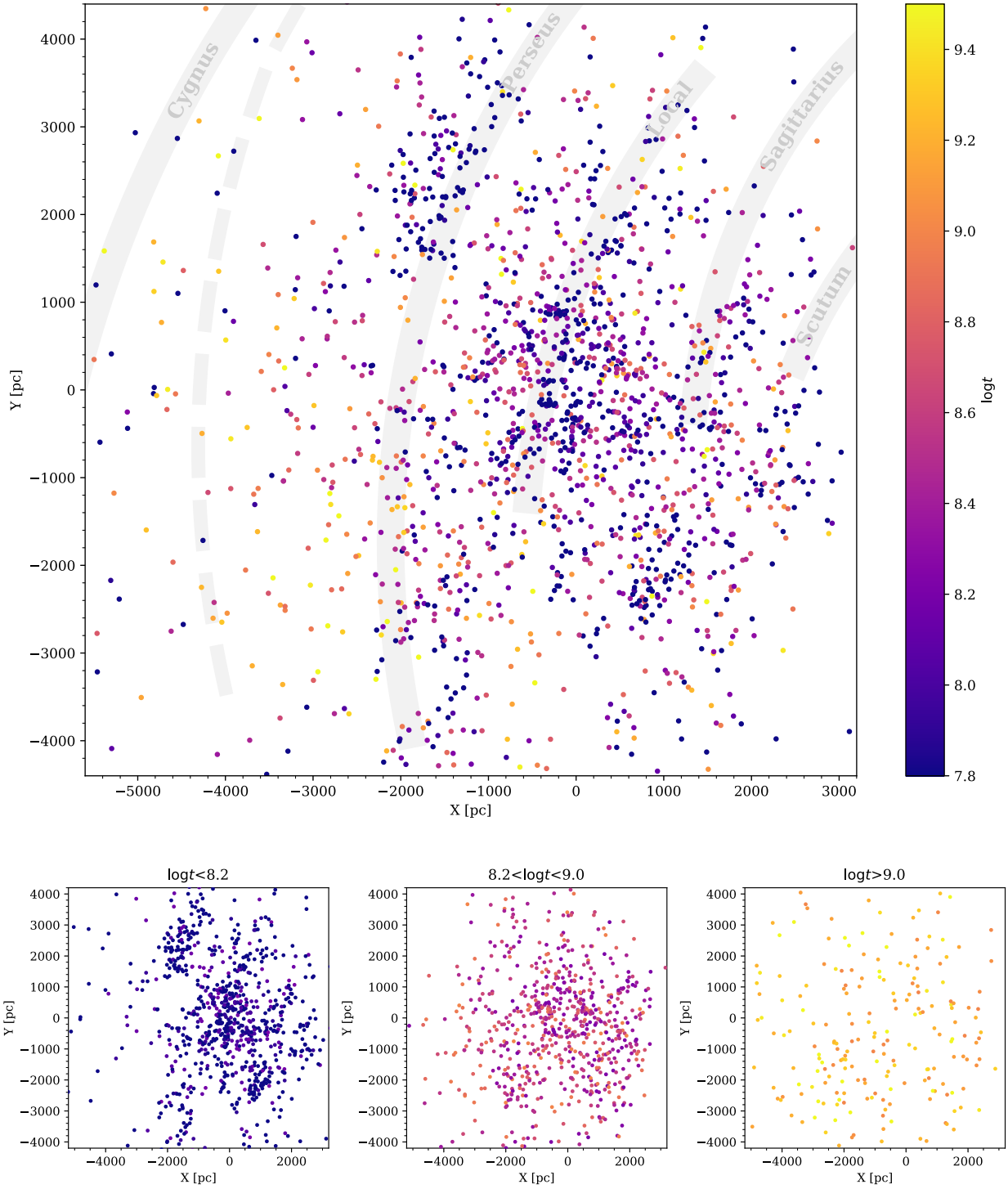


Fig. 8. Projection on the Galactic plane of the locations of clusters with derived parameters, colour-coded by age. *Top panel*: all ages. The shaded area shows the spiral arm model of Reid et al. (2014). The dashed arm is the revised path of the Cygnus arm in Reid et al. (2019). *Bottom row*: splits the sample into three age groups. The Sun is at (0,0) and the Galactic centre is to the right. The most distant objects were left out of the plot.

absence of red clump stars make it impossible to constrain their age. The increase in available training data (from e.g. TESS) and the flexibility of machine learning procedures, allowing for missing values and the empirical combination of measurements of a different nature, will make it possible to constrain the ages of such difficult objects.

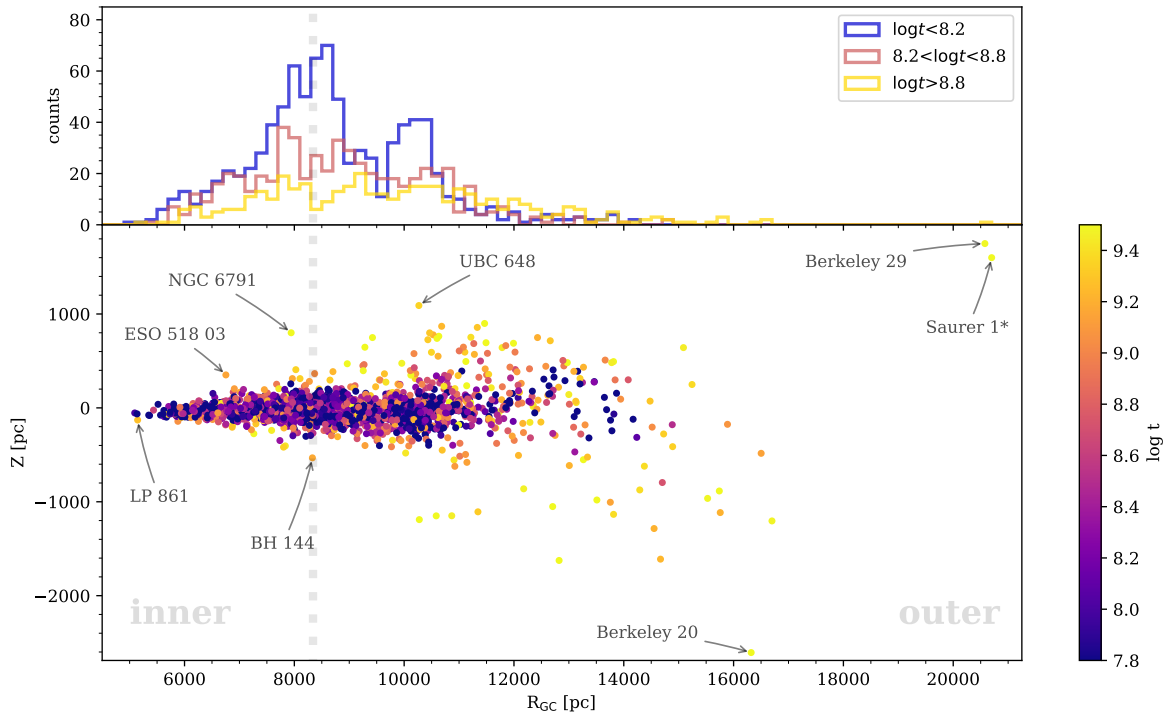


Fig. 9. *Top*: galactocentric distribution for three age groups. *Bottom*: distance from the Galactic plane against Galactocentric distance, colour-coded by age, for the clusters with derived parameters. The vertical dotted line shows the assumed Solar value of $R_{GC} = 8340$ pc (Reid et al. 2014). Our catalogue lacks Saurer 1 members, so we took its parameters from Carraro & Baume (2003).

5. Galactic structure

Using the derived distance modulus, we computed the (X, Y, Z) cartesian coordinates[6] of all clusters with available parameters. We show the projection of the cluster distribution on the Galactic plane in Fig. 8. We also computed the Galactocentric radius R_{GC}, assuming a Solar Galactocentric distance of 8340 pc[7], which is the value adopted by the spiral arm model of Reid et al. (2014). The R_{GC} versus Z distribution is shown in Fig. 9.

We show the distribution of extinction in Fig. 10. The sample of known clusters reaches much larger distances in the direction of the outer disc, especially for objects located far above the plane, but it is still limited by interstellar reddening at low Galactic latitudes.

5.1. Spiral structure

The spatial distribution of young clusters is known to correlate with the location of the spiral arms in the Milky Way (Morgan et al. 1953; Becker & Fenkart 1970; Dias & Lépine 2005). The projection of the cluster distribution is shown in Fig. 8. Its general aspect is similar to Fig. 11 in Cantat-Gaudin et al. (2018b), where groups of young clusters distribute preferentially along the locations of spiral arms delineated by Reid et al. (2014), but with important gaps and discontinuities.

In the region covered by the present study, the updated spiral arm model of Reid et al. (2019) is virtually identical. Most differences affect the first Galactic quadrant at distances that our sample of clusters does not reach, with the notable exception of the outer, Cygnus, arm. For this arm, Reid et al. (2019) fitted a significantly different location with a pitch angle of 3° and $R_{GC} \sim 11$ to 13 kpc in the anticentre direction (compared to 13.8° and 13 to 15 kpc in Reid et al. 2014). We show the revised arm as a dashed line in Fig. 8.

Our sample of *Gaia*-confirmed clusters only contains very few objects with $R_{GC} > 12$ kpc. The top panel of Fig. 9 exhibits two clearly visible peaks in the young cluster distribution, corresponding to the local arm and the Perseus arm. The Cygnus arm is not visible due to the lack of available tracers. Camargo et al. (2015) estimated the distance to several embedded clusters that were identified in WISE infrared images (Wright et al. 2010), and they propose that they trace the Cygnus arm at a Galactocentric distance of 13.5 to 15.5 kpc, which agrees with the more distant (Reid et al. 2014) model.

It has been noted (e.g. Vázquez et al. 2008) that the Perseus arm traced by clusters appears to be interrupted in the Galactic longitude range of ~140°–160°. Many clusters have been discovered in the Perseus arm region in the past decade, including two dedicated searches in *Gaia* DR2 (Cantat-Gaudin et al. 2019; Castro-Ginard et al. 2019), but all of them were found around the gap, rather than inside of it. This region of low density is visible around $(X, Y) = (-2000\,\text{pc} + 1000\,\text{pc})$ in the maps displayed in Fig. 8.

A natural explanation for the lack of detected objects in this direction could be that our view is obscured by interstellar dust, but this range of Galactic longitude does not correspond to a known region of high extinction (e.g. Lallement et al. 2019). The

[6] The Sun is at the origin. We note that X increases towards the Galactic centre, Y is in the direction of Galactic rotation, and Z is in the direction of the Galactic north pole.

[7] The most precise and recent estimate (Gravity Collaboration 2019) proposes a slightly smaller radius of ~8180 pc.

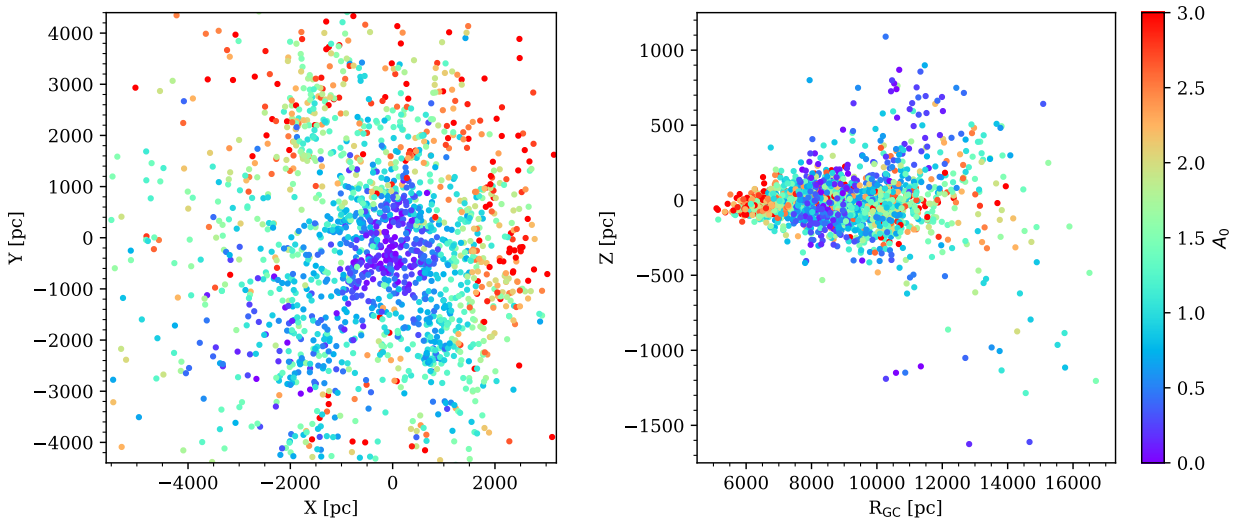


Fig. 10. Distribution of clusters in Galactic XY coordinates (*left*) and altitude versus Galactocentric radius (*right*), colour-coded by extinction A_0. In both panels, a few distant outliers were left out of the plotting window.

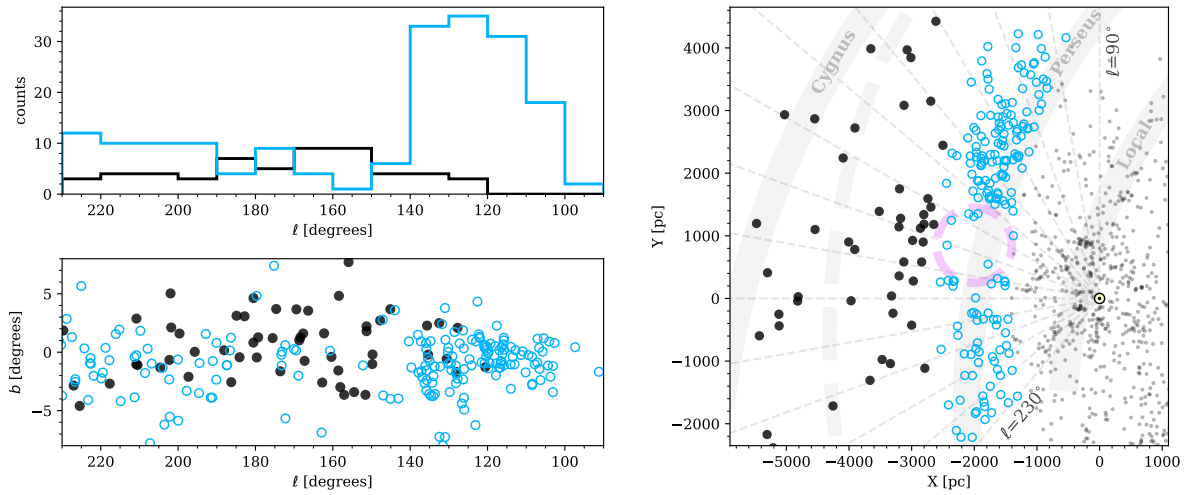


Fig. 11. *Top left*: distribution of Perseus arm clusters (arbitrarily selected as 10 kpc $< R_{GC} <$ 11 kpc) in cyan, and more distant clusters in black. *Bottom left*: galactic coordinates of the same clusters. *Right*: locations of the same clusters projected on the Galactic plane. The dashed circle has a diameter of 1200 pc.

strongest argument against extinction being responsible for this gap is illustrated in Fig. 11. While the number of clusters located at the distance of the Perseus arm drops for $\ell \sim 140°$ to 160°, the number of known clusters located behind the arm increases. It can be seen in Fig. 10 that the clusters located beyond the gap are only moderately reddened, with values of $A_0 \sim 1.5$ mag. This in fact suggests that the Perseus gap is a window of relatively lower extinction.

This gap is visible in the distribution of other young tracers, which are traditionally associated with spiral arms, and is in fact present in the HII map of Becker & Fenkart (1970) as well as in the HI map of Spicker & Feitzinger (1986), although the authors do not comment on it. The distribution of HII regions used by Hou & Han (2014) to trace the spiral structure is interrupted in the same region, and the gap can be seen (tentatively) in the Cepheid distribution of Skowron et al. (2019) as well as in the OB stars shown by Romero-Gómez et al. (2019), Poggio et al. (2018), and Jardine et al. (2019), and the high-mass star-forming regions of Reid et al. (2014, 2019).

A possibly similar gap, which is not as clear however, can be observed in the Sagittarius arm (Fig. 8), with an under-density of young clusters around $(X, Y) = (+1000\,\mathrm{pc}, -1000\,\mathrm{pc})$. Studying clusters in kinematical space could indicate that these arms are fragmenting, which is a phenomenon routinely seen in N-body simulations (e.g. Roca-Fàbrega et al. 2013; Grand et al. 2014; Hunt et al. 2015), and this would show that the Milky Way is not a grand design spiral galaxy, but rather a flocculent one.

We also see that the interarm region between the local arm and the Perseus arm is not as clear in the third quadrant as in the second quadrant, which is in agreement with Moitinho et al. (2006) and Vázquez et al. (2008), who propose that the local arm extends towards the Perseus arm along the $\ell = 245°$ line.

The presence of young clusters in the region between the Perseus and outer arms can also be interpreted as the trace of interam spurs, as reported by Molina Lera et al. (2019) and suggested by the HII maps of Hou & Han (2014). Such features are visible in external spiral galaxies (e.g. Corder et al. 2008; Elmegreen et al. 2018) and naturally occur in numerical simulations (e.g. Shetty & Ostriker 2006; Dobbs & Bonnell 2006; Pettitt et al. 2016).

5.2. Scale height

The fact that old clusters tend to be found at higher Galactic altitudes (further away from the plane) than young clusters has been noted by numerous observers (van den Bergh 1958; van den Bergh & McClure 1980; Janes et al. 1988; Janes & Phelps 1994; Phelps et al. 1994; Friel 1995), and this is visually obvious from our Fig. 9. The main cause for the thickening of the Galactic disc is the gradual velocity scatter, which is introduced by gravitational interactions with giant molecular clouds (first theorised by Spitzer & Schwarzschild 1951, 1953), although it is now understood that the effects of the spiral structure, Galactic bar, warp, and even minor mergers have contributed to the vertical heating of the disc (see e.g. the recent study of Mackereth et al. 2019, and references therein).

Various analytical parametrisations of the vertical density distribution are used in the literature (van der Kruit 1988; Dobbie & Warren 2020). A simple form often used for the cluster distribution is the exponential profile:

$$N(Z) = \frac{1}{h_z} \exp\left(-\frac{|Z - \langle z\rangle|}{h_z}\right), \quad (1)$$

where $\langle z\rangle$ is the mean offset of the Galactic plane with respect to the Sun and the h_z parameter is called the scale height. Many authors perform a fitting of the scale height in bins of age or Galactocentric radius. Rather than binning, we modelled it with a power-law dependence on age (t) and a linear dependence on the Galactocentric radius:

$$h_z = k + a \times \left(\frac{t}{100\,\mathrm{Myr}}\right)^{\alpha} + \rho \times (R_{\mathrm{GC}} - R_{\mathrm{GC},\odot}). \quad (2)$$

We sampled the parameter space using the Markov chain Monte Carlo sampler `emcee` (Foreman-Mackey et al. 2013), with flat priors on all parameters. The resulting posterior distribution is shown in Fig. 12.

The first free parameter in our model is $\langle z\rangle$, that is, the mean altitude of the entire sample considering that the Sun sits at altitude 0. The best fit value is $\langle z\rangle = -23 \pm 3$ pc, corresponding to a solar displacement of $z_0 = 23 \pm 3$ pc. This value is in line with estimates from star counts from Jurić et al. (2008; 25 ± 5 pc), Chen et al. (1999; 28 ± 6 pc), Chen et al. (2001; 27 ± 4 pc), or Maíz-Apellániz (2001; 24 ± 2 pc), for instance. We remark that studies making use of young tracers tend to report a slightly smaller solar displacement, which can be seen in Karim & Mamajek (2017; 17 ± 2 pc) or Reed (2006; 19.6 ± 2.1 pc), for instance, and previous estimates based on clusters such as in Buckner & Froebrich (2014; 18.5 ± 1.2 pc) and Joshi (2007; 13 to 20 pc) reported smaller values. The altitude of the Galactic mid-plane is known to vary with Galactocentric radius (sometimes called corrugation, see e.g. Gum et al. 1960; Lockman 1977; Spicker & Feitzinger 1986; Cantat-Gaudin et al. 2018b), which might be an additional reason why different samples yield slightly different values[8].

[8] We refer the interested reader to Karim & Mamajek (2017), who compiled a list of over 60 estimates published since 1918.

In the Solar neighbourhood, where the typical cluster age is ∼100 Myr, the cluster scale height of the best-fit model is 74 ± 5 pc, which is marginally compatible with the 64 ± 2 pc of Joshi et al. (2016). Our best-fit value of $\rho = 0.016 \pm 0.003$ (18 pc per kpc) is in good agreement with the value of 0.02 reported by Buckner & Froebrich (2014).

We also find that the scale height increases to several hundreds of parsecs for old clusters (also reported by Janes & Phelps 1994; Froebrich et al. 2010; Buckner & Froebrich 2014), with a power-law index of $\alpha = 1.3 \pm 0.2$. The mechanism often invoked to explain the steeper increase at higher ages is that clusters whose orbits do not reach high Z are destroyed at higher rates, which is due to crossing paths more often with giant molecular clouds (Moitinho 2010; Buckner & Froebrich 2014). Friel (1995) remarked that some old clusters reach such high altitudes that the encounter responsible for perturbing their orbit would likely disrupt the cluster in the process. Although Gustafsson et al. (2016) have shown that some clusters might survive such strong perturbations, there are no quantitative arguments to support that this mechanism is the only reason for the increase in scale height.

The phenomenon of heating has been studied more thoroughly for field stars than for clusters, but almost all studies have been performed in velocity space rather than positional space, making direct comparisons difficult. The time dependence of the vertical velocity dispersion in the Solar neighbourhood is often modelled as a power law ($\sigma_v \propto t^{\alpha}$). Theoretical models predict values of $\alpha < 0.3$ (Hänninen & Flynn 2002), while observations of field stars suggest an age exponent of $\alpha \sim 0.5$ (e.g. Wielen 1977; Holmberg et al. 2009; Aumer et al. 2016; Sharma et al. 2020), showing that other mechanisms have contributed to vertical heating such as mergers (Martig et al. 2014) or more efficient scattering by giant molecular clouds in the young Milky Way (Ting & Rix 2019). We refer the interested reader to Sect. 5.3.2 of Bland-Hawthorn & Gerhard (2016), who discuss recent estimates of the age-velocity dispersion relation.

The age-scale height relation we derive in this study cannot be directly compared to the age-velocity relation. It is not clear how a power-law increase of index 0.5 in velocity dispersion translates in positional space. The details of the relation between maximum velocity and maximum excursion from the Galactic plane (Z_{max}) depend on the assumed Galactic potential. For the `MWPotential2014` which was shipped with `galpy` (Bovy 2015), the relation is close to $Z_{\mathrm{max}} \propto v^{1.3}$, implying a steeper time dependence than a power law of index 0.5.

Radial migration and heating can also cause clusters to reach higher altitudes: Due to the shallower potential of the outer disc, their vertical velocity allows particles to reach larger excursion from the plane when their guiding radius is shifted outwards. If inward-migrating clusters are destroyed at higher rates than outward-migrating clusters (as suggested by e.g. Anders et al. 2017), then the mean Galactocentric radius and mean altitude of surviving clusters is expected to increase with age. Radial heating also contributes because particles on elliptical orbits reach higher altitudes near their apocentre.

The scale height of very young clusters appears to be rather large in the outer disc, with several of our clusters younger than 200 Myr reaching altitudes of 300 pc. Although the distances of these distant objects are less precise than for more nearby clusters, these results are compatible with the infrared findings of Camargo et al. (2015), who report seven embedded, and therefore very young, clusters that are further than 500 pc from the Galactic plane at $R_{\mathrm{GC}} \sim 14$ kpc. Our simple model assumes a linear increase in the scale height with Galactocentric radius,

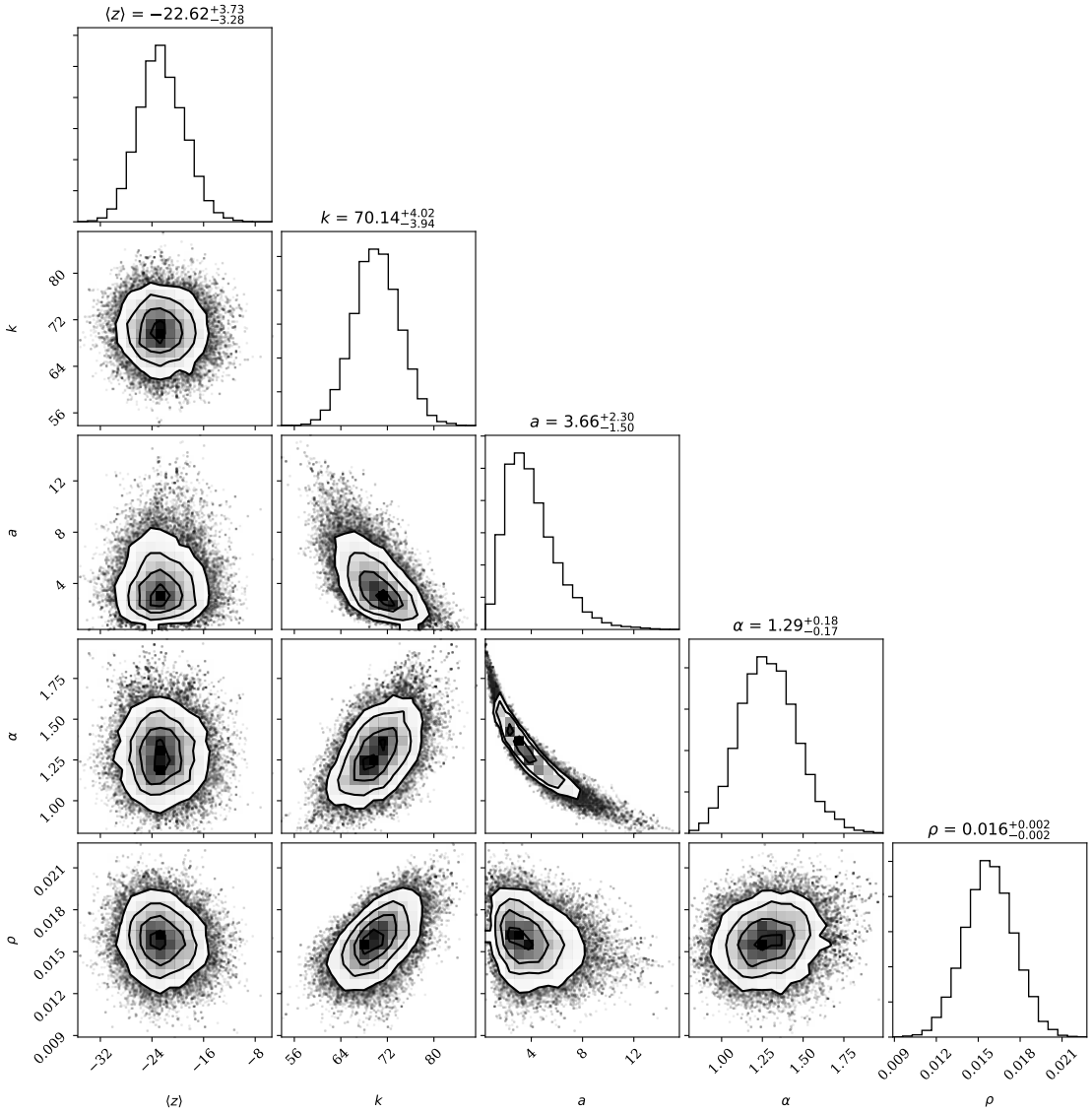


Fig. 12. Markov chain Monte Carlo sampling of the posterior distribution for the scale height model presented in Sect. 5.2, showing the last 2000 iterations of 32 walkers (64 000 points).

but Kalberla et al. (2007) and Kalberla & Dedes (2008), who could trace atomic hydrogen out to much larger distances than our cluster sample, show that the flaring of HI gas outside the Solar circle is better reproduced with an exponential function, and Wang et al. (2018) used a quadratic function.

Finally, if cluster disruption rates are lower in the outer disc, one would also expect scattering rates to be lower. Mathematically, this could be modelled by modifying Eq. (2) to allow the index α to vary with $R_{\rm GC}$. Including radial migration, heating, and disruption rates varying with $R_{\rm GC}$ and Z would make the model overly complicated and poorly constrained, with highly degenerate parameters.

Characterising the velocity distribution of clusters is out of the scope of this paper, but it would provide further insight on the processes of migration, heating, and disruption. Detailed chemical studies through high-resolution spectroscopy can also shed light on the origin of clusters. The old, metal-rich object NGC 6791 is a well-known case of a cluster migrating from the inner disc (Jílková et al. 2012; Carraro 2014; Martinez-Medina et al. 2018), but lesser-known or newly discovered clusters with discrepant altitudes (such as BH 144 or UBC 648, labelled in Fig. 9) might also shown evidence for radial migration.

5.3. Galactic warp

The Galactic mid-plane is known to deviate from the geometrical $b = 0°$ plane in the outer disc, which is a phenomenon called warp. The warping of the Galactic plane is particularly visible in the HI gas distribution (Burke 1957; Kerr 1957; Westerhout 1957; Levine et al. 2006; Kalberla et al. 2007) and is now known to be a common feature in disc galaxies (e.g. Sancisi 1976;

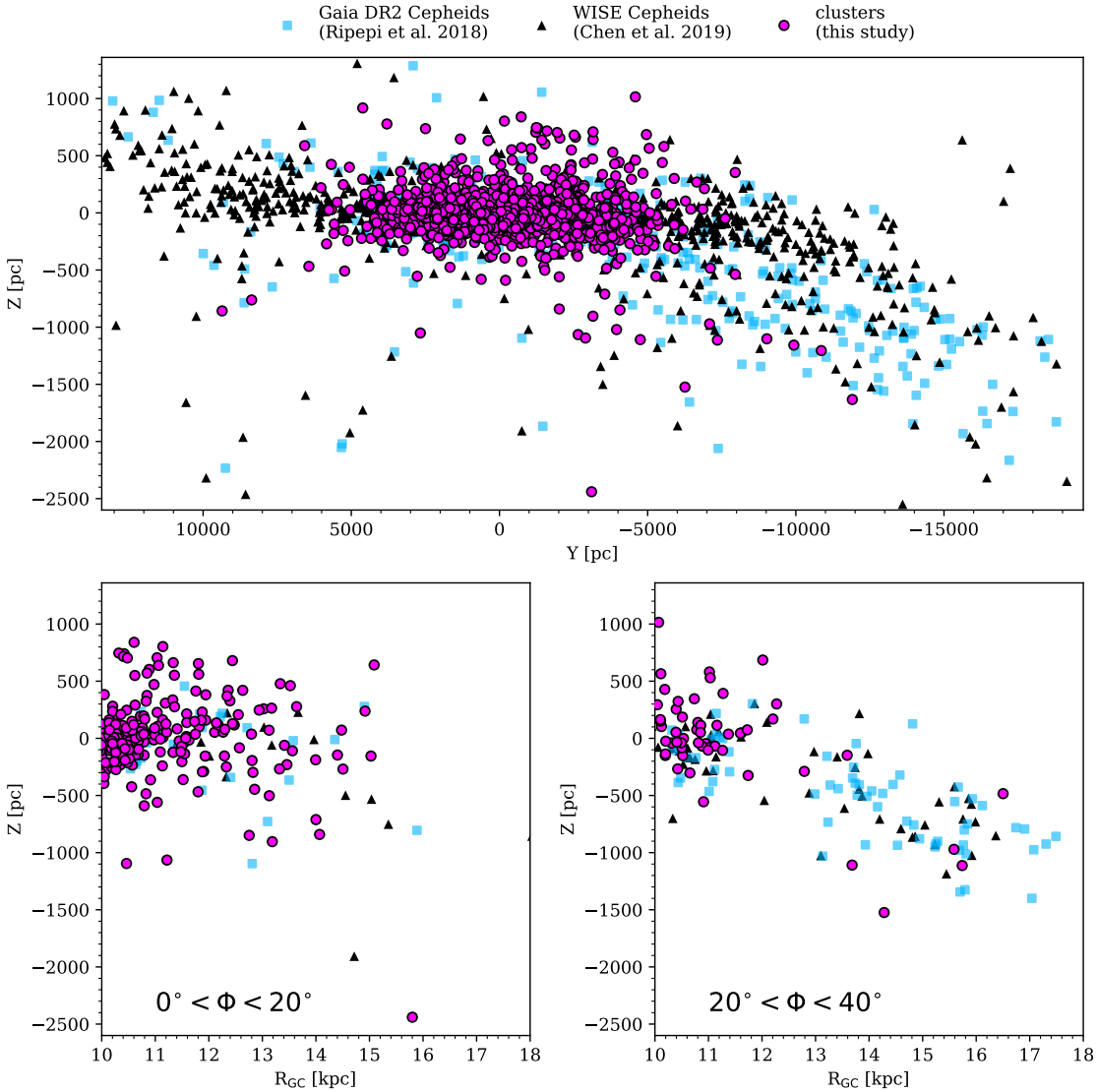


Fig. 13. *Top*: Y versus Z coordinates of our cluster sample and the Cepheids from Ripepi et al. (2019) and Chen et al. (2019). *Bottom*: galactocentric distance versus altitude Z in two ranges of Galactocentric angular coordinates, both in the third Galactic quadrant.

Briggs 1990; Sánchez-Saavedra et al. 2003). The warp is also visible in the distribution of molecular clouds (Wouterloot et al. 1990), dust (Marshall et al. 2006), stars (López-Corredoira et al. 2002; Moitinho et al. 2006; Vázquez et al. 2008; Reylé et al. 2009; Amôres et al. 2017; Chrobakova et al. 2020), and stellar kinematics (Poggio et al. 2018; Schönrich & Dehnen 2018); additionally, it was recently investigated by tracing the distribution of classical Cepheids (Skowron et al. 2019; Chen et al. 2019). These young (~20 to 120 Myr: Efremov 1978; Bono et al. 2005; Senchyna et al. 2015) and bright stars are visible at large distances and allow for precise distance determinations.

In Fig. 13 we compare the location of known clusters with classical Cepheids. The lower panels only include tracers in two bins of Galactocentric angular coordinates Φ, where Φ = 0° is the line passing through the Galactic centre and the Solar location, and Φ increases in the opposite direction to Galactic rotation (convention used in e.g. Ripepi et al. 2019; Skowron et al. 2019). The distant clusters in the third Galactic quadrant are on average older than 1 Gyr, and they follow the same southward trend as the young Cepheids. The number of known distant clusters is unfortunately too small to allow us to verify whether the Cepheid warp and the old cluster warp still coincide for Φ > 40°. In particular, no known clusters are located in the region of the northern warp.

6. Discussion

Among the clusters for which we can derive parameters, the closest to the Galactic centre is Ruprecht 126 ($\log t = 8.11$; $R_{GC} = 5230$ pc). Several known clusters might be located even deeper in the disc, according to their small parallax and apparent location, but their CMDs are too sparse and blurry to allow us to derive meaningful parameters and to constrain their distance with photometry. The deepest known clusters would be BH 222

(also studied by Piatti & Clariá 2002) and Gulliver 41, both of which are at $R_{GC} < 3$ kpc and lack estimated parameters in our catalogue.

We label in Fig. 9 several old clusters that stand out as outliers. One of them is the well-studied NGC 6791, an old metal-rich cluster whose likely origin is the bulge or the inner disc. Berkeley 20, Berkeley 29, and Saurer 1 are also well-known distant objects, which are currently located far from the Galactic plane. The object UBC 648 is a recently discovered cluster (Castro-Ginard et al. 2020), and it is also located far from the Galactic plane.

The cluster LP 861 was only recently discovered (Liu & Pang 2019) and is one of the innermost old clusters known. Other intermediate-age or old clusters were recently identified in the *Gaia* DR2 data, such as UBC 307, UBC 310, UBC 339, LP 866, and UFMG 2, which are all located at $R_{GC} < 6.5$ kpc and at very low altitudes. The only such objects known before *Gaia* were NGC 6005 (Piatti et al. 1998), NGC 6583 (Carraro et al. 2005), Ruprecht 134 (Carraro et al. 2006), and Teutsch 84 (Kronberger et al. 2006). These objects deserve further investigation in order to understand how they can survive to reach old ages in such a dense environment. They might be on very elliptical orbits, have recently migrated inwards, or their initial mass may have been sufficient for them to remain gravitationally bound.

We cannot presently probe the structure of the outer disc (e.g. the trace of the Cygnus arm or the geometry of the warp) with the sample of clusters identified in *Gaia* (with $G < 18$). As is visible in Fig 9, very few clusters are known at $R_{GC} > 14$ kpc and no clusters are known beyond 16.5 kpc, with the exception of Berkeley 29 and Saurer 1 which are near $R_{GC} \sim 20$ kpc. This lack of available tracers is due, at least in part, to an obscured line of sight preventing us from identifying distant objects near the Galactic plane. A near-infrared *Gaia*-like mission (Hobbs et al. 2016, 2019) would allow us to see through dust clouds and reveal obscured structures and embedded clusters. The upcoming ground-based LSST (LSST Science Collaboration 2009; Ivezić et al. 2019) will reach stars seven magnitudes fainter than *Gaia*, and it is expected to provide proper motions better than 1 mas yr^{-1} down to $G \sim 24$, allowing one to push the boundaries of cluster detection further than presently possible.

We note however that the distant outer disc clusters, especially in the third quadrant, are not strongly affected by extinction (Fig. 10). This suggests that the drop in density is not just an observational bias, but also a sign that few clusters populate the distant outer disc. Stellar population studies typically locate the disc truncation radius near 14 kpc (Robin et al. 1992), 15 kpc (Ruphy et al. 1996), or 16 kpc (Amôres et al. 2017). Due to the uncertainty on the completeness of our sample in the outer disc, we did not attempt to fit a radial density profile or try to identify a cut-off Galactocentric radius, but the observed cluster distribution visually agrees with a cut-off point near 14 kpc. The objects Berkeley 29 and Saurer 1, which are on the far edge of the disc, would therefore be outliers on very perturbed orbits, rather than representants of a cluster population forming at extreme Galactocentric distances. On the other hand, several distant disc clusters were recently discovered with a combination of *Gaia* data and deep ground-based photometry by authors searching for satellite systems (Koposov et al. 2017; Torrealba et al. 2019). The lack of clusters beyond $R_{GC} \sim 16$ kpc could therefore be an observational bias that future studies will be able to fill in.

This study focuses on the present-day location of clusters. The *Gaia* DR2 catalogue also allows us to determine proper motions for all of them and, therefore, estimate tangential velocities. Soubiran et al. (2018) have obtained mean radial velocities for several hundreds of clusters using the *Gaia* Radial Velocity Spectrometer (Cropper et al. 2018) and shown a smooth increase in vertical velocity dispersion with age. Further insight can be gathered by supplementing the scarce *Gaia* radial velocities with observations from other surveys (e.g. Carrera et al. 2019, with APOGEE and GALAH data). Although *Gaia* DR3 will contain significantly more radial velocities than DR2 (Brown 2019), the *Gaia* spacecraft only has limited spectroscopic capabilities. Ground-based spectroscopic surveys such as APOGEE (Majewski et al. 2017), *Gaia*-ESO (Gilmore et al. 2012; Randich et al. 2013), GALAH (De Silva et al. 2015), LAMOST (Cui et al. 2012), or the upcoming WEAVE (Dalton et al. 2012) and 4MOST (de Jong et al. 2012; Guiglion et al. 2019) will provide additional observations allowing for the full characterisation of the 3D velocities of many more objects, and they will shed light on the dynamical processes that drive the evolution of the spiral structure and the heating of the Galactic disc.

7. Summary and conclusion

This study relies almost exclusively on *Gaia* DR2 data. We characterise clusters whose members were identified with *Gaia* astrometry. We use an artificial neural network to estimate the age, distance modulus, and interstellar extinction of each cluster from the *Gaia* photometry of its members and their mean *Gaia* parallax. The training set was built using observed clusters with reliable parameters.

After visually inspecting the colour-magnitude diagrams and verifying the consistency of the parameter estimates, we end up with 1867 clusters with reliable parameters. The 3D distribution of clusters traces the structure of the Galactic disc, with warping and flaring in the outer disc. We clearly observe the known increase in cluster scale height with age. Various mechanisms contribute to this increase, and the current cluster locations are not sufficient at disentangling the effects of heating, migration, and location-dependent disruption rates. Establishing the 3D velocity vector and characterising the orbital parameters of clusters and their dependence with age will provide further insight on the evolutionary history of the Milky Way.

Projected on the Galactic plane, the locations of young clusters roughly align along the expected spiral pattern, and especially the local and Perseus arms. We argue that the apparent interruption in the Perseus arm is physical, and it is not due to an observational bias introduced by interstellar extinction. More kinematical data is needed in order to determine whether the Perseus arm is in the process of fragmenting. Our present sample does not contain a sufficient number of distant clusters to trace the path of the outer arm or constrain the geometry of the warp in the outer disc.

The catalogue presented in this paper is the largest homogeneous analysis of cluster parameters performed with *Gaia* data so far, with almost two thousand objects. The continuous discovery of new clusters and the development of data-driven methods that are capable of including other photometric passbands, astrophysical parameters from value-added catalogues, or rotation periods will allow for more precise and accurate cluster parameter estimates as well as a consistent account of observational errors.

Acknowledgements. We thank the referee for useful suggestions that helped clarify this paper. This work has made use of data from the European Space Agency (ESA) mission *Gaia* (www.cosmos.esa.int/gaia), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, 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 work was supported by the MINECO (Spanish Ministry of Economy) through grant ESP2016-80079-C2-1-R and RTI2018-095076-B-C21 (MINECO/FEDER, UE), and MDM-2014-0369 of ICCUB (Unidad de Excelencia "María de Maeztu"). TCG acknowledges support from Juan de la Cierva – Formación 2015 grant, MINECO (FEDER/UE). FA is grateful for funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 800502. AM acknowledges the support from the Portuguese Strategic Programme UID/FIS/00099/2019 for CENTRA. AV and AB acknowledge PREMIALE 2015 MITiC. DB is supported in the form of work contract FCT/MCTES through national funds and by FEDER through COMPETE2020 in connection to these grants: UID/FIS/04434/2019; PTDC/FIS-AST/30389/2017 & POCI-01-0145-FEDER-030389. The preparation of this work has made extensive use of Topcat (Taylor 2005), and of NASA's Astrophysics Data System Bibliographic Services, as well as the open-source Python packages `Astropy` (Astropy Collaboration 2013), `NumPy` (Van Der Walt et al. 2011), and `scikit-learn` (Pedregosa et al. 2011). The figures in this paper were produced with Matplotlib (Hunter 2007). Figure 12 was produced with `corner` (Foreman-Mackey 2016).

References

Ahumada, A. V., Cignoni, M., Bragaglia, A., et al. 2013, MNRAS, 430, 221
Amôres, E. B., Robin, A. C., & Reylé, C. 2017, A&A, 602, A67
Anders, F., Chiappini, C., Minchev, I., et al. 2017, A&A, 600, A70
Anders, F., Khalatyan, A., Chiappini, C., et al. 2019, A&A, 628, A94
Andrae, R., Fouesneau, M., Creevey, O., et al. 2018, A&A, 616, A8
Andreuzzi, G., Bragaglia, A., Tosi, M., & Marconi, G. 2011, MNRAS, 412, 1265
Anthony-Twarog, B. J., & Twarog, B. A. 1985, ApJ, 291, 595
Arenou, F., Luri, X., Babusiaux, C., et al. 2018, A&A, 616, A17
Arnason, R. M., Barmby, P., & Vulic, N. 2020, MNRAS, 492, 5075
Astropy Collaboration (Robitaille, T. P., et al.) 2013, A&A, 558, A33
Aumer, M., Binney, J., & Schönrich, R. 2016, MNRAS, 462, 1697
Barnes, S. A. 2007, ApJ, 669, 1167
Baron, D. 2019, ArXiv e-prints [arXiv:1904.07248]
Barrado y Navascués, D., Stauffer, J. R., & Jayawardhana, R. 2004, ApJ, 614, 386
Bastian, N., Kamann, S., Cabrera-Ziri, I., et al. 2018, MNRAS, 480, 3739
Becker, W., & Fenkart, R. B. 1970, in The Spiral Structure of our Galaxy, eds. W. Becker, & G. I. Kontopoulos, IAU Symp., 38, 205
Bland-Hawthorn, J., & Gerhard, O. 2016, ARA&A, 54, 529
Bono, G., Marconi, M., Cassisi, S., et al. 2005, ApJ, 621, 966
Bossini, D., Vallenari, A., Bragaglia, A., et al. 2019, A&A, 623, A108
Boucaud, A., Huertas-Company, M., Heneka, C., et al. 2020, MNRAS, 491, 2481
Bouma, L. G., Hartman, J. D., Bhatti, W., Winn, J. N., & Bakos, G. Á. 2019, ApJS, 245, 13
Bovy, J. 2015, ApJS, 216, 29
Bragaglia, A., & Tosi, M. 2006, AJ, 131, 1544
Bragaglia, A., Tosi, M., Andreuzzi, G., & Marconi, G. 2006, MNRAS, 368, 1971
Bressan, A., Marigo, P., Girardi, L., et al. 2012, MNRAS, 427, 127
Briggs, F. H. 1990, ApJ, 352, 15
Brown, A. G. A. 2019, https://doi.org/10.5281/zenodo.2637972
Buckner, A. S. M., & Froebrich, D. 2014, MNRAS, 444, 290
Burke, B. F. 1957, AJ, 62, 90
Camargo, D., Bonatto, C., & Bica, E. 2015, MNRAS, 450, 4150
Cantat-Gaudin, T., & Anders, F. 2020, A&A, 633, A99
Cantat-Gaudin, T., Vallenari, A., Zaggia, S., et al. 2014, A&A, 569, A17
Cantat-Gaudin, T., Vallenari, A., Sordo, R., et al. 2018a, A&A, 615, A49
Cantat-Gaudin, T., Jordi, C., Vallenari, A., et al. 2018b, A&A, 618, A93
Cantat-Gaudin, T., Krone-Martins, A., Sedaghat, N., et al. 2019, A&A, 624, A126
Carraro, G. 2014, in Properties and Origin of the Old, Metal Rich, Star Cluster, NGC 6791, eds. H. W. Lee, Y. W. Kang, & K. C. Leung, ASP Conf. Ser., 482, 245
Carraro, G., & Baume, G. 2003, MNRAS, 346, 18
Carraro, G., & Chiosi, C. 1994, A&A, 287, 761
Carraro, G., Méndez, R. A., & Costa, E. 2005, MNRAS, 356, 647
Carraro, G., Janes, K. A., Costa, E., & Méndez, R. A. 2006, MNRAS, 368, 1078
Carrera, R., Bragaglia, A., Cantat-Gaudin, T., et al. 2019, A&A, 623, A80
Castro-Ginard, A., Jordi, C., Luri, X., et al. 2018, A&A, 618, A59
Castro-Ginard, A., Jordi, C., Luri, X., Cantat-Gaudin, T., & Balaguer-Núñez, L. 2019, A&A, 627, A35
Castro-Ginard, A., Jordi, C., Luri, X., et al. 2020, A&A, 635, A45
Chen, B., Figueras, F., Torra, J., et al. 1999, A&A, 352, 459
Chen, B., Stoughton, C., Smith, J. A., et al. 2001, ApJ, 553, 184
Chen, X., Wang, S., Deng, L., et al. 2019, Nat. Astron., 3, 320
Chrobakova, Z., Nagy, R., & Lopez-Corredoira, M. 2020, A&A, 637, A96
Cignoni, M., Beccari, G., Bragaglia, A., & Tosi, M. 2011, MNRAS, 416, 1077
Collinder, P. 1931, Ann. Obs. Lund, 2, B1
Corder, S., Sheth, K., Scoville, N. Z., et al. 2008, ApJ, 689, 148
Cropper, M., Katz, D., Sartoretti, P., et al. 2018, A&A, 616, A5
Cui, X.-Q., Zhao, Y.-H., Chu, Y.-Q., et al. 2012, Res. Astron. Astrophys., 12, 1197
Curtis, J. L., Agüeros, M. A., Mamajek, E. E., Wright, J. T., & Cummings, J. D. 2019, AJ, 158, 77
Dalton, G., Trager, S. C., Abrams, D. C., et al. 2012, in WEAVE: The Next Generation Wide-field Spectroscopy Facility for the William Herschel Telescope, SPIE Conf. Ser., 8446, 84460P
Danielski, C., Babusiaux, C., Ruiz-Dern, L., Sartoretti, P., & Arenou, F. 2018, A&A, 614, A19
de Jong, R. S., Bellido-Tirado, O., Chiappini, C., et al. 2012, in 4MOST: 4-metre Multi-object Spectroscopic Telescope, SPIE Conf. Ser., 8446, 84460T
de Juan Ovelar, M., Gossage, S., Kamann, S., et al. 2020, MNRAS, 491, 2129
De Silva, G. M., Freeman, K. C., Bland-Hawthorn, J., et al. 2015, MNRAS, 449, 2604
Delgado, A. J., Sampedro, L., Alfaro, E. J., et al. 2016, MNRAS, 460, 3305
Dias, W. S., & Lépine, J. R. D. 2005, ApJ, 629, 825
Dias, W. S., Alessi, B. S., Moitinho, A., & Lépine, J. R. D. 2002, A&A, 389, 871
Dobbie, P., & Warren, S. J. 2020, ArXiv e-prints [arXiv:2003.05757]
Dobbs, C. L., & Bonnell, I. A. 2006, MNRAS, 367, 873
Donati, P., Bragaglia, A., Cignoni, M., Cocozza, G., & Tosi, M. 2012, MNRAS, 424, 1132
Donati, P., Beccari, G., Bragaglia, A., Cignoni, M., & Tosi, M. 2014a, MNRAS, 437, 1241
Donati, P., Cantat Gaudin, T., Bragaglia, A., et al. 2014b, A&A, 561, A94
Donati, P., Bragaglia, A., Carretta, E., et al. 2015, MNRAS, 453, 4185
Douglas, S. T., Curtis, J. L., Agüeros, M. A., et al. 2019, ApJ, 879, 100
Efremov, I. N. 1978, Sov. Astron., 22, 161
Elmegreen, B. G., Elmegreen, D. M., & Efremov, Y. N. 2018, ApJ, 863, 59
Evans, D. W., Riello, M., De Angeli, F., et al. 2018, A&A, 616, A4
Feinstein, A., Marraco, H. G., & Muzzio, J. C. 1973, A&AS, 12, 331
Fitzpatrick, E. L., & Massa, D. 2007, ApJ, 663, 320
Fluke, C. J., & Jacobs, C. 2020, WIREs Data Mining and Knowledge Discovery, 10, e1349
Foreman-Mackey, D. 2016, J. Open Sour. Softw., 1, 24
Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, PASP, 125, 306
Friel, E. D. 1995, ARA&A, 33, 381
Friel, E. D., Donati, P., Bragaglia, A., et al. 2014, A&A, 563, A117
Froebrich, D., Schmeja, S., Samuel, D., & Lucas, P. W. 2010, MNRAS, 409, 1281
Gaia Collaboration (Brown, A. G. A., et al.) 2018a, A&A, 616, A1
Gaia Collaboration (Babusiaux, C., et al.) 2018b, A&A, 616, A10
Gilmore, G., Randich, S., Asplund, M., et al. 2012, The Messenger, 147, 25
Grand, R. J. J., Kawata, D., & Cropper, M. 2014, MNRAS, 439, 623
Gravity Collaboration (Abuter, R., et al.) 2019, A&A, 625, L10
Guiglion, G., Battistini, C., Bell, C. P. M., et al. 2019, The Messenger, 175, 17
Gum, C. S., Kerr, F. J., & Westerhout, G. 1960, MNRAS, 121, 132
Gustafsson, B., Church, R. P., Davies, M. B., & Rickman, H. 2016, A&A, 593, A85
Hänninen, J., & Flynn, C. 2002, MNRAS, 337, 731
Hatzidimitriou, D., Held, E. V., Tognelli, E., et al. 2019, A&A, 626, A90
Heiter, U., Soubiran, C., Netopil, M., & Paunzen, E. 2014, A&A, 561, A93
Herschel, W. 1785, Philos. Trans. R. Soc. London Ser. I, 75, 213
Hobbs, D., Høg, E., Mora, A., et al. 2016, ArXiv e-prints [arXiv:1609.07325]
Hobbs, D., Brown, A., Høg, E., et al. 2019, ArXiv e-prints [arXiv:1907.12535]
Holmberg, J., Nordström, B., & Andersen, J. 2009, A&A, 501, 941
Hou, L. G., & Han, J. L. 2014, A&A, 569, A125
Hunt, J. A. S., Kawata, D., Grand, R. J. J., et al. 2015, MNRAS, 450, 2132
Hunter, J. D. 2007, Comput. Sci. Eng., 9, 90
Hur, H., Sung, H., & Bessell, M. S. 2012, AJ, 143, 41
Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111
Janes, K., & Adler, D. 1982, ApJS, 49, 425
Janes, K. A., & Phelps, R. L. 1994, AJ, 108, 1773
Janes, K. A., Tilley, C., & Lynga, G. 1988, AJ, 95, 771
Jardine, K., Poggio, E., & Drimmel, R. 2019, https://doi.org/10.5281/zenodo.3235352
Jeffery, E. J., von Hippel, T., van Dyk, D. A., et al. 2016, ApJ, 828, 79
Jeffries, R. D., & Oliveira, J. M. 2005, MNRAS, 358, 13
Jeffries, R. D., Jackson, R. J., Franciosini, E., et al. 2017, MNRAS, 464, 1456
Jílková, L., Carraro, G., Jungwiert, B., & Minchev, I. 2012, A&A, 541, A64
Jordi, C., Gebran, M., Carrasco, J. M., et al. 2010, A&A, 523, A48
Joshi, Y. C. 2007, MNRAS, 378, 768
Joshi, Y. C., Dambis, A. K., Pandey, A. K., & Joshi, S. 2016, A&A, 593, A116
Jurić, M., Ivezić, Ž., Brooks, A., et al. 2008, ApJ, 673, 864

Kalberla, P. M. W., & Dedes, L. 2008, A&A, 487, 951
Kalberla, P. M. W., Dedes, L., Kerp, J., & Haud, U. 2007, A&A, 469, 511
Karim, M. T., & Mamajek, E. E. 2017, MNRAS, 465, 472
Kerr, F. J. 1957, AJ, 62, 93
Kharchenko, N. V., Piskunov, A. E., Schilbach, E., Röser, S., & Scholz, R. D. 2013, A&A, 558, A53
Kingma, D. P., & Ba, J. 2014, 3rd International Conference for learning representations, San Diego
Koposov, S. E., Belokurov, V., & Torrealba, G. 2017, MNRAS, 470, 2702
Kounkel, M., & Covey, K. 2019, AJ, 158, 122
Kronberger, M., Teutsch, P., Alessi, B., et al. 2006, A&A, 447, 921
Krone-Martins, A., & Moitinho, A. 2014, A&A, 561, A57
Kumar, B., Sharma, S., Manfroid, J., et al. 2014, A&A, 567, A109
Lallement, R., Babusiaux, C., Vergely, J. L., et al. 2019, A&A, 625, A135
Leung, H. W., & Bovy, J. 2019, MNRAS, 483, 3255
Levine, E. S., Blitz, L., & Heiles, C. 2006, ApJ, 643, 881
Li, C., Sun, W., de Grijs, R., et al. 2019, ApJ, 876, 65
Lindegren, L., Hernández, J., Bombrun, A., et al. 2018, A&A, 616, A2
Liu, L., & Pang, X. 2019, ApJS, 245, 32
Lockman, F. J. 1977, AJ, 82, 408
López-Corredoira, M., Cabrera-Lavers, A., Garzón, F., & Hammersley, P. L. 2002, A&A, 394, 883
LSST Science Collaboration (Abell, P. A., et al.) 2009, ArXiv e-prints [arXiv:0912.0201]
Lynga, G. 1982, A&A, 109, 213
Lyra, W., Moitinho, A., van der Bliek, N. S., & Alves, J. 2006, A&A, 453, 101
Mackereth, J. T., Bovy, J., Leung, H. W., et al. 2019, MNRAS, 489, 176
Maíz-Apellániz, J. 2001, AJ, 121, 2737
Majewski, S. R., Schiavon, R. P., Frinchaboy, P. M., et al. 2017, AJ, 154, 94
Marino, A. F., Milone, A. P., Casagrande, L., et al. 2018, ApJ, 863, L33
Marshall, D. J., Robin, A. C., Reylé, C., Schultheis, M., & Picaud, S. 2006, A&A, 453, 635
Martig, M., Minchev, I., & Flynn, C. 2014, MNRAS, 443, 2452
Martinez-Medina, L. A., Gieles, M., Pichardo, B., & Peimbert, A. 2018, MNRAS, 474, 32
Meibom, S., Barnes, S. A., Platais, I., et al. 2015, Nature, 517, 589
Meingast, S., Alves, J., & Fürnkranz, V. 2019, A&A, 622, L13
Melotte, P. J. 1915, Mem. R. Astron. Soc., 60, 175
Moitinho, A. 2010, in Star Clusters: Basic Galactic Building Blocks Throughout Time and Space, eds. R. de Grijs, & J. R. D. Lépine, IAU Symp., 266, 106
Moitinho, A., Vázquez, R. A., Carraro, G., et al. 2006, MNRAS, 368, L77
Molina Lera, J. A., Baume, G., & Gamen, R. 2019, MNRAS, 488, 2158
Monteiro, H., & Dias, W. S. 2019, MNRAS, 487, 2385
Monteiro, H., Dias, W. S., & Caetano, T. C. 2010, A&A, 516, A2
Morgan, W. W., Whitford, A. E., & Code, A. D. 1953, ApJ, 118, 318
Naylor, T., & Jeffries, R. D. 2006, MNRAS, 373, 1251
Overbeek, J. C., Friel, E. D., Donati, P., et al. 2017, A&A, 598, A68
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, J. Mach. Learn. Res., 12, 2825
Perren, G. I., Vázquez, R. A., & Piatti, A. E. 2015, A&A, 576, A6
Pettitt, A. R., Tasker, E. J., & Wadsley, J. W. 2016, MNRAS, 458, 3990
Phelps, R. L., Janes, K. A., & Montgomery, K. A. 1994, AJ, 107, 1079
Piatti, A. E., & Clariá, J. J. 2002, A&A, 388, 179
Piatti, A. E., Clariá, J. J., Bica, E., Geisler, D., & Minniti, D. 1998, AJ, 116, 801
Poggio, E., Drimmel, R., Lattanzi, M. G., et al. 2018, MNRAS, 481, L21
Queiroz, A. B. A., Anders, F., Santiago, B. X., et al. 2018, MNRAS, 476, 2556
Randich, S., Gilmore, G., & Gaia-ESO Consortium 2013, The Messenger, 154, 47
Randich, S., Tognelli, E., Jackson, R., et al. 2018, A&A, 612, A99
Reed, B. C. 2006, JRASC, 100, 146
Reid, M. J., Menten, K. M., Brunthaler, A., et al. 2014, ApJ, 783, 130
Reid, M. J., Menten, K. M., Brunthaler, A., et al. 2019, ApJ, 885, 131
Reylé, C., Marshall, D. J., Robin, A. C., & Schultheis, M. 2009, A&A, 495, 819
Ricker, G. R., Winn, J. N., Vanderspek, R., et al. 2015, J. Astron. Telesc. Instrum. Syst., 1, 014003
Ripepi, V., Molinaro, R., Musella, I., et al. 2019, A&A, 625, A14
Robin, A. C., Creze, M., & Mohan, V. 1992, ApJ, 400, L25
Roca-Fàbrega, S., Valenzuela, O., Figueras, F., et al. 2013, MNRAS, 432, 2878
Romero-Gómez, M., Mateu, C., Aguilar, L., Figueras, F., & Castro-Ginard, A. 2019, A&A, 627, A150
Ruphy, S., Robin, A. C., Epchtein, N., et al. 1996, A&A, 313, L21
Salaris, M., Weiss, A., & Percival, S. M. 2004, A&A, 414, 163
Sánchez-Saavedra, M. L., Battaner, E., Guijarro, A., López-Corredoira, M., & Castro-Rodríguez, N. 2003, A&A, 399, 457
Sancisi, R. 1976, A&A, 53, 159
Sandage, A. 1988, PASP, 100, 293
Schönrich, R., & Dehnen, W. 2018, MNRAS, 478, 3809
Senchyna, P., Johnson, L. C., Dalcanton, J. J., et al. 2015, ApJ, 813, 31
Shapley, H. 1918, ApJ, 48, 154
Sharma, S., Hayden, M. R., Bland-Hawthorn, J., et al. 2020, MNRAS, submitted [arXiv:2004.06556]
Shetty, R., & Ostriker, E. C. 2006, ApJ, 647, 997
Siegel, M. H., LaPorte, S. J., Porterfield, B. L., Hagen, L. M. Z., & Gronwall, C. A. 2019, AJ, 158, 35
Skowron, D. M., Skowron, J., Mróz, P., et al. 2019, Science, 365, 478
Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, AJ, 131, 1163
Soubiran, C., Cantat-Gaudin, T., Romero-Gómez, M., et al. 2018, A&A, 619, A155
Spicker, J., & Feitzinger, J. V. 1986, A&A, 163, 43
Spitzer, L., Jr., & Schwarzschild, M. 1951, ApJ, 114, 385
Spitzer, L., Jr., & Schwarzschild, M. 1953, ApJ, 118, 106
Sun, W., de Grijs, R., Deng, L., & Albrow, M. D. 2019, ApJ, 876, 113
Tang, B., Geisler, D., Friel, E., et al. 2017, A&A, 601, A56
Taylor, M. B. 2005, in Astronomical Data Analysis Software and Systems XIV, eds. P. Shopbell, M. Britton, & R. Ebert, ASP Conf. Ser., 347, 29
Ting, Y.-S., & Rix, H.-W. 2019, ApJ, 878, 21
Ting, Y.-S., Hawkins, K., & Rix, H.-W. 2018, ApJ, 858, L7
Torrealba, G., Belokurov, V., & Koposov, S. E. 2019, MNRAS, 484, 2181
Tosi, M., Bragaglia, A., & Cignoni, M. 2007, MNRAS, 378, 730
Trumpler, R. J. 1925, PASP, 37, 307
Trumpler, R. J. 1930, Lick Obs. Bull., 420, 154
Twarog, B. A., & Anthony-Twarog, B. J. 1989, AJ, 97, 759
van den Bergh, S. 1958, ZAp, 46, 176
van den Bergh, S., & McClure, R. D. 1980, A&A, 88, 360
van der Kruit, P. C. 1988, A&A, 192, 117
Van Der Walt, S., Colbert, S. C., & Varoquaux, G. 2011, Comput. Sci. Eng., 13, 22
Vazquez, R. A., Baume, G., Feinstein, A., & Prado, P. 1996, A&AS, 116, 75
Vázquez, R. A., May, J., Carraro, G., et al. 2008, ApJ, 672, 930
von Hippel, T., Jefferys, W. H., Scott, J., et al. 2006, ApJ, 645, 1436
Wang, H.-F., Liu, C., Xu, Y., Wan, J.-C., & Deng, L. 2018, MNRAS, 478, 3367
Westerhout, G. 1957, Bull. Astron. Inst. Neth., 13, 201
Wielen, R. 1977, A&A, 60, 263
Wouterloot, J. G. A., Brand, J., Burton, W. B., & Kwee, K. K. 1990, A&A, 230, 21
Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, AJ, 140, 1868

b

MACHINE LEARNING-BASED DETECTION OF YOUNG STELLAR OBJECTS IN INFRARED DATA

This Appendix contains the published version of Kuhn et al. (2020).

Astrostatistics, or applying machine learning or statistical techniques for the knowledge discovery in Astronomy, can be applied for a variety of purposes. Throughout the thesis, we described how machine learning can be applied to data at optical wavelengths (*Gaia* DR2), to detect patterns in astrometric and photometric data which represents physical groupings of stars known as OCs. We also described how these objects can be used as tracers to extract information of bigger structures using different data mining techniques. In this Appendix, a set of Astrostatistics techniques is applied to data at near-infrared and infrared wavelengths, from the Spitzer space telescope, to detect and characterise young stellar objects (YSOs).

This work resulted in a catalogue of 117 446 YSOs, of which $\sim$ 90 000 are new identifications. This study represents the largest homogeneous catalogue of YSOs for the inner Galactic midplane.

Draft version November 30, 2020
Typeset using LaTeX **twocolumn** style in AASTeX63

arXiv:2011.12961v1 [astro-ph.GA] 25 Nov 2020

SPICY: The Spitzer/IRAC Candidate YSO Catalog for the Inner Galactic Midplane

Michael A. Kuhn,[1] Rafael S. de Souza,[2] Alberto Krone-Martins,[3,4] Alfred Castro-Ginard,[5] Emille E. O. Ishida,[6] Matthew S. Povich,[7,1] Lynne A. Hillenbrand[1]

for the COIN Collaboration

[1] *Department of Astronomy, California Institute of Technology, Pasadena, CA 91125, USA*
[2] *Key Laboratory for Research in Galaxies and Cosmology, Shanghai Astronomical Observatory, Chinese Academy of Sciences, 80 Nandan Rd., Shanghai 200030, China*
[3] *Donald Bren School of Information and Computer Sciences, University of California, Irvine, CA 92697, USA*
[4] *CENTRA/SIM, Faculdade de Ciências, Universidade de Lisboa, Ed. C8, Campo Grande, 1749-016, Lisboa, Portugal*
[5] *Institut de Ciències del Cosmos, Universitat de Barcelona (IEEC-UB), Martí i Franquès 1, 08028 Barcelona, Spain*
[6] *Université Clermont Auvergne, CNRS/IN2P3, LPC, F-63000 Clermont-Ferrand, France*
[7] *Department of Physics and Astronomy, California State Polytechnic University Pomona, 3801 West Temple Avenue, Pomona, CA 91768, USA*

(Received November 11, 2020; Revised November 25, 2020)

Submitted to the Astrophysical Journal Supplement Series

ABSTRACT

We present ∼120,000 Spitzer/IRAC candidate young stellar objects (YSOs) based on surveys of the Galactic midplane between $\ell \sim 255°$ and 110°, including the GLIMPSE I, II, and 3D, Vela-Carina, Cygnus X, and SMOG surveys (613 square degrees), augmented by near-infrared catalogs. We employed a classification scheme that uses the flexibility of a tailored statistical learning method and curated YSO datasets to take full advantage of IRAC's spatial resolution and sensitivity in the mid-infrared ∼3–9 μm range. Multi-wavelength color/magnitude distributions provide intuition about how the classifier separates YSOs from other red IRAC sources and validate that the sample is consistent with expectations for disk/envelope-bearing pre–main-sequence stars. We also identify areas of IRAC color space associated with objects with strong silicate absorption or polycyclic aromatic hydrocarbon emission. Spatial distributions and variability properties help corroborate the youthful nature of our sample. Most of the candidates are in regions with mid-IR nebulosity, associated with star-forming clouds, but others appear distributed in the field. Using Gaia DR2 distance estimates, we find groups of YSO candidates associated with the Local Arm, the Sagittarius-Carina Arm, and the Scutum-Centaurus Arm. Candidate YSOs visible to the Zwicky Transient Facility tend to exhibit higher variability amplitudes than randomly selected field stars of the same magnitude, with many high-amplitude variables having light-curve morphologies characteristic of YSOs. Given that no current or planned instruments will significantly exceed IRAC's spatial resolution while possessing its wide-area mapping capabilities, Spitzer-based catalogs such as ours will remain the main resources for mid-infrared YSOs in the Galactic midplane for the near future.

1. INTRODUCTION

The majority of young stellar objects (YSOs) in our galaxy are formed in massive star-forming complexes located near the Galaxy's midplane. The prevalence of star-forming regions in this part of the Galaxy is attested to by the spatially complex mid-infrared (mid-IR) nebulosity observed to permeate the entirety of the inner midplane and much of the outer midplane. For example, observations by the Spitzer Space Telescope (Werner et al. 2004) and the Wide-field Infrared Survey Explorer (WISE; Wright et al. 2010) have identified more than a thousand interstellar medium bubbles in these regions, most of which are associated with star formation activity (Churchwell et al. 2006, 2007; Simpson et al. 2012; Anderson et al. 2014; Bufano et al. 2018; Jayasinghe et al. 2019). Nevertheless, apart from a few dozen well-studied star-forming regions, the YSOs in these regions remain either mostly or wholly unknown. This is a consequence

Corresponding author: Michael A. Kuhn and Rafael S. de Souza
mkuhn@astro.caltech.edu, drsouza@shao.ac.cn

of observational difficulties at low Galactic latitudes, including high dust column densities along many lines of sight, which limit optical studies, high stellar densities, which may produce source confusion and increase the number of contaminants in catalogs, and lines of sight that pass through multiple star-forming regions at different distances (Feigelson 2018).

There are many scientific applications for reliable lists of YSOs generated uniformly for large segments of the sky rather than on a region-by-region basis. For example, it remains an open question whether nearly all stars are formed in dense groups or whether there is a significant population of stars formed in low-density environments (e.g., Carpenter 2000a; Bressert et al. 2010; Pfalzner et al. 2012; Gieles et al. 2012; Kuhn et al. 2015). Hence, catalogs that sample YSOs from both types of environment help to address this question. In addition, YSO catalogs, when combined with Gaia astrometric data, can be used to map out the kinematics of the youngest component of the Milky Way's thin disk. And, furthermore, with an increasing number of surveys searching large areas of the sky for transients, these catalogs would help in identifying outbursting YSOs and other YSO related variability (Hodgkin et al. 2013; Rosaria et al. 2018; Graham et al. 2019).

Our goal here is to make optimum use of Spitzer survey data from the inner Galactic midplane (between $\ell \sim 255°$ and 110° and $|b| < 1°$ to 3°) to identify YSOs out to several kpc in distance, using IR excess selection criteria that are independent of spatial clustering. Here, we focus on the 4-channel Infrared Array Camera (IRAC; Fazio et al. 2004) because this instrument provided the highest spatial resolution of any mid-IR imager with wide-area mapping capabilities over wavelengths from 3 to 9 μm. IRAC far exceeded the point-source sensitivity of *WISE* in the Galactic plane, because the latter was severely limited by both detector saturation and source confusion. The extensive IRAC observations of the Galactic plane were obtained as part of the Galactic Legacy Infrared Mid-Plane Survey Extraordinaire (GLIMPSE; Benjamin et al. 2003; Churchwell et al. 2009) along with several related Spitzer/IRAC programs that followed similar observing and data processing strategies.

Spitzer has proven effective at identifying candidate YSOs (e.g., Allen et al. 2004; Hartmann et al. 2005; Harvey et al. 2007; Simon et al. 2007; Gutermuth et al. 2009; Povich et al. 2011, 2013, and many others). However, these studies use differing criteria to select YSO candidates, ranging from simple cuts in color space to empirical probabilistic classification to fitting the spectral energy distributions (SEDs) with models of circumstellar dust actively infalling or accreting onto a central stellar object (e.g., Robitaille et al. 2006, 2007; Robitaille 2017). Our study employs a hybrid approach, which combines the strengths of SED fitting and principled statistical learning techniques.

An earlier GLIMPSE study ($\ell \sim 295°$–65°; Robitaille et al. 2008) identified ∼20,000 "intrinsically red sources" ($[4.5] - [8.0] \geq 1$), using strict photometric brightness and quality measures to guarantee that the infrared excesses they identify are real. However, they find that this selection criterion is sensitive not only to YSOs but also to intrinsically red contaminants, largely comprised of (post-)asymptotic giant branch stars (AGBs). They find that the 24 μm Spitzer/MIPS band is helpful for distinguishing between these cases. However, this band is not available for the vast majority of point sources detected in GLIMPSE (Gutermuth & Heyer 2015). In our study, we relax these criteria to identify significantly more YSOs candidates, but use patterns in the IRAC photometry to better distinguish between YSOs and contaminants. Our survey area also overlaps YSO catalogs for Cygnus X (Beerer et al. 2010; Winston et al. 2020) and the Spitzer Mapping of the Outer Galaxy (SMOG; Winston et al. 2019) survey; we use the latter as a benchmark to compare to our results.

This paper is organized as follows. Section 2 describes the datasets. Section 3 explains our statistical methodology. Section 4 introduces our YSO catalog. Color-color and color-magnitude diagrams of candidate YSOs and probable contaminants are examined in Section 5. The next sections describe properties of YSO candidates related to environment (Section 6), spatial clustering and kinematics (Section 7), and variability (Section 8). Comparisons with other catalogs are made in Section 9. Section 10 is the conclusion.

2. DATA

2.1. *IRAC catalogs*

The YSO selection is largely based on IRAC photometry from GLIMPSE (Benjamin et al. 2003; Churchwell et al. 2009) and related surveys that used similar observing strategies and data reduction methodologies.[1] These include GLIMPSE I (31,184,509 sources), GLIMPSE II (19,067,533 sources), and GLIMPSE 3D (20,403,915 sources), the Vela-Carina (2,001,032 sources) (Majewski et al. 2007; Zasowski et al. 2009), Cygnus X (3,913,559 sources) (Beerer et al. 2010), and SMOG (2,512,099 sources) (Winston et al. 2019) surveys. Spitzer's observations of the Galactic Center (Stolovy et al. 2006) were included in the GLIMPSE II Catalog. We use only the Spitzer photometry obtained during the cryogenic mission, which includes 4 mid-IR bands centered at 3.6, 4.5, 5.8, and 8.0 μm. We omit the GLIMPSE 360 data from the warm Spitzer mission that includes only the 3.6 and 4.5 μm bands.

The GLIMPSE team designed their reduction pipeline to provide reliable point-spread function (PSF) fitting

[1] The IRAC point-source catalogs were obtained from the NASA/IPAC archive at https://irsa.ipac.caltech.edu/data/SPITZER/GLIMPSE/overview.html

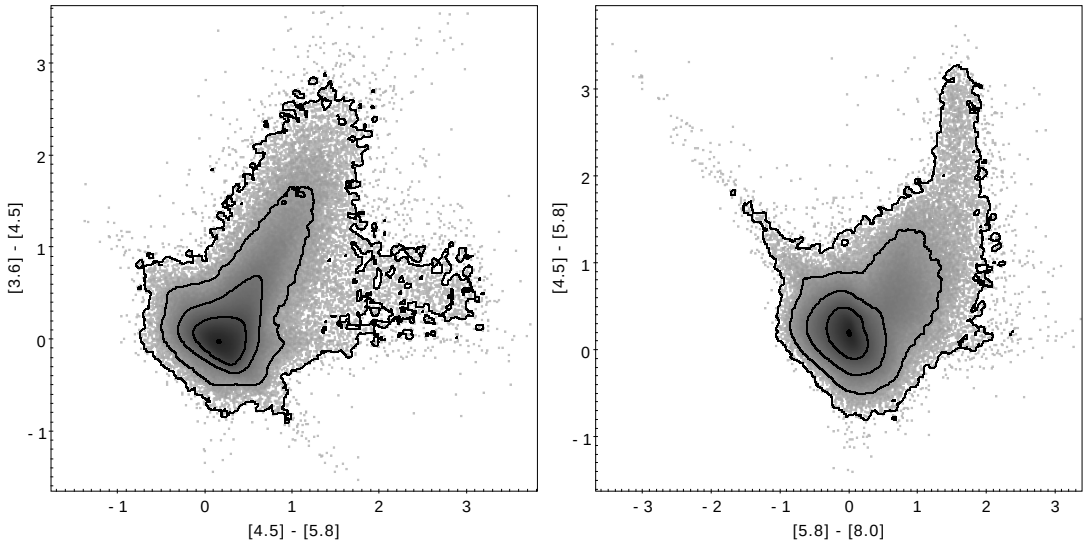


Figure 1. Colors of sources from the GLIMPSE Catalog. (Due to the high number of sources in the full tables, we display a random subsample for plotting convenience.) Contours are drawn at increases in density by a factor of 12. From these plots we see that the highest source density is at colors ∼0, but in both the [3.6] − [4.5] vs. [4.5] − [8.0] (left) and [4.5] − [5.8] vs. [5.8] − [8.0] (right) diagrams there is an excess of redder sources to the upper right. In the right panel, two additional features stand out. A streak from the origin to the upper left is an artifact resulting from source-extraction errors in the 5.8 μm band. To the upper right, there is a curved feature in the sources distribution, with [4.5] − [5.8] $\gtrsim$ 1.6 mag and [5.8] − [8.0] $\approx 1.6 \pm 0.25$ mag. We argue that these colors are affected by PAH emission (Section 5.7).

photometry in crowded fields with spatially varying nebular emission (Benjamin et al. 2003; Kobulnicky et al. 2013) – conditions that are common in IRAC images of star-forming regions. They provide two source lists for each survey, the "Catalog" which is more reliable, and the "Archive" which is more complete[2]. Following, Povich et al. (2013), we use the "Catalog" photometry. We make no additional cuts on quality flags, but certain flags are discussed in Appendix A. The Spitzer/IRAC images have PSFs with full widths at half maximum of 1.66″ at 3.6 μm, 1.72″ at 4.5 μm, 1.88″ at 5.8 μm, and 1.98″ at 8.0 μm. This is significantly better than the ∼6″ resolution provided by the WISE survey over a similar wavelength range (Wright et al. 2010), giving IRAC a distinct advantage in crowded fields in the Galactic midplane. The catalogs from the GLIMPSE team also include many stars that are missing from the Spitzer Enhanced Imaging Products (SEIP) due to GLIMPSE's better treatment mid-IR nebulosity. PSF photometry is also more accurate than SEIP aperture photometry in regions with variable backgrounds (Fang et al. 2020).

The GLIMPSE I, II, 3D, Galactic Center, and Vela-Carina survey observations consisted of 2–5 × 1.2 s integrations at each positions, while the Cygnus X and SMOG surveys used a 0.4+10.4 s high-dynamics range mode. Given that Cygnus X and SMOG are deeper than the rest of the data, we impose uniformity by omitting sources in these fields that are either brighter or fainter than the limits for the main GLIMPSE survey.[3] The magnitudes of our selected YSO candidates range from [3.6] = 8.3–14.9 mag, [4.5] = 7.3–13.7 mag, [5.6] = 6.4–12.9 mag, and [8.0] = 5.5–12.2 mag (1%–99% quantiles), and the median photometric uncertainties are 0.062, 0.073, 0.075, and 0.060 mag in these bands, respectively.

Even in the full GLIMPSE catalogs, the presence of red sources and several catalog artifacts can be seen in color-color diagrams (Figure 1). The highest concentration of sources have colors close to 0 (expected for normal stars without IR excess), but numerous sources form a distribution extending to the upper right in these

[2] http://www.astro.wisc.edu/sirtf/docs.html

[3] The bright limits for GLIMPSE are 7, 6.5, 4.0, and 4.0 mag in the 3.6, 4.5, 5.8, and 8.0 μm bands, respectively. The 3σ detection limits are 15.5, 15.0, 13.0, and 13.0 mag in these bands, but completeness declines precipitously before reaching these limit (http://www.astro.wisc.edu/sirtf/GQA-master.pdf). Completeness in the GLIMPSE Catalog is a strong function of both crowding and background sky level, with structure in the background playing a larger role than photon noise (Kobulnicky et al. 2013). In the main GLIMPSE survey, the magnitude distribution of Catalog sources peaks at [4.5] ≈ 13.6 mag, before declining. We adopt a faint limit of [4.5] = 14.5 mag (where density has decreased by a factor of ∼5) based on this magnitude distribution and because no source fainter than this is selected as a YSO candidate.

plots, which is made up of both YSOs and other red mid-IR sources (e.g., evolved stars, galaxies, etc.; Appendix B). In the $[4.5]-[5.8]$ vs. $[5.8]-[8.0]$ diagram, a streak can be seen extending from the origin to the upper left. This streak appears to be related to erroneous photometry in the 5.8 μm band for a low fraction of the GLIMPSE sources, and it extends from the origin because this is where the source density is highest. Another prominent feature in this diagram is a finger-like structure extending upward at $[5.8]-[8.0] \approx 1.6$, which we attribute to PAH emission (Section 5.7).

2.2. *Cross-Matches to Near-IR Catalogs*

Near-infrared JHK_s photometry from the Two Micron All Sky Survey (2MASS; Skrutskie et al. 2006) is already included in the GLIMPSE (and extensions) data products.2MASS has a spatial resolution of $\sim 2''$, which is comparable to the Spitzer/IRAC PSF. For our sample, 2MASS is nearly complete down to $J \sim 15.4$ mag, $H \sim 14.2$ mag, and $K_s \sim 13.0$ mag, with median photometric uncertainties of 0.038, 0.040, and 0.035, respectively. While these limiting magnitudes correspond well with the limits of the GLIMPSE surveys, in practice YSOs are often found in regions of high interstellar reddening, where 2MASS may not be deep enough to detect NIR counterparts of red GLIMPSE sources.

Deeper NIR catalogs with higher spatial resolution are available from the United Kingdom Infra-Red Telescope (UKIRT) Infrared Deep Sky Survey (UKIDSS; Lawrence et al. 2007) and the Visible and Infrared Survey Telescope for Astronomy (VISTA) Variables in the Vía Láctea survey (VVV; Minniti et al. 2010) for the northern and southern portions of the Galactic plane, respectively, with overlap around the Galactic Center. These catalogs are deeper than 2MASS, but are saturated for brighter sources. We use the UKIDSS catalog from the Galactic Plane Survey (Lucas et al. 2008) and the averaged VVV photometry for multiple epochs from the VVV Infrared Astrometric Catalog (VIRAC DR1; Smith et al. 2018). For both deeper NIR surveys, the photometry extends to $J \sim 19$ mag, $H \sim 18$ mag, and $K_s \sim 16$ mag with formal photometric uncertainties <0.01 mag.

IRAC and UKIDSS/VVV were cross matched using a $1''$ match radius in TOPCAT (Taylor 2005). Photometric measurements from UKIDSS/VVV were omitted if they did not have the flag $mergedClass = -1$ (stellar) or if they had magnitudes $J < 11$, $H < 12$, $K < 10.5$ (UKIDSS) or $J < 12.5$, $H < 13$, $K_s < 11.5$ (VVV), for which saturation effects start to affect photometry. The higher spatial resolutions of the NIR catalogs mean that it is possible for multiple NIR sources to be associated with individual IRAC sources; however, examination of IRAC+UKIDSS matching by Morales & Robitaille (2017) have found that the NIR flux is usually dominated by a single counterpart.

In our analysis we perform the YSO candidate selection independently on cross-matches of IRAC+2MASS, IRAC+UKIDSS, and IRAC+VVV, and the results of these separate selections are merged.

2.3. *Ancillary Data*

The Gaia mission (Gaia Collaboration et al. 2016), in its second data release (Gaia DR2; Gaia Collaboration et al. 2018), has provided optical broad band photometry (Evans et al. 2018) for the whole sky along with exquisite astrometric measurements (Lindegren et al. 2018) for more than 1.3 billion stars. These data on their own can be used for selecting possible pre–main-sequence stars (e.g., Zari et al. 2018), but for our study we use them as ancillary data to better understand the parallax (ϖ) and proper motions ($\mu_{\alpha*}, \mu_\delta$), distributions of the IR-excess selected YSO candidates. From the YSOs candidate list (Section 4), 33% have Gaia counterparts with the full 5-parameter astrometric solution (Lindegren et al. 2018). A match rate below 50% is expected because many YSOs are enshrouded by dust and thus not optically visible.

Longer wavelength photometry is available from both Spitzer's MIPS Galactic Plane Survey (MIPSGAL; Carey et al. 2009; Gutermuth & Heyer 2015) at 24 μm and the WISE All-Sky Data Release (Wright et al. 2010) at 22 μm. In the Galactic plane AllWISE is affected by high numbers of spurious sources, particularly in the longer wavelength bands, so we follow the catalog cleaning recommendations from Koenig & Leisawitz (2014) and apply the signal-to-noise and χ^2 quality cuts on the profile-fit photometry given in their Equations 1–4. For MIPSGAL, we use the "Catalog" instead of the "Archive." The WISE photometry, and to a lesser extent the MIPS photometry, is strongly affected by crowding and nebulosity in the regions around the Galactic plane that we are investigating, leading to fewer reliable source detections in these areas. The application of the quality cuts from Koenig & Leisawitz (2014) leave visible holes in the spatial distribution of WISE sources surrounding the clusters of YSO candidates that we identify with the IRAC photometry. Only 30% of our IRAC candidates have counterparts in the 24 μm MIPS band, and 8% have reliable 22 μm WISE photometry. Because of the unavailability of longer wavelength data for the majority of our sample, we do not uses these bands for identifying candidates, but only for post-selection examination of the sample. For both catalogs we used a cross-match radius of $1.2''$.

2.4. *Published YSO Catalogs*

YSOs identified in earlier studies of star-forming regions within GLIMPSE (and extensions) can can be used to train a classifier to find similar types of objects. We use YSOs identified as part of the Massive Young Star-Forming Complex Study in Infrared and X-ray (MYStIX; Feigelson et al. 2013), in addition to a

similar, earlier study of the Carina Nebula (Townsley et al. 2011). From the combined lists, we have included probable YSOs from the Carina Nebula, NGC 6611, M17, NGC 6530, M20, NGC 6357, NGC 6334, RCW 38, RCW 36, and DR 21, ranging from ∼0.7 to 2.7 kpc in distance. Povich et al. (2011, 2013) performed the IR excess detection for these projects using Spitzer data based on a strategy that included SED fitting of both reddened stellar atmospheres and the Robitaille YSO models, color cuts to remove certain types of contaminants, spatial filtering to remove objects that are not clustered, and visual examination of SEDs.

We select all objects from the MYStIX IR-excess catalogs classified as both a YSO ($Cl = 0$) and a probable member ($Mm = 1$). More recently, Gaia DR2 has become available, so, for the subset of sources with Gaia parallaxes (∼30% of the sample), we refine the sample further by removing any source with a parallax that is discrepant from the median parallax of the group by >2 times the reported parallax error.

In addition to the aforementioned studies of multiple regions and large areas, GLIMPSE has been used in hundreds of papers about individual (or several) star-forming regions. A few representative examples of these include Zavagno et al. (2006), Watson et al. (2008), Povich et al. (2009), Dewangan & Ojha (2013), Samal et al. (2014), Mallick et al. (2015), and Povich et al. (2016).

3. METHODOLOGY

YSOs make up a minuscule fraction of the nearly fifty million sources detected in the Spitzer/IRAC surveys included in this project. This means that selection of YSO candidates requires rejection of numerous contaminants (mostly field stars) along similar lines of sight. The first steps in our procedure, in which we reject sources that can be explained without IR excess, are nearly identical to those from Povich et al. (2013). These steps greatly reduce the sample size and are based on well established stellar atmosphere models. However, in the next steps – classification of the remaining sources – rather than fitting models of YSO SEDs as Povich et al. (2013) do, we use their resulting MYStIX YSO sample to train our random forest classifier.

A data-driven approach offers some advantages. For instance, we use IRAC photometry of actual stars as a training set instead of artificial photometry generated from theoretical YSO models. This means that the method will tend to avoid classifying an object with unusual colors as a YSO even if these colors can be reproduced by a physically unrealistic configuration of a star, disk, and envelope that exists in a grid of theoretical YSO models. Furthermore, it takes significantly less computational time to apply a trained classifier to millions of stars than it does to fit each of them with several categories of parametric YSO model. Nevertheless, the YSO SED fitting method does play an important role in generating training sets for the classifier (Section 3.2).

3.1. *Removing Sources without Significant IR Excess*

In the Galactic midplane, many background stars are affected by high levels of foreground extinction, so any source that is either insufficiently red or whose red colors can plausibly be explained by reddening alone is dropped from further scrutiny.

Cuts on IRAC colors and color uncertainties can remove many objects that have no chance of being selected as reliable IR excess objects. We apply the rules recommended by Povich et al. (2011, 2013), decreasing the number of sources in our sample by a factor of ∼10. All retained sources must be detected in at least 4 out of the 7 IR bands, two of which must be 3.6 and 4.5 μm. Sources are kept if there is the suggestion of IR excess in the $[3.6]-[4.5]$ color using the criterion

$$[3.6]-[4.5]-0.408 > \mathrm{error}([3.6]-[4.5]), \quad (1)$$

where "error" denotes uncertainty in color, calculated by adding the photometric uncertainties for the two bands in quadrature. The value 0.408 is the expected reddening of this color with $A_V \approx 30$ mag of extinction. Sources are also kept if they have have photometric measurements in the 5.8 and 8.0 μm bands and either meet both the criteria

$$|[4.5]-[5.8]| > \mathrm{error}([4.5]-[5.8]) \quad (2)$$
$$[5.8]-[8.0] > \mathrm{error}([5.8]-[8.0]), \quad (3)$$

or

$$|[4.5]-[5.8]| \leq \mathrm{error}([4.5]-[5.8]) \quad (4)$$
$$|[5.8]-[8.0]| \leq \mathrm{error}([5.8]-[8.0]). \quad (5)$$

These rules, optimized from experience with GLIMPSE data, ensure that determination of IR excess is based on more than just the 8.0 μm band, which can occasionally give a spuriously bright measurement.

We fit the JHK+IRAC SEDs of the remaining sources with reddened Castelli & Kurucz (2003) stellar atmosphere models, using the Indebetouw et al. (2005) extinction law. The fitting procedure takes into account the statistical photometric uncertainties on the data, which happen to be of similar size for both 2MASS and IRAC photometry. The UKIDSS and VVV datasets provide JHK photometry for many objects that were not detected in 2MASS, allowing many more sources to be included. However, the statistical measurement uncertainties for most UKIDSS and VVV sources are far more precise than for 2MASS or IRAC. Given that we are mostly interested in detecting deviations from a reddened stellar atmosphere model in the IRAC bands and we want similar selection performance for each dataset, we re-scale all UKIDSS and VVV error bars that are

smaller than the median 2MASS error bars to be equal to the median 2MASS error bars. The sources that are poorly fit by the reddened stellar photosphere, with χ^2 per data point >4, comprise the target set for our random forest classifier.

Overall, these pruning steps leave 319,251 2MASS+IRAC, 188,701 UKIDSS+IRAC, and 257,334 VVV+IRAC sources with possible IR excess as inputs to our classification step below.

3.2. *Training Sets*

The training data includes both MYStIX IR-excess sources (Section 2.4) that we label "YSO" and sources unlikely to be YSOs that we label "contaminant" (see discussion of contaminants below). Although lists of members are more complete in some of the nearest star-forming regions (e.g., Ophiuchus or Taurus; Evans et al. 2009a; Luhman 2018), we choose to use MYStIX because these massive star-forming complexes may better represent the regions we expect to probe in the Galactic midplane, i.e. at greater distances, with higher extinction, and in more extreme environments. Furthermore, many of the MYStIX regions lie within the survey region of GLIMPSE and its extensions, meaning that homogeneous data products are available for both the training and target sets.

Contaminants can include both sources that occur in star-forming regions (e.g., non-stellar sources such as nebular knots and shocked emission) and sources that are smoothly distributed on the sky (e.g., AGB stars and galaxies; Robitaille et al. 2008; Gutermuth et al. 2009) – see Appendix B for discussion of these objects. Povich et al. (2011, 2013) used a variety of techniques to remove these objects from their catalogs, including SED fitting of Robitaille et al. (2007) models, color cuts, and visual inspection. We label any object within the field of view of these studies that was not classified as a probable young star by either IR or X-ray criteria as a field object.

To enlarge our sample of contaminants, we identify several fields near the Galactic plane that have no signs of star formation (Appendix B.3) and label objects in these fields as non-YSOs. These fields were selected to include lines of sight at multiple Galactic longitudes, in the midplane and up to several degrees above or below it, and with different amounts of Galactic extinction.

Training sets are generated separately for each combination of NIR+IRAC data due to the differences in NIR filters. For 2MASS+IRAC, the training set contains 2,865 YSOs, 3,436 field objects in the MYStIX fields, and 7,718 other field objects for the 2MASS+IRAC dataset. For UKIDSS+IRAC these numbers are 919, 2,000, and 1,128, and for VVV+IRAC they are 1,266, 1,459, and 2,595, respectively.

The distributions of training-set object in color space is discussed in Appendix C. The full IRAC catalog of $\sim 5 \times 10^7$ sources includes a few outliers in region of color space that are not well sampled by either the labeled YSOs or labeled non-YSOs in the training set. (The limits used to identify these outliers are given in Appendix C.) Given that we have little basis to assign such objects to either category, we are cautious and do not include these objects in our final YSO list.

3.3. *Missing data*

When data are combined across multiple catalogues, it is almost certain that missing data will occur, as is the case here. Figure 2 depicts the missing pattern for the 2MASS+IRAC, UKIDSS+IRAC, and VVV+IRAC training sets, from which only 57%, 46% and 29% of objects, respectively, have complete information. About 20% of the 2MASS+IRAC objects are missing three colors at once, and JHK_s are often missing together. The VVV+IRAC dataset has the most missing data, with 57% of the objects missing at least three bands. While UKIDSS+IRAC is the most complete, more than 35% of their rows have at least two missing colors. Thus, a naive removal of rows presenting missing values would throw away a non-negligible amount of valuable information.

As a final pre-processing step before training our YSO classifier we employed a multiple copula imputation. This decomposes joint probability distributions into their marginal distributions and a function, the copula, that couples them (Nelsen 2010). Copulas have been used previously in astronomy, for example, to construct likelihood functions for weak lensing analysis (Sato et al. 2011; Lin et al. 2016) and to infer bivariate luminosity and mass functions (Andreani et al. 2018). Previous tests suggest that this method outperforms other popular approaches, such as multiple imputation via chained equations (van Buuren & Groothuis-Oudshoorn 2011) and Amelia (Honaker et al. 2011), in terms of bias and coverage, especially in cases where the variables are not normally distributed (Hoff 2007). The underlying idea of copula imputation is to derive conditional density functions of the missing variables given the observed ones through the corresponding conditional copulas, and then impute missing values by drawing observations from them. Finally, the choice of performing imputation before training the random forest models has been previously assessed by other studies (Jaeger et al. 2020), which have shown it to reduce the variance in model error estimate, without any detectable change in precision. The imputation method was implemented using the SBGCOP package (Hoff 2018) within the R language (R Core Team 2019). Copulas were fit simultaneously to both training and target datasets.

Overall, the imputed data preserves the coverage of the original dataset (Appendix C). Nevertheless, we do not advocate for the use of these colors in other contexts. They are treated here as nuisance parameters to enable classification of the entire dataset.

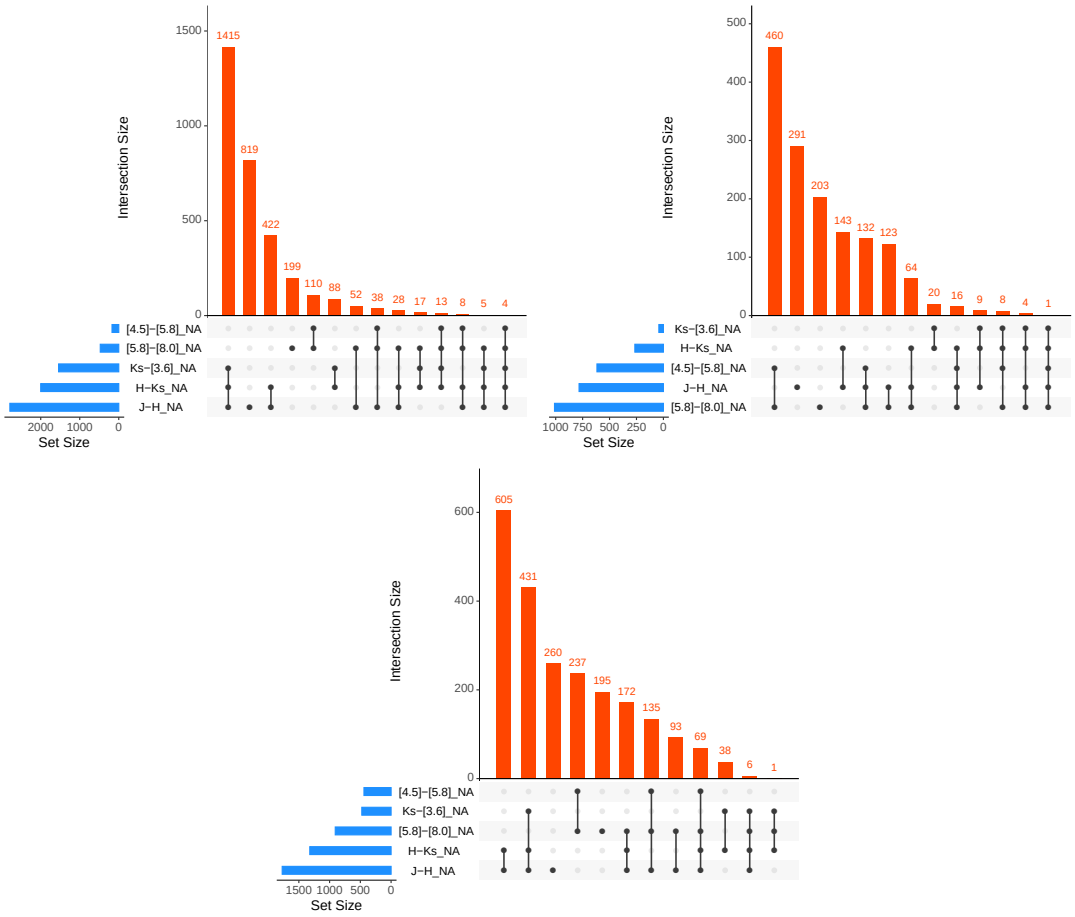


Figure 2. Missing data pattern for 2MASS+IRAC (upper left), UKIDSS+IRAC (upper right), and VVV+IRAC (bottom) from the labeled training set. Blue bars are the number of missing colors, the connected black dots indicate combinations of missing colors, and the red histograms indicate the number of instances these combinations are missing. Note that the bars are sorted by number of examples, and the order differs between the plots.

3.4. *Tree-based Classification*

Decision trees are learning algorithms that resemble the natural flow of human decision making. At each node of the tree, the algorithm randomly selects one feature and, based on the distribution of training data, determines the decision boundaries that best separate different classes. Training objects are then propagated along their branch of the tree to the next node, where a new feature is selected. The process is repeated until the tree reaches a pre-determined depth or until all objects in a leaf belong to the same class (for a detail description see Rokach & Maimon 2014). This basic concept has given rise to successful algorithms in many different fields. However, a single decision tree trained on an entire dataset is prone to overfitting, presenting low accuracy results whenever faced with data not used in training. This problem can be overcome by randomizing different stages of the tree construction and combining many independent estimators in a more robust classifier. This type of approach belongs to the wider class of ensemble models.

Ensemble methods (e.g. Sagi & Rokach 2018) are regression algorithms, constructed from the combination of many weak classifiers that, when considered together, provide a robuster estimate than any of their constituents. Random forests (Ho 1995; Breiman 2001) are one such algorithm, composed of many decision trees, each constructed independently. The final classification is determined via majority vote, considering all trees in the forest. In this context, the probability of being a YSO is approximated by the percentage of trees in the ensemble voting for a YSO candidate – we call this probability estimate the "YSO score." Random forests have been successfully used to classify YSOs in

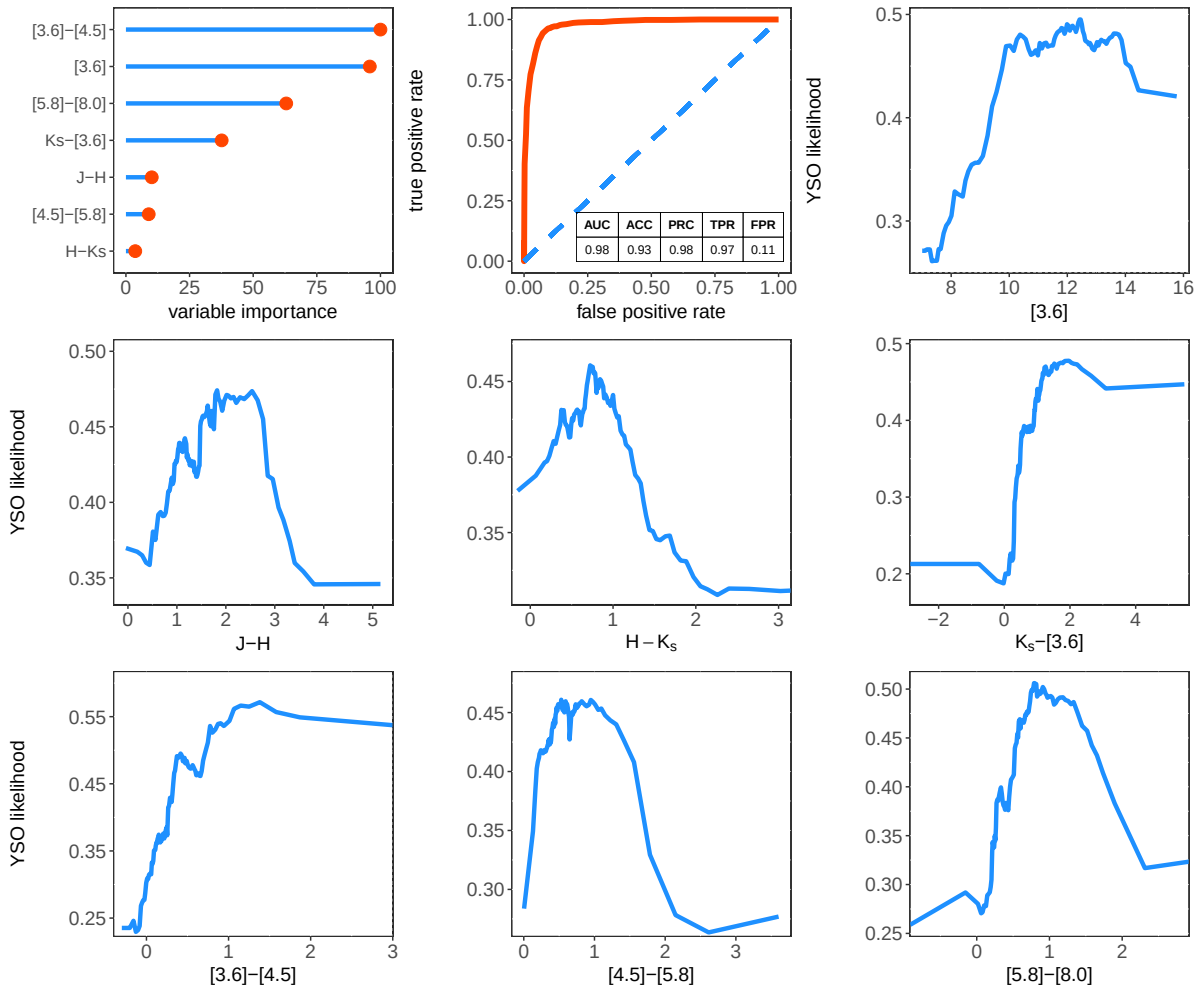


Figure 3. Top left: Estimated importance of colors and magnitudes in the random forest model fit to the 2MASS+IRAC data. Top center: Receiver operating characteristic (ROC) curve, with values for area under the curve (AUC), accuracy (ACC), precision (PRC), true positive rate (TPR), and false positive rate (FPR) using a threshold of $p = 0.5$ for classifying sources as YSOs. Other panels: Mean YSO score from the classifier as a function of each feature.

smaller scale studies, including with missing data imputation (e.g., Ducourant et al. 2017; Melton 2020).

We construct the YSO random forest classification using the following covariates: $J - H$, $H - K_s$, $K_s - [3.6]$, $[3.6] - [4.5]$, $[4.5] - [5.8]$, $[5.8] - [8.0]$, and the 3.6 μm band magnitude. Each model was independently trained for 2MASS + IRAC, UKIDSS + IRAC, and VVV + IRAC datasets. The random forest was employed using the CARET R package, with 1500 trees, which was sufficient to guarantee a stable solutions. As a sanity check, we tested few other regression models (including generalized additive models, support vector machine, gradient boosting machines, and conditional random forest), but no significant difference in the final YSO candidate set was found. This suggests that random forest (or other typical non-linear classifiers) captures the data complexity well enough without the need for highly complex models.

3.5. *Performance metric*

The receiver operating characteristic (ROC) curve (Figure 3, top center) provides a visually and quantitative approach for assessing the accuracy of a binary classifier. The curve plots the true positive rate (TPR) versus the false positive rate (FPR) for different values of the decision boundary, i.e., the classifier score used for deciding whether an object is a YSO. This curve lets us examine the performance of the classifier under unequal error costs, i.e., scenarios where the cost of a false

positive is different from a false negative. The quality of a ROC curve can be assessed by the area under the curve (AUC). Higher values of AUC correspond to more accurate classifiers, while a value of 0.5 corresponds to a random guess. Other measures include the accuracy (ACC), the number of true positives divided by the total population, and the precision (PRC), the number of true positives divided by the number of true positives plus false positives. The model statistics given on the ROC plot indicate good performance.

Variable importance (Figure 3, top left) is evaluated via out-of-bag samples, which consists of random samplings of the data that are left out of each tree. This is calculated by measuring variations in the prediction error when the out-of-bag data are permuted solely among a specific color, leaving the others unchanged. The process is then repeated across all trees. The final result is a measure of the incremental error for a given color when compared with the unperturbed colors for all the 1500 trees over the entire forest.

The subsequent panels of Figure 3 display partial dependence plots (PDP; Greenwell 2017). PDPs are useful for visualizing the relationship between individual features and the response while accounting for the average effect of the other predictors in the model. The shape and steepness of the curves are indicators of the predictor's relative influence. Note the sharp behaviour of $[3.6] - [4.5]$, one of the best indicators of YSOs candidates.

4. CATALOG

All objects with YSO scores >50% from any of the three random forests are classified as candidate YSOs, while other sources are regarded as probable contaminants. Among the Spitzer sources with IR excess, there are 117,446 candidate YSOs and 180,997 probable contaminants. The candidates are listed in Table 1 with the designation Spitzer/IRAC Candidate YSO (SPICY).

Figure 4 shows how the candidates are distributed within the footprints of the Spitzer surveys. Many of the candidate YSOs are concentrated toward the Galactic midplane while others form prominent clumps. More detail is visible in the zoomed-in maps from the atlas (Figure 5), which have been labeled with the locations of the H II bubbles from WISE (Anderson et al. 2014) and massive YSOs from MSX (Lumsden et al. 2013). These maps show that the YSO candidate distribution can be resolved into stellar clusters and associations, along with a non-negligible number of widely distributed objects. The locations of dense groups of YSOs are often correlated with the WISE bubbles and the MSX sources.

With the new YSO candidates, some previously unrecognized stellar groups become apparent. In Figure 6, we show an image containing one such group located in the Vela-Carina portion of the survey and designated G271.6-0.5 (left side of the image). To the south west

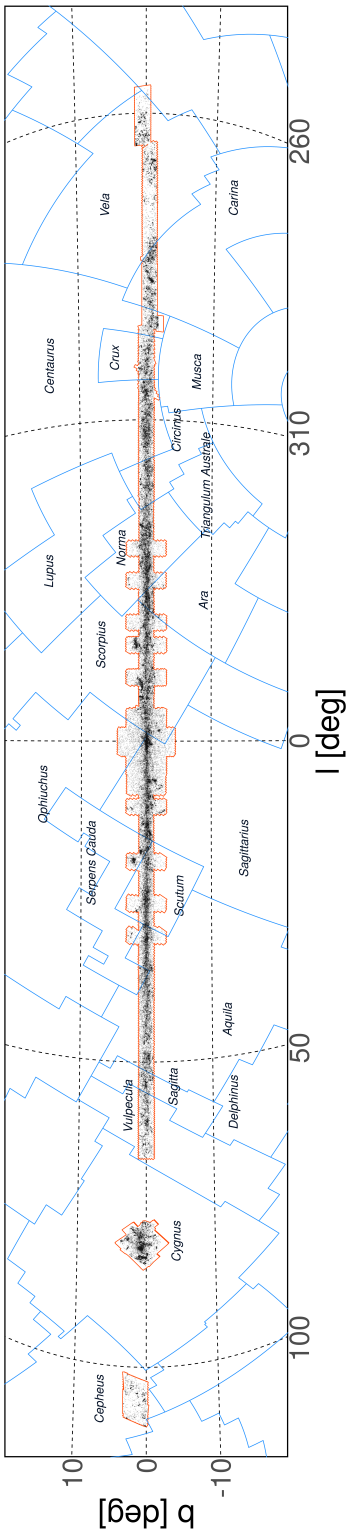


Figure 4. Spatial distribution of the candidate YSOs (black points) within the Spitzer/IRAC survey regions (outlined in black). The data are plotted in Galactic coordinates and the constellation boundaries are shown in blue. Candidate YSOs tend to be concentrated toward the midplane and/or in spatial clusters.

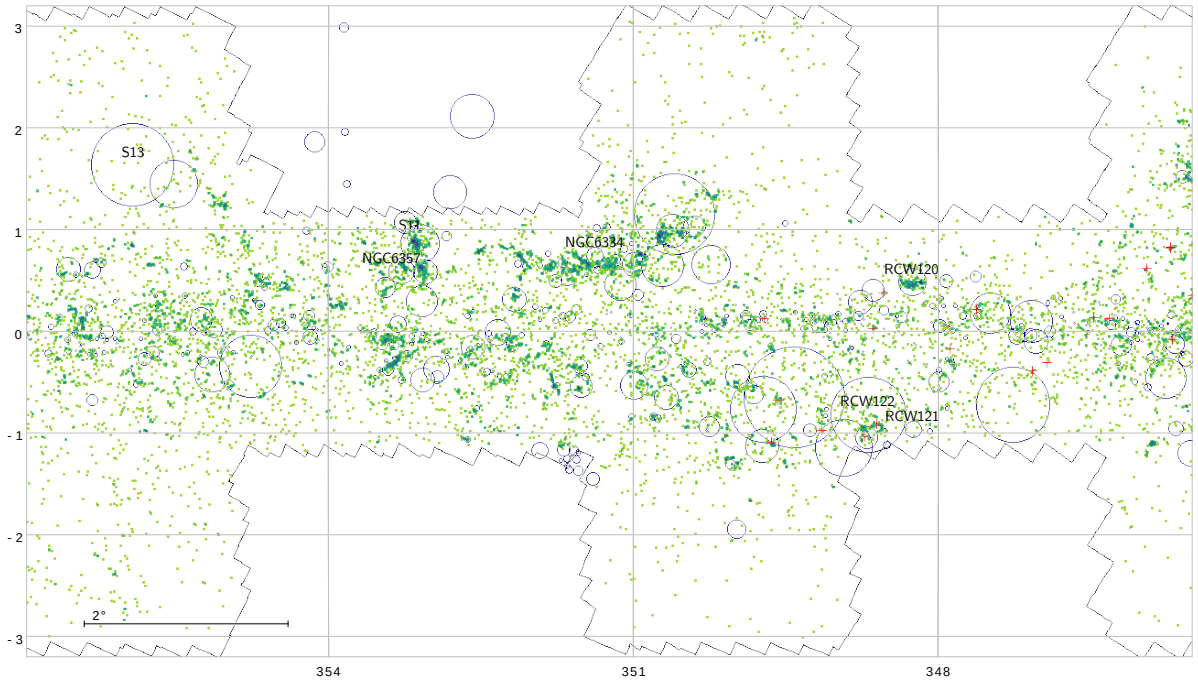


Figure 5. This figure set (19 components) provides an atlas of the Galactic midplane with locations of the YSO candidates (green points) and the boundaries of the IRAC surveys (black lines). Overlapping points produce darker shades of green, using a square-root scale and the "viridis" color pallet. For context, we also show outlines of H II bubbles from WISE (blue circles; Anderson et al. 2014), massive YSOs from MSX (red crosses; Lumsden et al. 2013), and labels of select star-forming regions.

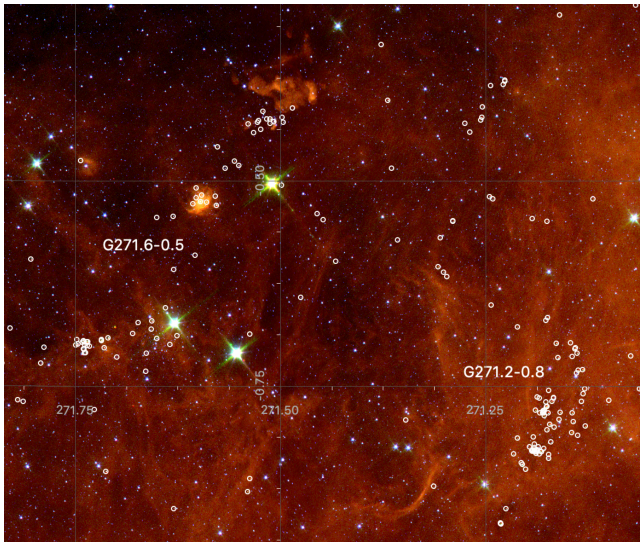


Figure 6. Spitzer/IRAC image (Vela-Carina survey) with our YSO candidates marked by the white circles. The image is composed of the 3.6 μm (blue), 5.8 μm (green), and 8.0 μm (red) images. The image captures two groups of stars, the previously un-studied group G271.6-0.5 and a neighboring group G271.2-0.8.

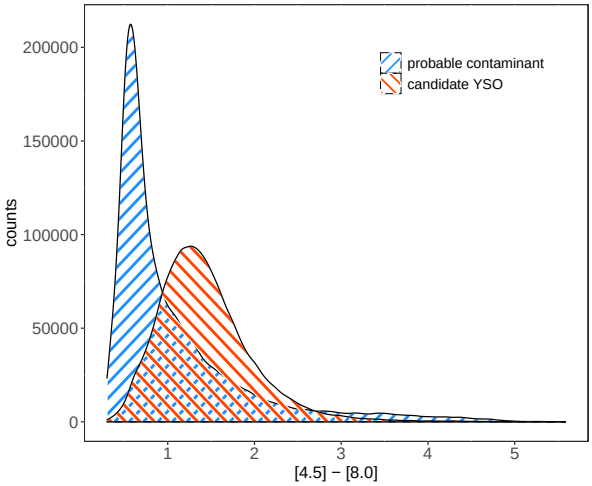


Figure 7. Distributions of $[4.5]-[8.0]$ color for the candidate YSOs (red stripes) and the probable contaminants (blue stripes). Overall, the candidate YSOs tend to be redder than the probable contaminants. Densities of sources in both samples are approximately equal at $[4.5]-[8.0] \approx 1$ mag, the limit imposed in the study by Robitaille et al. (2008), but in our sample 18% of the YSO candidates are bluer than this limit and 25% of the probable contaminants are redder. Probable contaminants also outnumber candidate YSOs at colors $[4.5]-[8.0] \gtrsim 3.5$ mag.

of this group, a previously identified, but little studied, star-forming region, G271.2-0.8 can also be seen.

The SPICY catalog is the largest homogeneous sample of YSO candidates available to date for the inner regions of the Milky Way. It seems unlikely that this mid-IR list of YSOs will be superseded in the near future given that no existing or planned mid-IR instrument exceeds Spitzer's spatial resolution in tandem with its wide-area mapping capabilities. The catalog is intended for both use in addressing questions about star formation on Galactic scales and assistance in searches for interesting individual YSOs. However, some contaminants inevitably remain, and formal assessment of contamination requires followup observations (e.g., spectrographic surveys). Nevertheless, the properties of these stars, including their colors, the environments in which they are found, their spatial and kinematic distributions, and their photometric variability (discussed in Sections 5–8), are useful for corroborating the results of the random forest classifier and may give a qualitative sense of the level of remaining contamination.

5. COLOR AND MAGNITUDE DISTRIBUTIONS

By examining the IR color and magnitude distributions for classified objects, we gain insight into how the classifier makes its decisions and how it compares to other selection criteria used in previous studies.

Figure 7 shows the distribution of $[4.5]-[8.0]$, one of the main features used in the earlier study by Robitaille et al. (2008). The sources we input into the classifier have a bimodal distribution in this color, but each of the output classes has a unimodal distribution, with the probable contaminants making up the bluer peak and the YSO candidates making up the redder peak. The densities are approximately equal at $[4.5]-[8.0] \approx 1$, the threshold used by Robitaille et al. (2008), but we also find a substantial number of objects of both classes (but particularly the contaminants) crossing the threshold.

The four panels of Figure 8 show the magnitude distributions of the classified sources in each IRAC band. The input distributions are bimodal, with lower peaks near the brightness limits and higher peaks at fainter magnitudes. The peaks at bright magnitudes can be attributed to an artifact of the χ^2 fitting step because bright sources tend to have smaller magnitude uncertainties, and thus a smaller deviation is capable of leading to a formally "bad fit." The classifier has identified the majority of sources associated with the bright peaks as probable contaminants. The distributions of the candidate YSOs are all unimodal, with peaks at fairly faint magnitudes, and heavy tails extending to brighter magnitudes. The contaminants also exhibit a second peak at magnitudes slightly fainter than the peak for YSOs.

The YSO magnitude distributions appear reasonable, given that we would expect most of them to be low-to-intermediate mass objects at distances of one to several kpc, with a low number of brighter objects that could either be massive YSOs or nearby objects. The tendency to classify the faintest objects as probable contaminants may inherit a bias from the MYStIX training set which only includes YSOs out to ∼3 kpc. However, the differences in colors of the faintest objects (examined below) imply that they may be intrinsically different.

Figures 9–12 show various JHK+IRAC color-magnitude and color-color diagrams. Candidate YSOs (red points) overlap probable contaminants (blue points) in each of these projections. Nevertheless, the locations in these diagrams with greatest source density are different for the two classes. We show reddening vectors indicating the effect of $A_K \approx 1$ mag (∼9 mag in the V band) of extinction, adopting the reddening law from Rieke & Lebofsky (1985) for JHK and Indebetouw et al. (2005) for the IRAC bands. We also plot curves for the near-IR stellar colors for stellar models without additional IR excess. For graphical display, we have merged the 2MASS, UKIDSS, and VIRAC photometry, converting UKIDSS and VIRAC to the 2MASS system using the first-order transformations from Hodgkin et al. (2009) and Soto et al. (2013), and picking the most reliable photometry for each source.

Some of these color spaces have been used in previous studies for selecting YSOs based on cuts on color. For

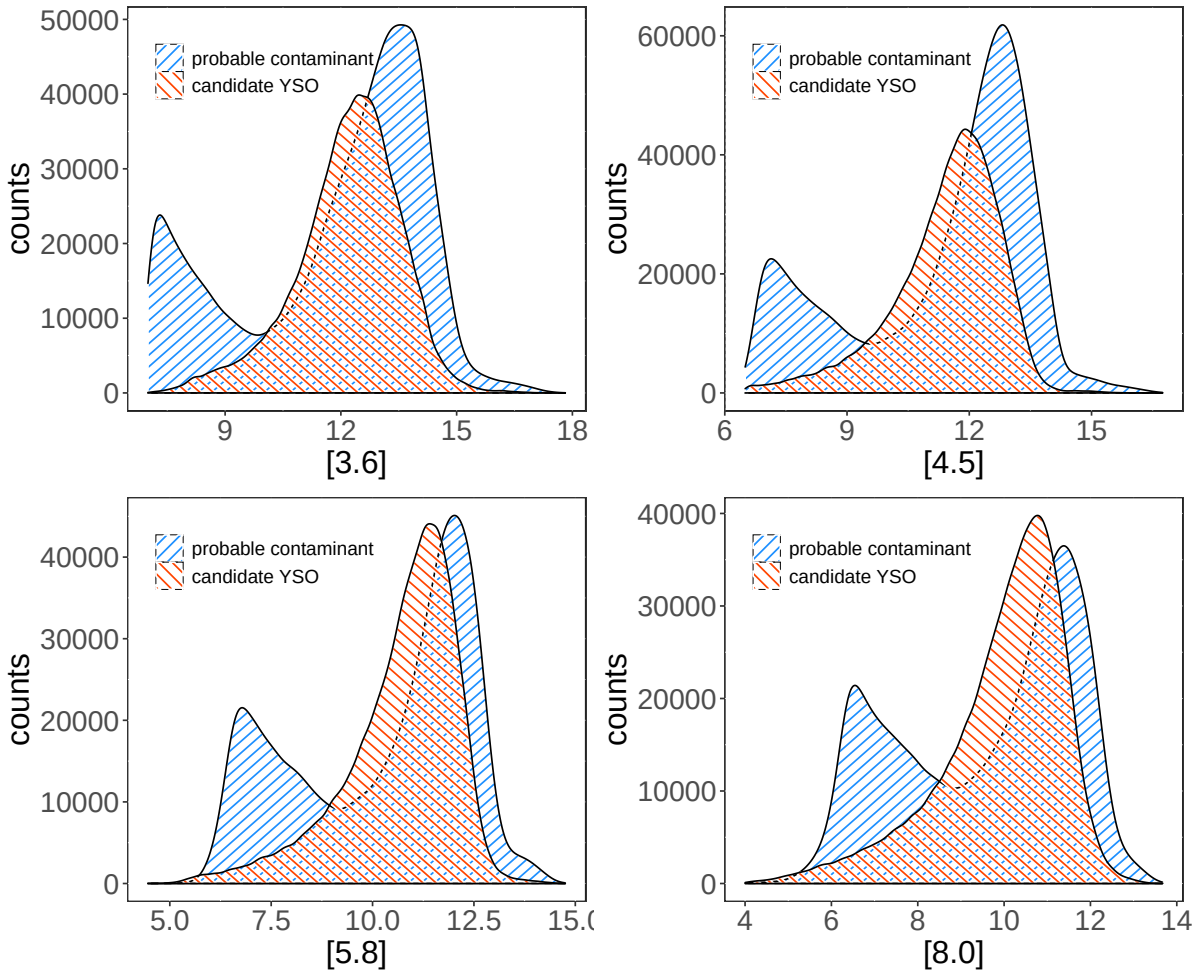


Figure 8. Distributions of IRAC magnitudes for the candidate YSOs (red stripes) and the probable contaminants (blue stripes). The distributions for probable contaminants are all multimodal, while the YSO candidates each have a single mode toward the fainter end of the distribution, and a heavy tail consisting of brighter sources.

example, the selection boundaries between YSOs and contaminants used by Gutermuth et al. (2009) are depicted as gray lines in several of the diagrams, including [4.5] vs. [4.5] − [8.0] (Figure 9, left panel), [3.6] − [4.5] vs. [4.5] − [5.8], and [4.5] − [5.8] vs. [5.8] − [8.0] (Figure 11, upper panels).

In the following subsections, we examine the IR criteria used for classification, evidence from Gaia that stars are pre–main-sequence, properties of the stars at 24 μm, YSO evolutionary classes, and the effects of various IR absorption and emission features.

5.1. *Color-Magnitude Diagrams*

On the J vs. $J-H$ diagram (Figure 9, left), both candidate YSOs and probable contaminants occupy a triangular region of color–magnitude space, where the upper edge of the triangle is approximately parallel to the reddening vector. The YSO candidates are densest around $J \sim 15.5$ mag and $J - H \sim 1.3$ mag, whereas the probable contaminant distribution is multi-modal, with one peak just blueward of the peak of the YSO candidates, and another strip of stars along the upper right edge of the triangle. The stars in this strip, which are more luminous than the typical YSO candidate with the same $J - H$ color, lie in the region of the diagram that would be occupied by reddened post–main-sequence stars.

On the [4.5] vs. [4.5] − [8.0] diagram (Figure 9, right), the YSO candidates form a smooth distribution ranging from the bright limit at [4.5] = 6.5 mag to ∼14 mag, where sensitivity declines, with the peak of the distribution at [4.5] ∼ 12.2 mag and [4.5]−[8.0] ∼ 1.2. A ∼1 Myr old (pre-)main-sequence star with a mass in the range

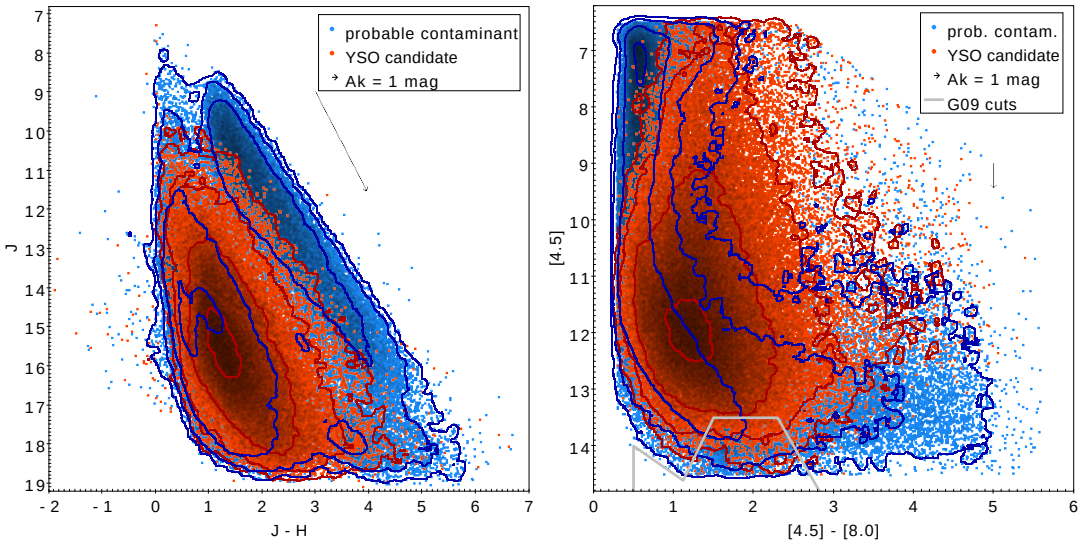


Figure 9. Infrared color-magnitude diagrams, J vs. $J-H$ (left) and [4.5] vs. [4.5] − [8.0] (right), with candidate YSOs (red) and probable contaminants (blue). In low-density parts of the scatter plot, individual points are drawn, but in areas with overlapping points, darker colors indicate higher density. We also include contours at evenly spaced logarithmic increases in density. The arrow indicates the approximate shift produced by extinction of $A_K = 1$ mag assuming the Indebetouw et al. (2005) reddening law. The gray polygon demarcates the region used by Gutermuth et al. (2009) to select contaminants.

0.4–10 $M_\odot$ at a distance of ∼1-2 kpc would have an an unreddened photospheric magnitude $9 \lesssim [4.5] \lesssim 14$ mag (Bressan et al. 2012) – approximately where were find the bulk of the YSO candidates. The probable contaminant distribution peaks at both bright and faint magnitudes. The bright contaminants form a band that tends to be bluer than the YSOs in [4.5] − [8.0], while the faint contaminants tend to have redder [4.5] − [8.0] colors.

The gray lines on the [4.5] vs. [4.5]−[8.0] diagram were defined by Gutermuth et al. (2009) to separate dusty AGNs from YSOs in their studies of nearby star-forming regions. Although many of the faintest 4.5 μm sources in our sample have been classified as probable contaminants, the region defined by Gutermuth et al. (2009) for selecting AGNs does not appear to separate our classes well. This apparent discrepancy may arise because Gutermuth et al. (2009) examined deeper Spitzer surveys of relatively nearby star-forming clouds at higher Galactic latitudes, where more AGN are expected to be detected, whereas GLIMPSE is less sensitive to this type of contaminant. Furthermore, GLIMPSE includes more distant star-forming regions in which legitimate YSOs will present fainter observed [4.5] magnitude distributions.

5.2. *Color-Color Diagrams*

Figure 10 shows the distributions of sources in $J-H$, $H-K_s$, and $H-[4.5]$. On the JHK_s diagram, we include a representative isochrone for ∼1 Myr unreddened stellar models. Most of the objects are shifted to the upper right from this curve, in the approximate direction of the reddening vector. However, the distribution of the YSO candidates spreads to redder $H-K_s$ colors, which would be expected for stars with K_s-band excess. Objects with very red $J-H > 5$ colors are largely classified as contaminants.

On the $H-K_s$ vs. $K_s-[4.5]$ diagram, we show both a 1 Myr isochrone for (pre–)main-sequence stars and a 1 Gyr isochrone that also includes post–main-sequence stars. The red-giant branch extends upward to stars with redder $H-K_s$ colors than the (pre–)main-sequence, allowing these groups to be better separated. On this plot, the reddening vector points to the upper right. If we consider a line parallel to the reddening vector, starting from the tip of the asymptotic giant branch (as shown by the gray line in the figure), we would expect that many of the stars lying above this line could be evolved stellar contaminants. This is consistent with what the classifier finds; most objects above this line are classified as probable contaminants, while the candidate YSOs are more abundant below this line. The slope of the Indebetouw et al. (2005) reddening vector is not precisely parallel to the upper edge of the source distribution; this may arise due to systematic uncertainties in the reddening law or could be a property of IR colors of highly obscured evolved stars.

Figure 11 shows four projections of sources in IRAC color-color space. On the [3.6] − [4.5] vs. [4.5] − [5.8] diagram, the YSO candidates are smoothly distributed, with a peak in density around [3.6] − [4.5] ∼ 0.5 and [4.5] − [5.8] ∼ 0.4, and a tail that extends up and to the

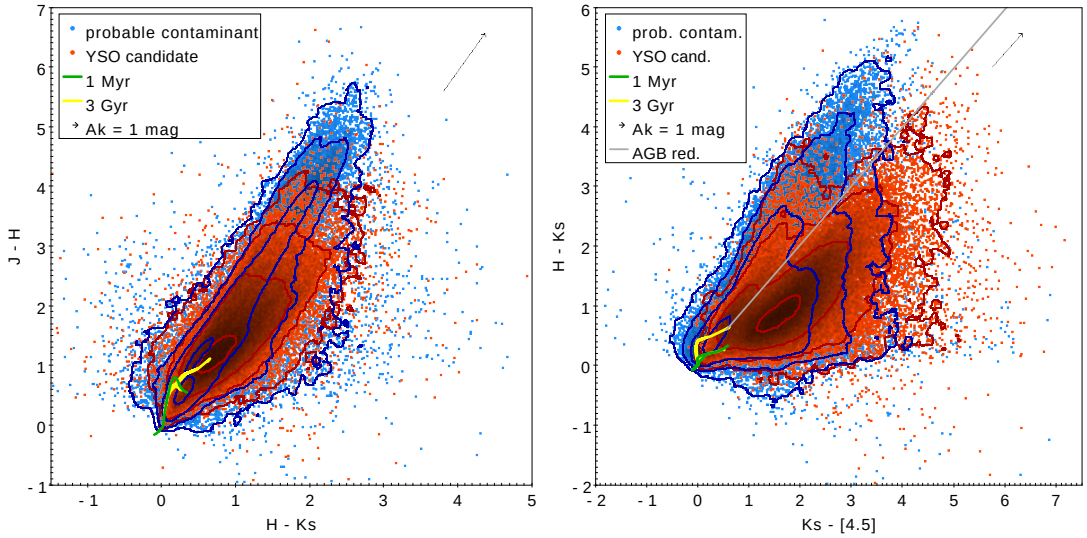


Figure 10. Color-color diagrams for $J-H$ vs. $H-K_s$ (left) and $H-K_s$ vs. $K_s-[4.5]$ (right) with candidate YSOs (red) and probable contaminants (blue). Curves indicate 1 Myr (green) and 3 Gyr (yellow) isochrones (Bressan et al. 2012), with models for the AGB phase (Marigo et al. 2013) included in the 3 Gyr isochrone. In the right panel, a gray line parallel to the Indebetouw et al. (2005) reddening vector extends from the tip of the AGB. In our sample, objects classified as contaminants predominate above this line while objects classified as candidate YSOs are more abundant below.

right. In contrast, the contaminant distribution peaks slightly bluer in $[3.6]-[4.5]$, and the distribution appears bifurcated, with some sources being redder in $[3.6]-[4.5]$ while others are redder in $[4.5]-[5.8]$. This bifurcation may be related to the types of contaminants. For example, the contaminants to the upper left roughly correspond to a region of color space identified by Gutermuth et al. (2009) (and indicated by the gray boundary) as containing sources produced by knots of shocked emission from H_2, while the contaminants to the lower right were associated with knots of PAH emission. Both areas outlined by Gutermuth et al. are dominated by objects that we classify as probable contaminants, but the edges of the YSO distribution also overlaps these boundaries.

On the $[4.5]-[5.8]$ vs. $[5.8]-[8.0]$ diagram the peak density of YSO candidates is redder in both colors than the peak density of probable contaminants. The objects with the most extreme $[5.8]-[8.0]$ colors are nearly all classified as contaminants. These lie in a region of the diagram identified by Gutermuth et al. (2009) (gray boundary lines) as being dominated by unresolved star-forming galaxies. This diagram also includes a finger comprised of both YSO candidates and contaminants, extending to high $[4.5]-[5.8]$ values ranging from $\sim$2 to $\sim$3.5, but with $[5.8]-[8.0]$ colors in a restricted range $1.5 \lesssim [5.8]-[8.0] \lesssim 2.6$. Previously published YSO catalogs (e.g., Gutermuth et al. 2009; Rebull et al. 2011) have included a few YSOs in this region of color space; however, the high number of sources identified when examining the entire inner Galactic midplane makes this feature appear much more pronounced. These stars have colors similar to the PAH nebulosity found in star-forming regions (Povich et al. 2013); however, visual inspection of a sample of these sources suggests that the majority are *bona fide* point sources in all four IRAC bands.

In the bottom two panels of Figure 11, the peaks of the YSO candidate distributions are redder than the peaks of the probable contaminant distributions for each IRAC color. Nevertheless, while the objects with reddest $[3.6]-[4.5]$ tend to be YSO candidates, the objects with most extreme red $[4.5]-[8.0]$ or $[5.8]-[8.0]$ colors are almost all classified as contaminants.

Figure 12 shows the $J-H$ colors, which are the most sensitive to extinction, versus the IRAC colors, which are the most sensitive to IR excess. In $J-H$, the peaks of the density distributions are slightly redder for the contaminants than for the YSO candidates, but in IRAC colors, the peaks are significantly redder for the YSO candidates than the contaminants. In both cases, the objects with most extreme red colors tend to be classified as contaminants. However, of the reddest IRAC sources do not appear on these plots because they lack J-band magnitudes.

On all these color-color plots, the blue ends of the distributions are artificially truncated by the selection rules imposed to ensure that the IR excesses are real. Thus, our catalogs will not be sensitive to certain classes of YSOs, including some YSOs with anemic disks or pre–main-sequence stars without disks.

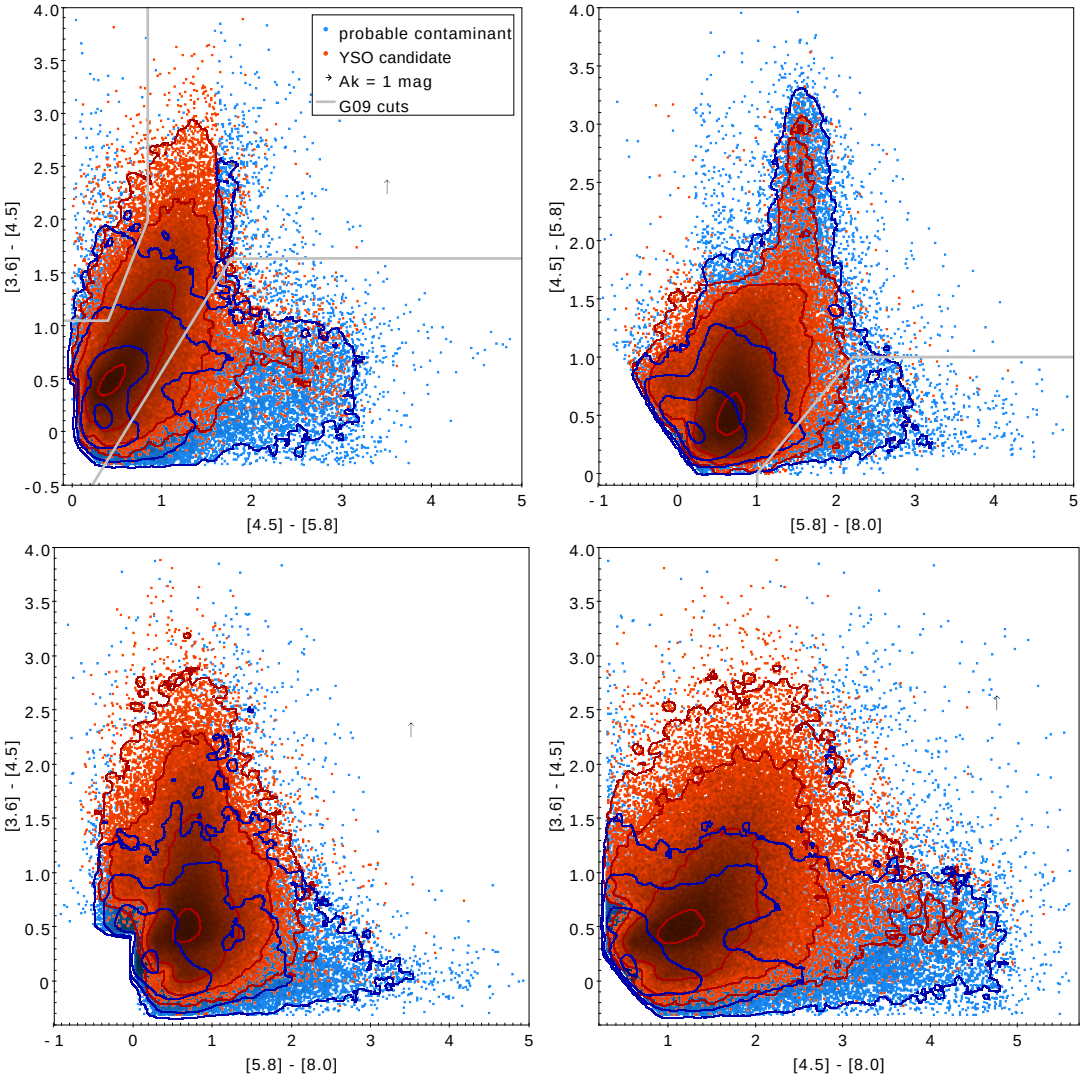


Figure 11. Color-color diagrams in the IRAC bands. The YSO candidates (red points) and probable contaminants (blue points) partially overlap in each of these projections, but differences are visible in their distributions. The short length (or absence) of the $A_K = 1$ mag reddening vectors (black arrows) implies that extinction would need to be extreme to significantly change these distributions. The PAH feature is distinctly visible on the $[4.5] - [5.8]$ vs. $[5.8] - [8.0]$ diagram. The Gutermuth et al. (2009) criteria are indicated by gray lines for comparison.

5.3. *Optical Color-Magnitude Diagram*

Less than half the YSO candidates are optically visible, for example Gaia DR2 detects ∼36,000 of them, which comprise ∼30% of the entire sample. The candidates detected by Gaia tend not to be as red in the mid-IR as other candidates (e.g., $[3.6] - [4.5] \lesssim 1$ and $[4.5] - [8.0] \lesssim 1.2$).

Figure 13 shows a Gaia color-magnitude diagram for the visible YSO candidates. Absolute G-band magnitudes, computed using Gaia parallaxes ϖ, are plotted against Gaia $G - RP$ colors. Only sources with signal-to-noise $\varpi/\sigma_\varpi > 3$ are included, meaning that the sample of 7686 sources is small compared with the total number of YSO candidates. Nevertheless, this sample is useful for evaluating whether the optically bright candidates have properties consistent with pre–main-sequence stars.

On the Gaia color-magnitude diagram, we show isochrones for young stars at several ages ranging from 1 Myr to 50 Myr from the Bressan et al. (2012) mod-

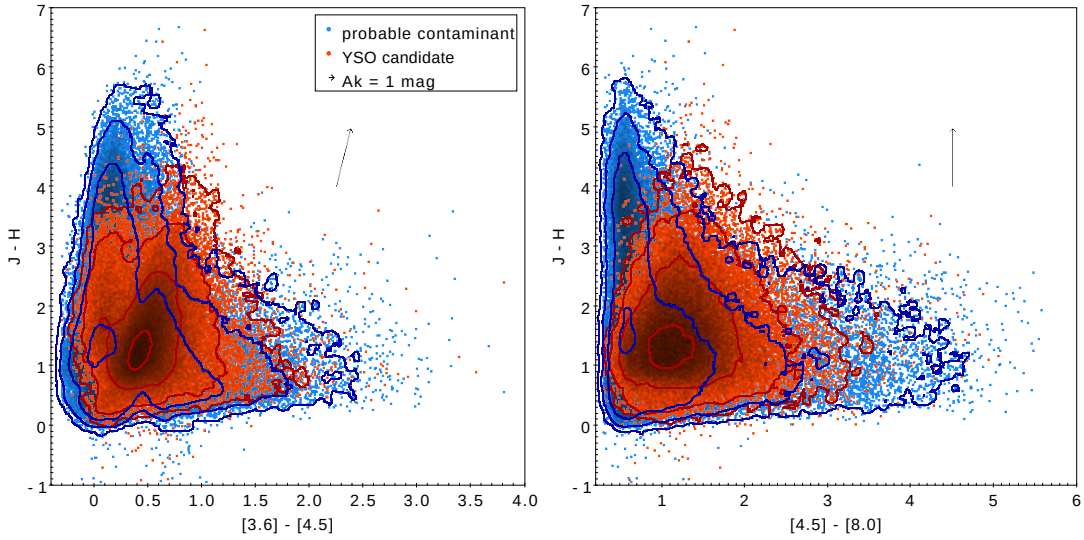


Figure 12. Color magnitude diagrams showing $J-H$ (the color most sensitive to reddening) vs. $[3.6]-[4.5]$ (left) and $[4.5]-[8.0]$ (right), which are both useful for selecting YSOs. Symbols and lines are the same as in Figure 9.

els. We also indicate the effects of reddening, which would shift points down and to the right on this diagram. The wide Gaia bands mean that the effect of reddening depends on the spectrum of the object, so we show three approximate reddening vectors for three colors; more discussion of how this affects selection of pre–main-sequence stars can be found in Herczeg et al. (2019) and Kuhn et al. (2020). Nearly all the candidates lie above the 50 Myr isochrone, and the majority also lie above the 1 Myr isochrone, which is consistent with most of these candidates being very young pre–main-sequence stars.

5.4. *24 Micron Photometry*

When photometry is available in the MIPS 24 μm band (or the W4 band at 22 μm), it can be useful for corroborating classifications based on IRAC. For example, the SED at $\sim$24 μm tends to be more steeply declining for AGB stars, where IR excess is produced in hot dusty winds, in contrast with YSOs' relatively cooler disks and envelopes.

Figure 14 shows the candidate YSOs and contaminants in $J-K_s$ vs. $[4.5]-[24]$ colors. These colors may be useful for distinguishing between AGB stars and YSOs because the typical AGB star has a precipitous rise in the JHK bands followed by a drop in the mid-IR. The figure shows that the YSO candidates tend to be in the middle of the $[4.5]-[24]$ distribution. The objects with $[4.5]-[24] \lesssim 2.4$ are almost all classified as probable contaminants; however, there is a red tail to the probable contaminant distribution, with a high percentage of objects redder than $[4.5]-[24] \gtrsim 7$ also being considered contaminants. An examination of the spatial distribution (not shown) of the probable contaminants in this red tail reveals that many of them are non-clustered higher latitude objects in the Galactic bulge. Of the contaminants with low $[4.5]-[24]$, the distribution of $J-K_s$ ranges from $\sim$0 to $\sim$9, extending redward of the YSOs. Such red $J-K_s$ colors combined with relatively blue $[4.5]-[24]$ colors would be consistent with our hypothesis that many of these probable contaminants are AGB stars.

5.5. *SED Class*

Spectral index in the infrared, defined as

$$\alpha = \frac{\mathrm{d}\log(\lambda f_\lambda)}{\mathrm{d}\log\lambda}, \tag{6}$$

is frequently used to assess the evolutionary stages of YSOs (e.g., Lada 1987; Andre & Montmerle 1994; Evans et al. 2009b; Rebull et al. 2014). However, the value of α depends on what spectral range is used, with the largest available range typically being favored by most studies. Furthermore, the calculation of spectral index may also be affected by reddening. To estimate α values that are minimally affected by reddening, we use the wavelength range from 4.5 μm to 24 μm, since interstellar extinction in these bands is smaller than at shorter wavelengths and the reddening curve is flatter (Indebetouw et al. 2005; McClure 2009; Xue et al. 2016). For these bands,

$$\alpha_{[4.5]-[24]} \approx 0.55\,([4.5]-[24]) - 2.94 \tag{7}$$

$$\alpha_{[4.5]-W4} \approx 0.58\,([4.5]-W4) - 2.92 \tag{8}$$

$$\alpha_{[4.5]-[8.0]} \approx 1.64\,([4.5]-[8]) - 2.82. \tag{9}$$

Where available, we prefer the α estimate based $[4.5]-[24]$, followed by $[4.5]-W4$, and finally $[4.5]-[8.0]$.

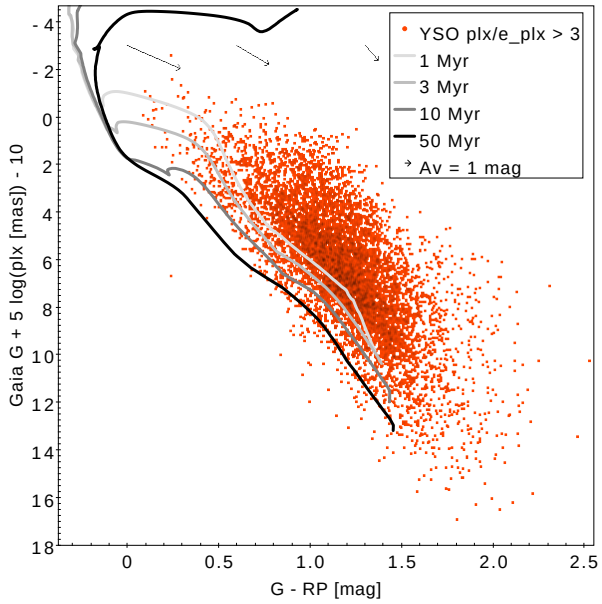


Figure 13. Absolute Gaia G-band magnitude vs. $G - RP$ color for candidate YSOs with $\varpi/\sigma_{\varpi} > 3$. The curves are unreddened isochrones with ages of 1, 3, 10, and 50 Myr from Bressan et al. (2012). The arrows indicate approximate Gaia reddening vectors using the Cardelli et al. (1989) and O'Donnell (1994) extinction curves with $R_V = 3.1$. The broad Gaia bands mean that these vectors vary with color, so we show three vectors estimated using stellar spectra with intrinsic colors of $G - RP = 0$, 0.6, and 1.3. Nearly all of these candidate YSOs are in the region of this color–magnitude diagram consistent with the pre–main sequence.

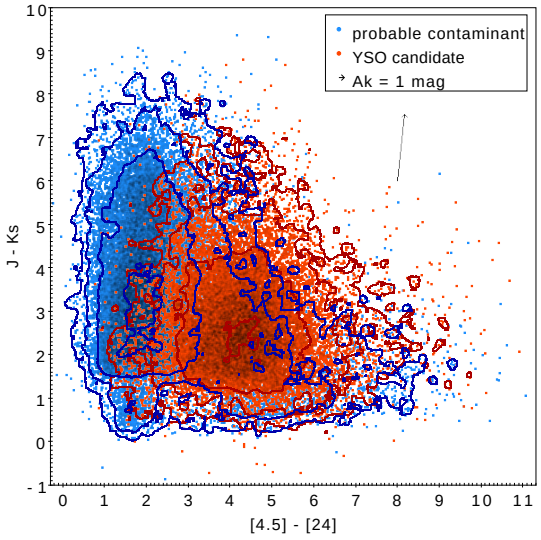


Figure 14. The $J-K_s$ vs. $[4.5]-[24]$ color-color diagram for candidate YSOs and probable contaminants. This diagram may be useful for verifying separation between AGB stars and YSOs. AGB stars typically have steep red SED shapes in the near-IR, but turn over to a Rayleigh-Jeans tail around 24 μm. We find the sources with reddest $J - H$ colors, but not as red $[4.5] - [24]$ colors are mostly classified as probable contaminants, consistent with being AGB stars.

For YSOs suspected of having strong silicate absorption or PAH emission (Sections 5.6–5.7) we do not use the $[4.5] - [8.0]$ color to estimate YSO class because either feature could affect the 8.0 μm band.

Figure 15 shows the distribution of spectral indices calculated for candidate YSOs. Based on these estimates there are 15,943 Class I ($\alpha > 0.3$), 23,810 flat spectrum ($0.3 \leq \alpha < -0.3$), 59,949 Class II ($-0.3 \leq \alpha < -1.6$), and 5,352 Class III ($\alpha \leq 1.6$) YSOs, using the α boundaries from Greene et al. (1994). In addition there are 12,392 candidate YSOs with uncertain class due to missing photometry. This classification scheme roughly reflects the YSO evolutionary sequence from deeply embedded sources with massive envelopes (Class I and flat spectrum) to stars with disks (Class II) and systems where the disk has mostly dispersed (Class III). However, viewing geometry may also affect the assigned YSO class; for example, a YSO that would otherwise be considered Class II may have a Class I SED if viewed at high inclination (Williams & Cieza 2011). Finally, we clarify that even though some classification schemes regard Class III sources as having no IR excess (see Evans et al. 2009b), in our scheme Class III implies weak, but detectable excess.

5.6. *Possible Silicate Absorption*

Broad silicate dust absorption or emission features, centered at ∼9.7 and ∼18 μm, are frequently detected in the mid-IR spectra of YSOs (e.g., Furlan et al. 2006, 2008, 2011; Oliveira et al. 2010). The 9.7 μm feature overlaps the IRAC 8 μm band, so these features can affect YSO colors observed by IRAC.

In the color-color diagram shown in Figure 16 (left panel), a group of ∼2000 YSO candidates stand out due to their unusually blue $[5.8] - [8.0] < 0$ colors – these objects are flagged in Table 1. Given the lack of a red color in $[5.8] - [8.0]$, the classification of these stars as YSO candidates was based mainly on their $[3.6]-[4.5] \gtrsim 0.5$ and $[4.5] - [5.8] \gtrsim 0.5$ colors, both of which tend to be redder than most of the other YSO candidates.

Figure 17 shows three example SEDs that we have fit with YSO models from Robitaille (2017) – for each source the ten best-fitting convolved models are indicated by the gray lines. Robitaille (2017) include multiple configurations of disks and/or envelopes, so we used the simplest model forms capable of explaining the data:

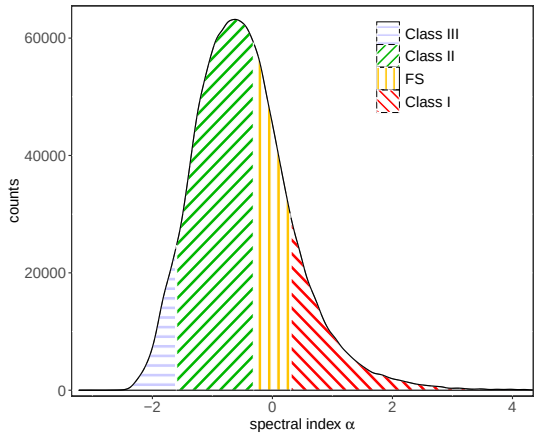


Figure 15. Distribution of spectral index α for YSO candidates, subdivided into YSO class using the customary demarcations at $\alpha = -1.6$, -0.3, and 0.3. The shape of the distribution will be the product of the prevalence of the YSO classes, with Class II/III YSOs being more common than Class I/flat SED YSOs due to the longer lifetimes of the later evolutionary stages (e.g., Evans et al. 2009a), and our sensitivity to each class, which may be lower for YSOs with smaller IR excesses (e.g., Class III) and for deeply embedded YSOs (e.g., Class I).

a star and disk model (sp-s-i[4]) for SPICY 75228; a star, disk, and envelope models with variable inner radius (spu-hmi) for SPICY 85135; and a star and disk model with variable inner radius (sp-h-i) for SPICY 99415. Although these fits are not all formally good given the reported photometric uncertainties, they illustrate the range of SED morphologies that could produced the colors that we observe. Each case requires a strong silicate absorption feature at 9.7 μm to reproduce the lower 8.0 μm band emission. The best models also tend to also have nearly edge-on inclination to provide the high absorbing column density.

Silicate absorption in YSO SEDs can come from the object itself or from foreground interstellar dust (van Breemen et al. 2011). We would expect a YSO with strong intrinsic silicate absorption to have a substantial disk or envelope that can produce the extinction, and Forbrich et al. (2010) find that YSOs with positive spectral indices are more likely to have strong silicate absorption. Figure 16 (right panel) shows our strong silicate absorption candidates on a plot of $[3.6]-[4.5]$ vs. $[4.5]-[24]$. The color $[4.5]-[24]$ is a good indicator for SED spectral index that is not affected by silicate absorption. Most of the objects with possible silicate absorption have $[4.5]-[24] > 5$, higher than average for the YSOs, but ∼20 sources have $[4.5]-[24]$ colors bluer than this. In the interstellar medium, the relation between the optical depth of the 9.7 μm feature and optical extinction is approximately $\tau_{9.7} \sim A_V/20$ (Roche & Aitken 1984; Chiar et al. 2007; Shao et al. 2018). Although most stars in our sample would not have sufficiently high foreground extinction for the feature to become optically thick, this can be achieved along lines of sight that pass through dense molecular clouds or near the Galactic Center.

[4] The designations correspond to models from Robitaille (2017).

5.7. *Possible PAH Emission*

Another salient feature in the $[4.5]-[5.8]$ vs. $[5.8]-[8.0]$ diagram is the finger-like structure at $[5.8]-[8.0] \approx 1.6$. Most of the sources with these colors in the full IRAC catalogs were classified as probable contaminants, but a minority (∼490 objects) were classified as YSO candidates. These colors match those expected for sources dominated by PAH emission bands. For an astronomical PAH emission spectrum, the ratio of flux in the 5.8 μm band to the 8.0 μm band ranges from 0.31 to 0.41 (Draine & Li 2007, and references therein), corresponding to $[5.8]-[8.0] = 1.6$–1.9. There is little PAH emission in the 4.5 μm band, leading to a red $[4.5]-[5.8]$ color. Candidates are flagged in Table 1 for strong PAH emission if they meet the criteria $[4.5]-[5.8] > 2$ and $[5.8]-[8.0] > 1$, which are based on the observed morphology of this feature in color space.

Although IR nebulosity in star-forming regions is dominated by PAH emission, inspection of the flagged YSO candidates suggests that most are valid point sources in all 4 IRAC bands, not spurious detection of nebular knots. For example, ∼90% of these sources have $M = 2$ detections in both the 5.8 μm and 8.0 μm bands, indicating reliable detections. We examined the images of a subset of these objects by eye and found that even in cases with surrounding nebulosity, the sources themselves appeared to match the PSF. PAH emission may be intrinsic to massive YSOs with sufficiently high ultraviolet luminosities (e.g., Whitney et al. 2013). Spitzer IRS spectroscopy of massive YSOs has shown PAH emission to be nearly ubiquitous and correlated with YSO luminosity (Oliveira et al. 2013).

In the MYStIX IR-excess catalog that we used for training, Povich et al. (2013) aggressively filtered sources with PAH emission to avoid contamination by nebular PAH knots. To be included, they required sources to exhibit red $K_s - [4.5]$ colors (avoiding bands with PAH emission), as would be expected for a massive YSO. This requirement will be reflected in our classifications via the random forest classifier. In the SPICY catalog, YSO candidates with possible PAH emission have median $K_s - [4.5] = 3.3$ compared with a median for the entire sample of 2.0.

Figure 16 (right panel) shows the 24 μm emission for these objects. The YSO candidates with possible PAH emission have $[4.5]-[24]$ colors ranging from 5 to 10,

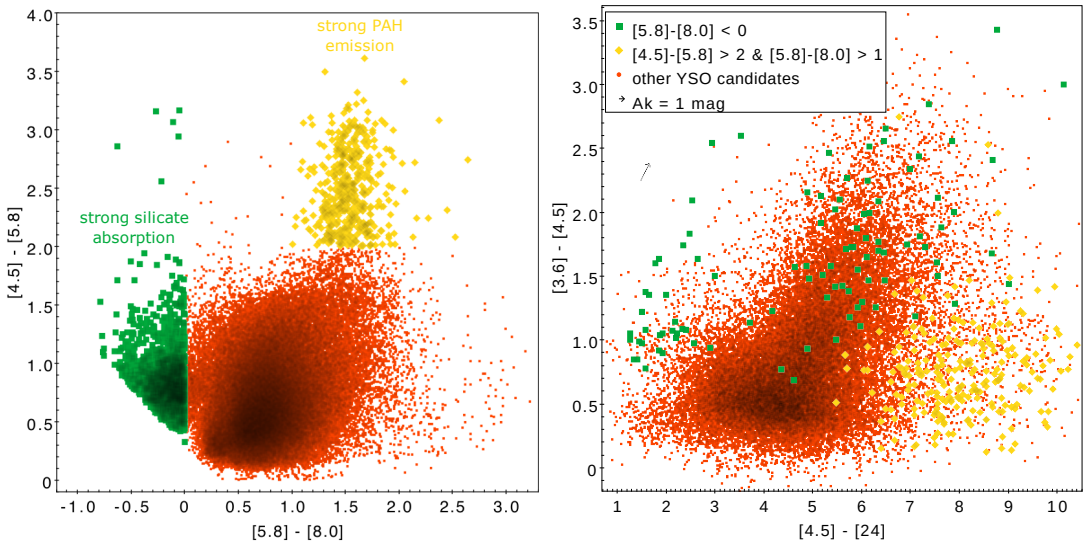


Figure 16. Left: IRAC color-color diagram with YSO candidates with possible strong silicate absorption and PAH emission labeled. Right: YSO candidates on the $[3.6]-[4.5]$ vs. $[4.5]-[24]$ diagram, with several subcategories selected. YSO candidates with low values of $[5.8]-[8.0]$ colors are shown as green squares and YSO candidates with suspected PAH emission are yellow-orange diamonds. Both groups have redder than average $[4.5]-[24]$ colors, which is consistent with these classes of sources being deeply embedded.

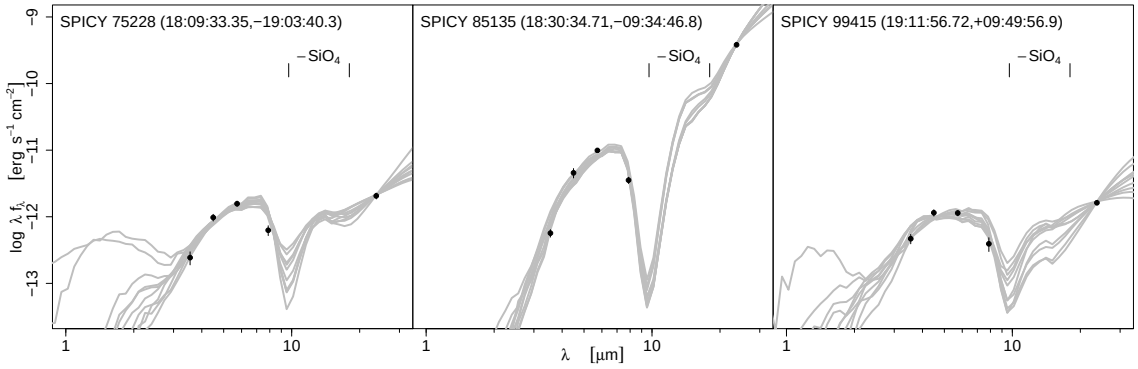


Figure 17. SEDs for three YSO candidates with $[5.8]-[8.0] < 0$. The black points represent the photometry in the IRAC bands, with 1σ error bars. The gray lines are the 10 best-fitting convolved YSO models from Robitaille (2017). Although not all of these models provide formally good fits, they represent the SED shapes necessary to produce the observed IRAC colors. These models demonstrate that strong silicate absorption features (centers indicated by tic marks) are necessary to reproduce observed photometry.

much higher than the average for YSOs. This result is consistent with these objects being massive YSOs.

6. ENVIRONMENT

Many dynamical processes sculpt the interstellar medium in star-forming regions and affect the spatial relationship between the clouds and young stars (Shu et al. 1987; McKee & Ostriker 2007). Infrared nebulosity, however, can be considered a strong proxy for star formation, as newly-formed massive stars illuminate the primordial clouds in the star-forming complex. The Spitzer images reveal features ranging from IR dark clouds to bright PAH-dominated nebulosity, which can trace the photodissociation regions at the edges of clouds and bubbles (e.g., Churchwell et al. 2009; Pari & Hora 2020).

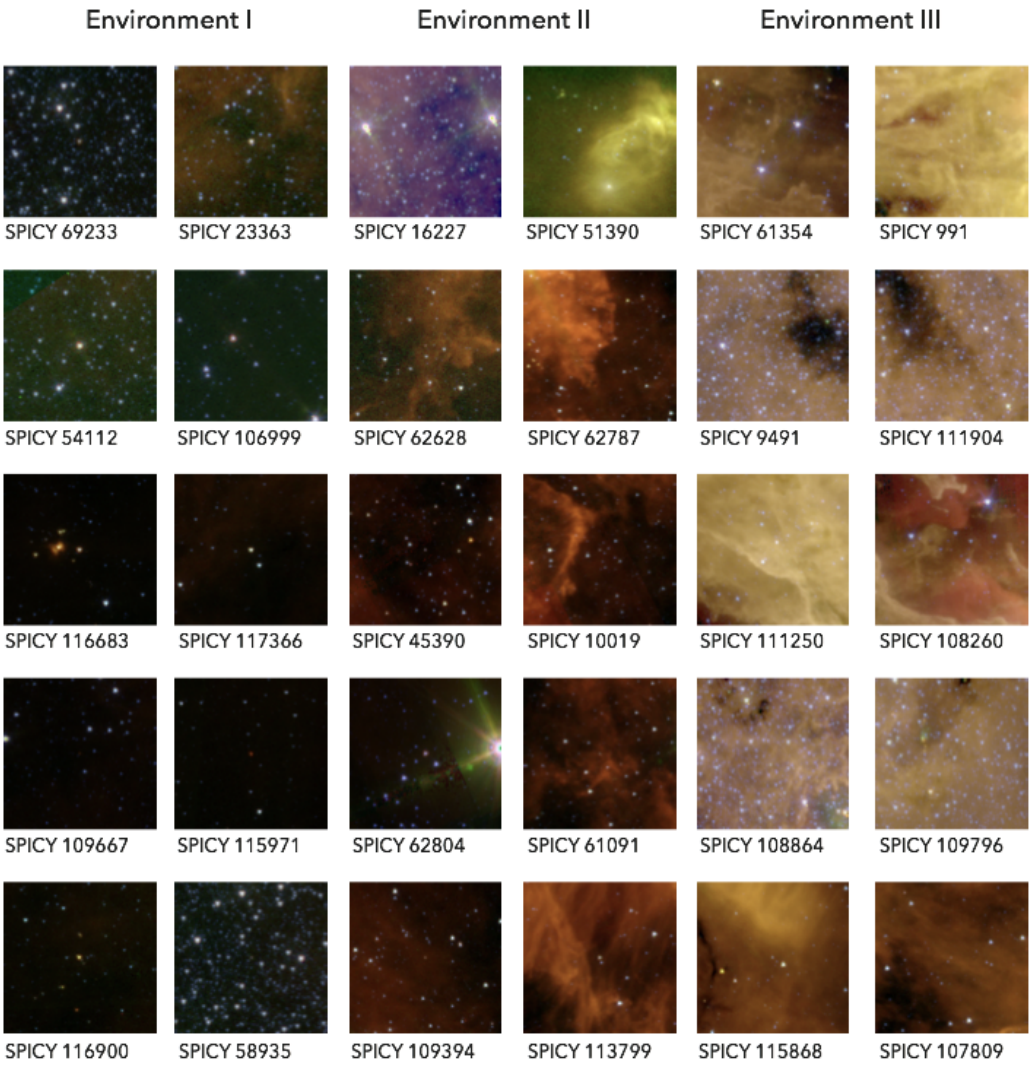


Figure 18. Color $3' \times 3'$ cutouts (IRAC 3.6 μm in blue; 5.8 μm in green; 8.0 μm in red) centered around a sample of YSO candidates from the SPICY Album. The first two columns show examples of stamps from the Environemnt 1 class, corresponding to images with no or minimal nebulosity. The two middle columns correspond to examples of stamps from the Environment 2 class, which is a mixed class mostly containing regions at the transition between the Environments 1 and 3 or miss-classifications from those two extreme classes. The last two columns show examples of regions classified as Environment 3, that clearly correspond to cloud-like environments.

To facilitate the study of the local environments around YSOs, we have created an album of $3' \times 3'$ image cutouts[5] in all four Spitzer bands with additional false-color image files ready for visual inspection or generic image processing frameworks (e.g., Yang et al. 2012). The false-color images (see examples in Figure 18) were created with a heuristic based on Lupton et al. (2004), mapping the IRAC 3.6 μm to the blue channel, 5.8 μm to green and 8.0 μm to red. Here, we applied a hyperbolic arcsin transform to each IRAC band, and we selected the range of the color intervals from the mode of the distribution of the lower 2×10^{-2}% of the pixels for the minimum value, and the mode of the distribution of the upper 6×10^{-5}% pixels for the maximum. These values were chosen to optimize the visual experience while minimizing information loss and excluding extreme outlier pixels. The modes were estimated using the Venter (1967) estimator, as implemented by the MODEEST package (Poncet 2019).

The SPICY album comprises a total of 117,224 PNG stamps. A total of 222 YSOs candidates from the SPICY catalog miss their stamps due to numerical problems in the original FITS files and/or the lack of response from the IPAC archive in one or more bands at the time of the album creation. Its size is 251 GB, and all PNG and FITS files are publicly distributed and archived long-term at Zenodo hosted at CERN facilities.

6.1. *A Simple Characterization*

Below, we demonstrate an example application for these cutouts, using a simple unsupervised image clustering strategy to characterize environments in which the YSOs candidates are found.

We avoid clustering in the pixel space because it is not invariant to image translations and rotations, which are properties that any proper content-based image clustering solution should have. Two candidate transforms that can introduce these these properties via the power spectrum are wavelets (as used in Krone-Martins et al. 2019, for a similar application) and Fourier transforms (e.g., Kauppinen et al. 1995; van der Schaaf & van Hateren 1996); here, we adopt the latter. This is partially motivated because the Fourier power spectra is linked to the turbulent properties of the star formation medium (e.g., Elmegreen & Scalo 2004), revealing signatures of different physical phenomena.

We first compute the 2D Fourier power spectra of each cutout in each IRAC band. Then, we compute 1D radially medianized power spectra from each of the original 2D power spectra and concatenate these 1D power spectra to form a vector for each YSO candidate. Next, we organize the vectors of all environments into a single matrix and perform principal components analysis (PCA; Pearson 1901; Hotelling 1933), from which we select the most relevant dimensions (see also Ishida & de Souza 2013; de Souza et al. 2014, for PCA variants), which acts as feature compression (see e.g., Sasdelli et al. 2016). Finally, we model the distribution using a multivariate Gaussian mixture model (GMM; Pearson 1894; Scrucca et al. 2016; de Souza et al. 2017; Melchior & Goulding 2018) in the space defined by the first two principal components of the power spectra and the modes of the pixel values in each cutout, which we transform using an inverse sinh function. The distribution in this space is complex and requires many (25) Gaussian components, with model selection using the Bayesian information criterion (Schwarz 1978).

Visual inspection shows that the GMM components tend to correspond to three types of environments: those that are nebulosity-free (or minimal nebulosity) environments, mixed environments, and cloud-like environments. These are labeled environments 1, 2, and 3 in Table 1. We also found outliers on the boundary of the distribution. Examination of the cutouts showed that the outliers correspond to severe image reconstruction errors and/or missing data in one or more bands and are located at the edges of the surveys.

A total of 66,539 stamps, or ∼57%, of the valid stamps, were classified as cloud-like, while 32,790 stamps, or ∼28%, were classified as nebulosity-free. The mixed class and the outliers correspond to 15,462 and 2,433 stamps, or ∼13% and ∼2%, respectively. These numbers indicate that most YSO candidates are indeed in cloud-dominated environments, as would be expected for YSOs in star-forming regions. However, the number of candidates in environments presenting diffuse nebulosity or even no detectable nebulosity is not negligible.

The cloud-like environments are most prevalent in the inner regions of the Galaxy, between approximately $\ell = 300°$ and $50°$ and $|b| \leq 1°$. Cloud-like environments are also associated with large star-forming complexes outside this coordinate range, including some of the star-forming regions in Cygnus X.

The mixed environment is most prevalent further from the Galactic Center (e.g., $\ell \leq 310°$ or $\ell \geq 30°$). Some stellar associations include both cloud-like and mixed classes (e.g., the Carina Nebula).

The image cutouts with no or minimal nebulosity are found throughout the entirety of the survey, except within ∼1° of the Galactic Center. Stars in this environment are the most evenly distributed, but even among these stars several clusters can be seen. For example, the Sco OB1 Association is made up of both cloud-like and nebulosity-free classes.

[5] To produce the album, we constructed an infrastructure to query the IPAC archive at http://irsa.ipac.caltech.edu that tracks the FITS transfers and that also tracks and verifies the local generation of the PNG stamps. This infrastructure makes use of a PostgreSQL database (PostgreSQL Global Development Group 2020) and is parallelized. However, we kept the number of parallel data transfers from IPAC low to avoid overloading their servers, enabling the extraction of all IRAC images and the construction of all the stamps in about three days.

This simple application can certainly be significantly improved by adopting tailored methodologies and signal representations, for instance, curvelet transforms (e.g. Candès & Donoho 2000; Starck et al. 2003), to characterize the signal power contained in filamentary structures, and customized clustering methods. Moreover, a proper physical characterization of the YSO environment requires consideration of effects of, for example, the distances to the objects, accounting for the distinct physical scales probed, differences of PSFs in different IRAC bands and their impacts on the power spectra, projection effects of the nebulous matter in the plane of the sky, etc. However, as we show here, even a simple analysis already reveals that, curiously, a significant fraction of the YSO candidates in the SPICY catalog do not seem to be lying in environments dominated by clouds.

7. SPATIAL CLUSTERING

Spatial aggregation is a well recognized property of YSOs (e.g., van den Bergh 1964; Carpenter 2000b; Allen et al. 2007) that can be observed in the distribution of our YSO candidates (Figure 5). However, the best clustering algorithm to use for stars, or even what is the most meaningful definition of star cluster, is not obvious (see Kuhn & Feigelson 2017; Gouliermis 2018; Ascenso 2018). For example, different groups may have vastly different numbers and densities of stars, and the spatial distributions are complex and often fractal-like. Thus, different cluster analysis methods that yield different segmentations may be appropriate for different scientific applications (e.g., Everitt et al. 2001).

We choose the algorithm "Hierarchical Density-Based Spatial Clustering of Applications with Noise" (HDBSCAN; Campello et al. 2013), which has been successfully applied to Gaia DR2 data to detect hundreds of new open clusters (e.g., Kounkel & Covey 2019; Castro-Ginard et al. 2020). The HDBSCAN algorithm (Campello et al. 2013) allows for groups of stars with different numbers, densities, and morphologies, it permits stars to not belong to any group, and it can be applied across the entire survey area in a uniform way providing a reasonable looking clustering solution. We apply this algorithm to each of the contiguous survey regions using a Python implementation[6]. The main parameter we choose is the minimum number of stars in a group, which we set to $n = 30$, and we run the algorithm using the "excess of mass" method for cutting the tree. The algorithm is run on the Galactic ℓ and b coordinates for each contiguous segment of the IRAC survey area. The resulting groups are not necessarily gravitationally bound systems, but rather collections of YSO candidates that appear to be spatially aggregated. The two groups, labeled in Figure 6, were selected using this algorithm. A list of these groups, along with their properties, are provided in Table 2.

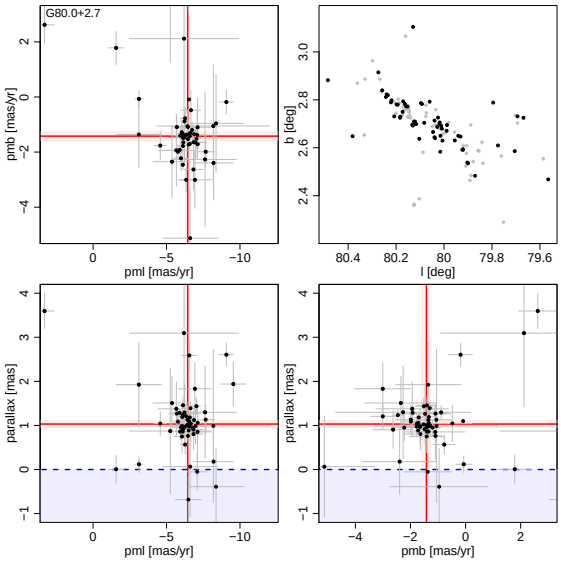


Figure 19. Scatter plots of astrometric properties for YSO groups; G80.0+2.7 (a group in the Cygnus X field) is shown as an example. These plots include proper-motion vs. proper motion, parallax vs. proper motion, and (ℓ, b) positions. The estimated means from the hierarchical Bayesian model are shown by the red lines, and 2σ formal uncertainties are illustrated by the pink shaded areas. Values excluded by our prior are shaded blue. Stars with Gaia DR2 5-parameter astrometry are black circles with gray 1σ error bars, and stars with only position information are gray diamonds.
(The complete figure set (406 images) is available in the online version.)

The method found 406 stellar groups, collectively including 58,084 (= 49%) of the YSO candidates. This suggests that roughly half the YSO candidates are spatially clustered while the other half are more widely distributed. The choice of n does affect the solution, particularly whether groups are subdivided into smaller groups are unified into larger groups. However, the percentage of stars in groups stays relatively constant (i.e. 47–56%) when n is varied from 15 to 60. The median angular diameters of the groups increase from $\sim$0.2° to $\sim$0.7° over this range of n, with a value of $\sim$0.4° at $n = 30$. We pick $n = 30$ because the resulting solution appears to avoid chaining together unrelated stars over large areas of the sky, but the groups are large enough to include enough Gaia sources for their astrometric properties to be estimated. An examples of a complicated structure identified by us as a single group is the Carina Nebula complex, an association made up of multiple star clusters. We also note that HDBSCAN

[6] https://hdbscan.readthedocs.io/en/latest/api.html

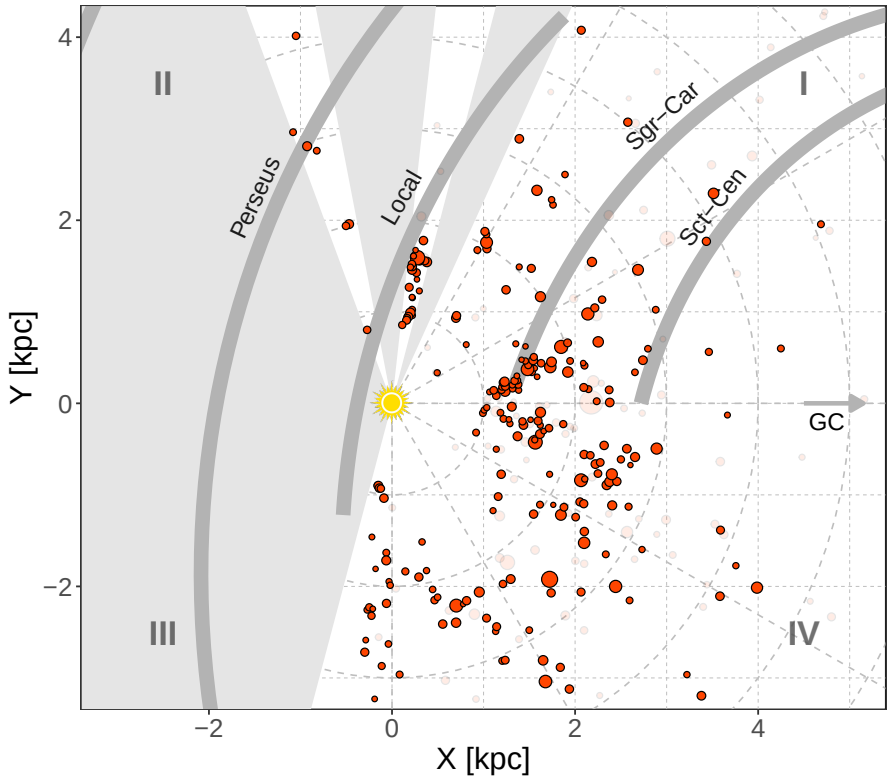


Figure 20. Spatial distribution of YSO groups in heliocentric Galactic XY coordinates. Groups with more reliable distances are depicted by the darker red circles, while those with more uncertain distances ($\varpi/\sigma_\varpi < 2$ or flagged) are lighter red. Circle sizes are proportional to the total numbers of members. The approximate centers of spiral arms from Reid et al. (2019) are indicated by the gray curves. The Sun's location is indicated by the yellow symbol. Wedges of the XY plane not covered by the catalog are shaded gray. The Galactic quadrants and the direction of the Galactic Center are labeled. For conversion of Gaia DR2 parallaxes to distance in this figure, we use the −0.0523 mas zero-point offset estimated by Leung & Bovy (2019)

collects the over-density of candidate YSOs toward the Galactic center into a single group labeled "G0.2-0.1"; these stars do not all come from the same star-forming regions but this region of the Galaxy is challenging for our algorithms.

7.1. *Galactic XY Distribution*

To estimate the heliocentric distances ($d_\odot$) to each of the YSO groups we employ a hierarchical Bayesian model (Hilbe et al. 2017) to account for the measurement errors in parallaxes and presence of outliers. The use of robust statistics is particularly suitable given the non-negligible presence of unknown contaminants in each group. Normality assumptions are sensitive to noise and outliers, which may result in a biased estimate of the mean distance. Replacing a Gaussian likelihood by a t-distribution is a relatively easy fix. The t-distribution has an extra ν parameter called degrees of freedom, which controls close the distribution resembles the normal distribution. Larger values $\nu > 30$ essentially recover the normal distribution, while smaller values result in a distribution with heavier tails. This extra flexibility enables it to adapt to the extra noise in the data, without introducing a bias in the underlying relationship.

The model formulation for the robust estimate is given below, where we define a t-likelihood for the observed ϖ, and suitably vague priors on all the model parameters: uniform for $d_\odot$ over 25 kpc, and a gamma (Γ) prior (to ensure positivity) for ν.

$$\begin{aligned}
&\varpi_{\rm i} \sim \mathscr{T}(1/{\rm d}_\odot, \sigma^2_{\varpi_{\rm i}}, \nu),\\
&\nu \sim \Gamma(2, 0.1),\\
&d_\odot \sim {\rm Uniform}(0, 25),\\
&i = 1 \ldots n_{\rm Gaia}
\end{aligned} \tag{10}$$

The index i runs over the members of each group $n_{\rm Gaia}$ with Gaia astrometric information. Although distance is constrained to be positive, our likelihood model permits the parallax measurements for individual stars, ϖ_i, to be either positive or negative. We evaluate the model using a Gibbs sampler, for which we use the JAGS[7] package (Plummer 2017) within the R language. We initiate three Markov Chains by starting the Gibbs sampler at different initial values. Initial burn-in phases were set to 5,000 steps followed by 5,000 integration steps for each YSO group, which are sufficient to guarantee the convergence of each chain.

Table 2 provides group parallaxes and proper motions estimated from the posterior medians. Uncertainties are estimated from the mean absolute deviation (MAD) of the posterior (scaled to approximate 1σ uncertainties) and added in quadrature to the ± 0.04 mas and ± 0.07 mas yr^{-1} spatially correlated systematic errors on DR2 zero points (Lindegren et al. 2018). Out of 406 groups, 402 have some Gaia astrometry, giving at least a rough estimate of parallax and proper motion. Of these, most groups include at least $n_{\rm Gaia} = 10$ members having Gaia 5-parameter astrometric solutions, enabling estimates that are more precise than those based on individual stars.

For each group, we show scatter plots of stellar proper motions, parallaxes, and positions (Figure 19), with the groups' mean parallaxes and proper motions indicated. In most cases, the stars form a single clump in $\mu_{\ell\star}$–μ_b–ϖ space, suggesting that most of the group members are spatially and kinematically associated. In other cases (e.g., G77.8+1.0), multiple clumps are apparent, which may imply that distinct stellar groups with chance alignment have been merged by the HDBSCAN algorithm. In the example G77.8+1.0, the estimated properties correspond to the more distant but more numerous of the two groups. We visually inspected all groups in Figure 19 and flag those in Table 2 where problems could affect the interpretation of the Bayesian models, such as the suggestion of multimodality, groups that appear dominated by field stars, or a single data point with too much leverage. We also flag groups with $n_{\rm Gaia} < 3$.

The locations of these stellar groups in heliocentric Galactic XY coordinates are plotted in Figure 20. For this plot, we converted parallaxes to distance using an average -0.0523 mas Gaia DR2 zero-point correction estimated by Leung & Bovy (2019). Then $X = d_\odot \cos(b)\cos(\ell)$, and $Y = d_\odot \cos(b)\sin(\ell)$. The estimated centers of spiral arms from Reid et al. (2019) are also indicated. Fainter red points are groups with $\varpi/\sigma_\varpi < 2$ or groups that have been flagged. Few YSO groups are detected within 1 kpc, but the footprint of the GLIMPSE (and GLIMPSE extensions) surveys exclude many of the nearest star-forming regions, which are located more than several degrees above or below the Galactic midplane. The bulk of the YSO candidates for which we have accurate measurements have heliocentric distances that range from 1 to 3 kpc. There may be some bias in the distances we are sensitive to because this range resembles the range in distance of objects in our training set. Nevertheless, there are groups that appear to lie beyond ∼3 kpc, but Gaia-based distance become more uncertain at this range.

The YSO groups are not distributed smoothly within the Galaxy, but instead reveal Galactic structure. The survey areas intersect several several spiral arms, and crucially provide information about Galactic structure at the boundary between Quadrants I and IV, where structures traced by $v_{\rm lsr}$ measurements of gas become degenerate. We discuss the relation of the stellar groups to the spiral arms below.

Local (Orion) Arm: In Quadrant I and II, the local Arm intersects both the SMOG field and the Cygnus X field. In SMOG, there is one group at the approximate distance of this arm. We find the YSO groups in Cygnus X to be spread linearly, spanning a factor of ∼2 in distance (1–2 kpc). This situation is consistent with looking down the length of the arm. Cygnus X is thought to lie at an end of a long molecular filament (Alves et al. 2020; Zucker et al. 2020) that contains multiple prominent star-forming regions (e.g., Orion, Taurus, the North America Nebula). The additional length of Cygnus X may increase the total length of this structure by 50%. A similar result is reported by Xu et al. (2016). In Quadrant IV, a chain of groups is located at $X \sim 0$ and extends from ∼1 to ∼3 kpc in distance toward the constellation Vela. The orientation of this chain suggests that these groups could connect with the Local Arm.

Sagittarius-Carina Arm: Numerous star-forming regions can be found in the inner 20° of the Galaxy, including some famous regions like the Trifid Nebula, the Lagoon Nebula, and NGC 6334. Some of these groups (including the aforementioned famous regions) form a coherent, chevron-shaped structure, that has a vertex pointing toward us at a distance of ∼1.2 kpc and edges that extend away with lengths of ∼1 kpc. The vertex is at the approximate distance of the Sagittarius-Carina Arm from Reid et al. (2019), implying that the arm is an active site of star-formation activity. How-

[7] http://cran.r-project.org/package=rjags

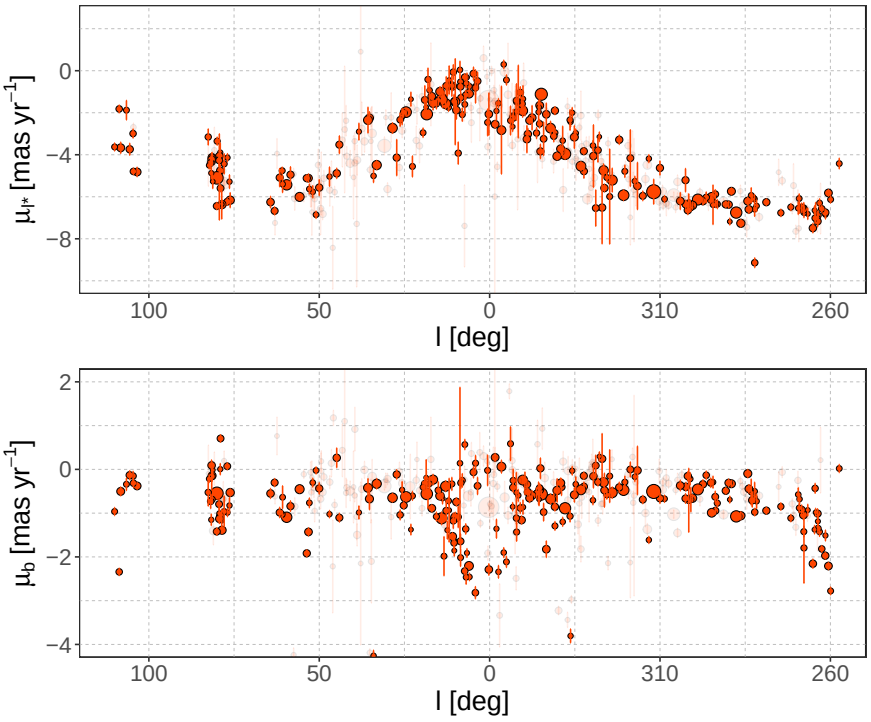


Figure 21. Position-velocity diagrams for the YSOs candidates: $\mu_{\ell^\star}$ vs. ℓ (top) and μ_b vs. ℓ (bottom). The distributions are the cumulative effects of Galactic rotation, Solar motion, and peculiar velocities of YSO associations. Size and shading of circles is the same as Figure 20. Error bars combine the statistical uncertainties on group motions with the ± 0.07 mas yr^{-1} Gaia DR2 systematic uncertainty.

ever, the angles of the edges of the chevron are inconsistent with the angle of the arm predicted by Reid et al. (2019). The linear structure making up the edges of the chevron cannot be a result of the "Fingers of God" effect because it is not oriented along our line of sight. We find relatively little sign of YSO groups associated with the Sagittarius-Carina Arm at Galactic longitudes beyond $\ell > 30°$ in Quadrant I or between $300 < \ell < 330°$ in Quadrant IV.

Scutum-Centaurus Arm: This arm is less clearly delineated by stellar groups than the others, possibly owing to large distance uncertainties at the distance of this arm. However, there is an increase in the density of groups near this arm in Quadrants I and IV.

Perseus Arm: This arm is intersected by the SMOG field, and three groups have distance estimates consistent with the center of this arm from Reid et al. (2019). The large distance of this arm may decrease our sensitivity to YSOs associated with it.

Inter-arm: There are multiple stellar groups that appear to be located between the spiral arms from Reid et al. (2019). For example, within $\sim 10°$ of the Galactic Center, many groups appear located between the Sagittarius-Carina and Scutum-Centaurus arms. In Quadrant I, several groups are located between the Local arm and the Sagittarius-Carina Arm.

7.2. *Galactic Rotation*

The procedure for calculation of mean proper motions for the groups is similar to the calculation of heliocentric distances, using a weakly Gaussian prior for $\mu_{\alpha^\star,0}$ and $\mu_{\delta,0}$ instead.

Figure 21 displays the proper motions in Galactic longitude and latitude as function of ℓ. The expected distribution for stars in circular Galactic orbits would be governed by Galactic rotation, parameterized by the Oort (1927) constants, and effects from Solar motion, which are distance dependent (Bovy 2017). On the $\mu_{\ell^\star}$ vs. ℓ diagram, to first order in the Galactic plane, this would be a sinusoid with period 180° for Galactic rotation added to a sinusoid with period 360° for Solar motion. This overall structure appears to dominate the $\mu_{\ell^\star}$ vs. ℓ plot

for our YSO groups, but there are hints of deviations that we will discuss in more detail in a subsequent paper.

On the μ_b vs. ℓ diagram, several structures can be seen. For example, between $\ell \sim 5$–$25°$, there is a diagonal chain of groups in position–proper-motion space. These groups correspond to one of the edges of the chevron-like structure detected in the XY diagram. The dispersion in μ_b is slightly higher around $\ell \approx 75°$ (Cygnus X) and around $\ell \approx 260°$ (Vela) – both of which correspond to extended structures along the line of sight.

7.3. *Spatial distribution of YSOs by Class*

YSO candidates of all SED classes are spatially clustered, but the classes corresponding to earlier evolutionary stages tend to be more strongly clustered, as can be seen in the section of the Galactic midplane in Figure 22 (left panel). In this region near the young cluster NGC 6823, Class I sources mostly lie within the densest groups, but the other classes are comparatively more distributed. The K function (Ripley 1976) can be used to quantitatively compare the relative strength of clustering for these populations. We used the SPATSTAT package (Baddeley 2017) to estimate K as a function of angular separation r for Class I, flat spectrum, Class II, and Class III objects with 95% confidence intervals estimate using the bootstrap method of Loh (2008). On this log-log plot (Figure 22, right panel), at angular separations of several arcminutes, the slope for Class I YSOs is significantly flatter than for flat spectrum YSOs, which is also significantly flatter than for Class II and III YSOs, implying that the earlier stages are more clustered. This finding agrees with numerous other examinations of the spatial distribution of sources by YSO class (e.g., Sung et al. 2009; Samal et al. 2010; Buckner et al. 2020). We note that even some candidate Class I YSOs appear isolated, for example ∼100 of these objects (<1% of the Class I YSOs) are separated from their nearest neighbors in our catalog by more than $10'$.

Figure 23 shows the smoothed distributions of sources of various classes in both Galactic longitude and latitude. In longitude, the normalized distributions of YSOs of all SED classes are similar, whereas, in latitude, the distributions of earlier classes (e.g., Class I and flat spectrum) are more strongly concentrated near the midplane than the later classes (e.g., Class II and III). This may be a result of the dispersal of YSOs, if stars are born in regions nearest the midplane, and then drift away. For example, a YSO traveling at a tangential velocity of $\sim$2 km s^{-1} at a distance of ∼2 kpc could travel 0.25° from its point of origin in ∼5 Myr. This is enough to flatten the distribution of b shown in the figure but not the distribution of ℓ.

The sources with strong silicate absorption are more concentrated toward the Galactic Center than other YSOs. This could be an effect of the higher interstellar dust column densities in this direction. The YSOs with strong PAH emission also appear to be preferentially concentrated toward the inner Galaxy, but the peak of the distribution appears to be in star-forming regions around $\ell \sim 330°$.

8. OPTICAL VARIABILITY

Optical variability, with amplitudes ranging from several tenths of a magnitude to outbursts of multiple magnitudes, is associated with YSOs (e.g., Joy 1945; Herbig 1954; Cody & Hillenbrand 2018) and has even been used as a criterion for identifying previously unrecognized YSOs (e.g., Contreras Peña et al. 2017). Thus, strong optical variability from YSO candidates in our Spitzer selected sample can be regarded as corroborating evidence for the youth of these objects. To investigate which sources show optical variability, we use photometric measurements from the Zwicky Transient Facility (ZTF; Bellm et al. 2019), which is sensitive to a variety of variability phenomena from YSOs with its cadence of approximately 1 observation per night (Graham et al. 2019), including dips due to occultation from circumstellar dust, variations in accretion rate, magnetic flares, and rotational modulation due to large star spots.

We cross-match our YSO candidates to the ZTF DR3 (Masci et al. 2019) catalog using a match radius of $1''$, and use ZTF sources with at least 10 measurements in the r band between April 2018 and June 2019, excluding observations from the high cadence deep-drilling program; the median number of observations is ∼130. This yields 7,585 YSO candidates with usable ZTF light curves. This represents a relatively small fraction of our entire catalog because many of the Spitzer sources are not detected in the optical and ZTF is only available for the Northern Hemisphere. Nevertheless, in absolute numbers of sources this sample is moderately large and useful for statistical analysis. To characterize variables, we calculate the r-band light curve's standard deviation $\sigma_{\rm var}$, the mean magnitude $\bar{r}$, and skewness of the distribution.

Figure 24 shows $\sigma_{\rm var}$ vs. $\bar{r}$ for our YSO candidates (red points) as well as ∼2400 randomly selected field stars (gray points) from the same region of the sky. For the field stars, we have estimated the running median $\sigma_{\rm var}$ as a function of $\bar{r}$, shown as the solid black line. 88% of the YSO candidates have $\sigma_{\rm var}$ values greater than this line, indicating that the YSO candidates have higher variability on average than the field stars. Following Fang et al. (2020), we use the median $\sigma_{\rm var}$ for field stars to delineate variability thresholds for the YSO candidates. Objects where $\sigma_{\rm var}$ is more than 3 times the median value for field stars are considered to show high variability (above the dotted line), objects between 2–3 times the median value are considered to show moderate variability (between the dashed and dotted lines), and objects <2 times the median value have low or insignificant variability (below the dashed line). Using these definitions we find 1695 with high variability, 914 with

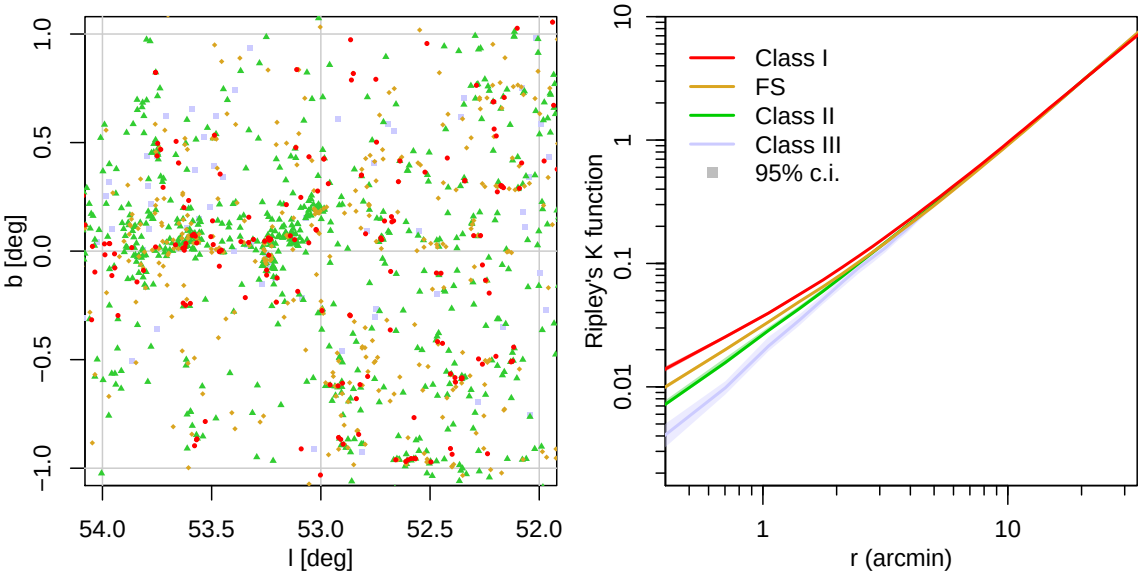


Figure 22. Diagnostics of spatial clustering for stars of different classes. The same color-coding is used to represent each class in both panels. Left: The spatial distribution of YSO candidates in a $\sim 2^\circ \times 2^\circ$ sample area in Sagitta. Right: Ripley's reduced second moment $K(r)$ function, with 95% confidence intervals calculated for all stars. The flatter slope for the Class I and flat-SED YSOs at small angular separations implies these sources are more strongly clustered than the Class II and Class III objects. All classes exhibit some spatial clustering, but the earlier evolutionary classes tend to be more clumped. Nevertheless, examples of isolated YSOs of all classes can also be found.

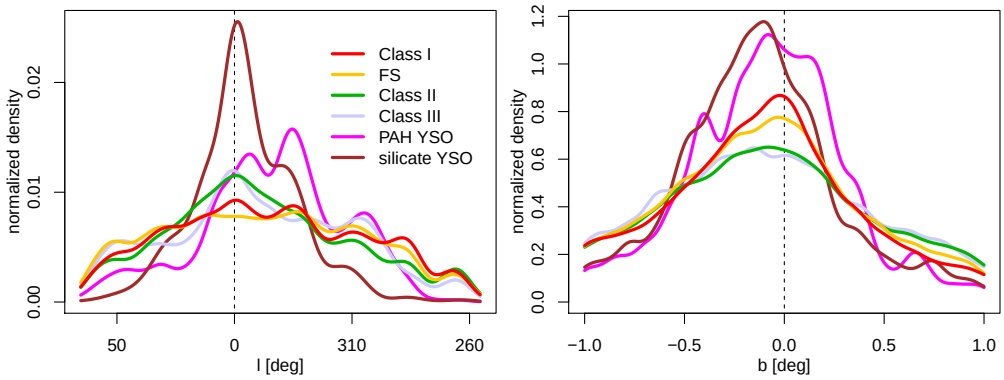


Figure 23. The distributions of ℓ and b for stars of different YSO classes and stars with probable PAH emission or silicate absorption. Densities are calculated with a 5° kernel in ℓ and a 0.05° kernel in b. Only GLIMPSE I, II, 3D, and Vela-Carina data are included in this plot. In Galactic longitude, the distributions of Class I/FS/II/III stars are similar, but in Galactic latitude, the younger YSO classes are more strongly peaked toward the midplane. Sources with silicate absorption are strongly peaked near the Galactic center, while the peak in PAH source density is offset by $\sim 30^\circ$ from the center.

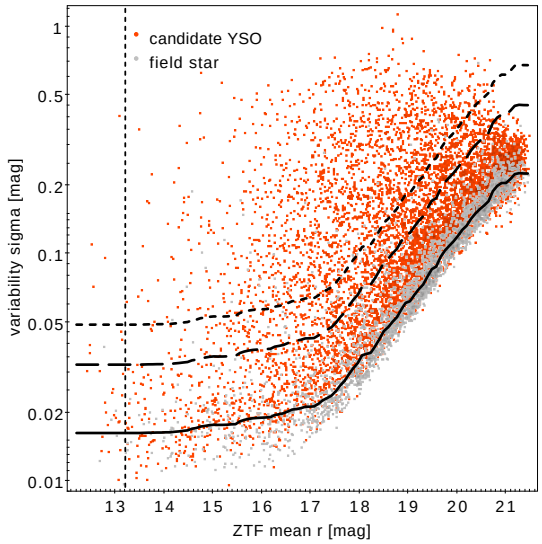


Figure 24. The standard deviation of the ZTF light curve's variability in the r-band ($\sigma_{\rm var}$) vs. mean r for 7,585 YSO candidates (red) and ∼2,400 randomly selected field stars (gray) from the same areas of the sky. The solid, black line is the median $\sigma_{\rm var}$ as a function of magnitude for the field stars, and the lines above are 2 times (dashed) and 3 times (dotted) this level. We label stars above the dotted line as having "strongly variability," between the dashed and dotted lines "moderately variability," and below the dashed line "low or insignificant variability."

moderate variability, and 4976 with low or insignificant variability.

We find a slight tendency for strong or moderate YSO candidates to be more spatially clustered than weak or non-variable YSO candidates. For example, 56% of stars with strong variability are members of HDBSCAN groups, while 51% of stars with moderate variability are members, and 47% of stars with weak or no variability are members. If non-variable candidates have a higher probability of being non-YSO contaminants, this could influence the observed trend, since contaminants are not expected to be clustered. However, even among the YSO candidates with high variability, 44% are not members of HDBSCAN groups, providing further evidence suggesting that many of the relatively isolated candidates may still be legitimate YSOs.

Figure 25 shows a sample of light curves from sources showing high variability. Many of these stars exhibit dipping features with sharp bottoms, a morphological feature typically associated with extinction from dust (possibly in the circumstellar disk) briefly passing in front of the stars. The timescales of the dips seen in these ZTF light curves can range from several days to multiple months. Some light curves show long timescale trends, like the gradual brightening seen SPICY 110421. Other stars' light curves exhibit outbursts, with SPICY 116663 shown as an example with a particularly large >3 mag amplitude. These features are similar to the categories of YSO variability identified by Cody & Hillenbrand (2018), albeit many of the structures identified in their K2 study occur on a shorter timescale than we are sensitive to with the cadence of ZTF. Some our candidates YSOs also exhibit periodic behavior, which is thought to be associated with the rotation periods of the stars due either to star spots or material orbiting at the co-rotation radius (Herbst et al. 1994; Stauffer et al. 2017). SPICY 108092 is an example of one such star, which includes periodic rotation along with dips, bursts, and long time-scale changes.

The ZTF YSO light curves exhibit considerable diversity in their morphologies. Given that the objects in the SPICY catalog were selected in a uniform way independent of their variability, this dataset may be useful as a training set for future efforts to develop a classifier of YSOs based on optical variability.

9. COMPARISON TO OTHER YSO CATALOGS

Both we and Robitaille et al. (2008) search GLIMPSE catalogs for YSOs, but our catalog extends the search to much fainter magnitudes. We have less stringent source quality criteria, we do not impose an *ab initio* $[4.5]-[8.0] \geq 1$ color cut, and, most significantly, we include sources fainter than the flux limits imposed in their catalog. Their 10 mJy limit in the 8.0 μm band ($[8.0] < 9.52$ mag) would discount 73% of our YSOs. In the overlapping regions, the GLIMPSE I and II survey areas, we identify >4 times more YSO candidates than the red sources from Robitaille et al. (2008) catalog.

Unlike our catalog, most of the sources from Robitaille et al. (2008) are bright enough to have been detected by observations at 24 μm. For these objects, they use a simple heuristic set of color criteria to separate YSO and AGB candidates: sources with $[4.5] \leq 7.8$ or $[8.0]-[24] < 2.5$ are considered likely AGB stars, while sources with $[4.5] > 7.8$ and $[8.0]-[24] \geq 2.5$ are likely YSOs. They acknowledge that a division like this is likely to produce erroneous classifications in either direction. Out of 16,670 "red sources" from Robitaille et al. (2008) that they labeled either YSO (9,387) or AGB (7,283), 13,290 (80%) were re-identified as candidate YSOs by our analysis, including 8,637 (92%) of those labeled YSO and 4,653 (64%) of those labeled AGB. Assuming the Robitaille et al. (2008) classifications are accurate, and extrapolating to fainter objects, this could suggest that up to ∼35% of our YSO candidates are misclassified. However, of the 4,653 sources classified as YSOs by us but as AGB by Robitaille et al. (2008), 33% are members of the spatial clusters from Section 7, suggesting that some of the objects they label AGB stars could be YSOs.

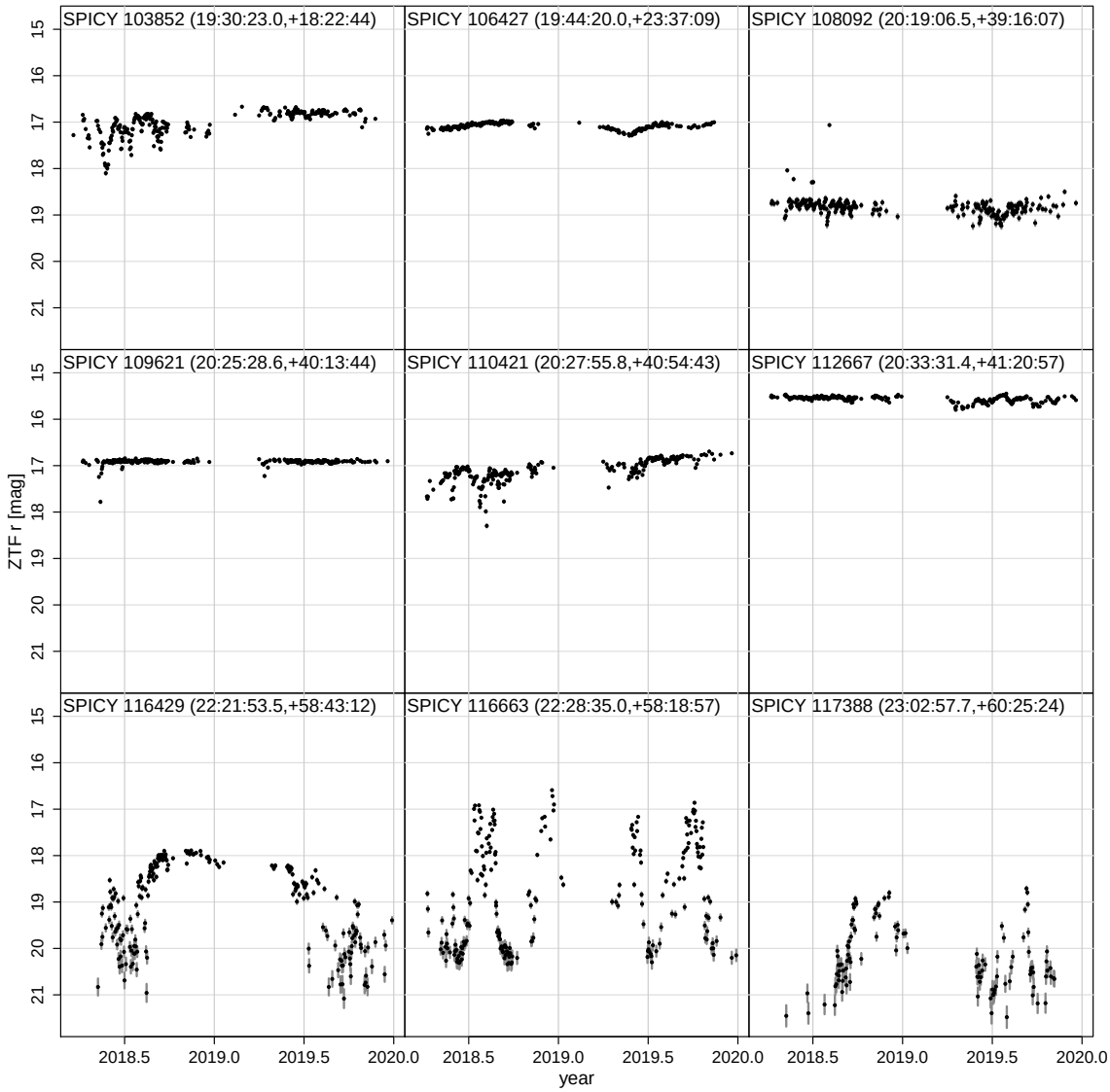


Figure 25. Sample ZTF light curves for several YSO candidates showing strong variability. These light curves exhibit diverse behaviors, including dipping, slow variation in brightness, outbursts, and periodicity.
(The ZTF light curves for 7,585 SPICY sources are provided as data behind the figure.)

As discussed in the introduction, a variety of strategies have been used to identify YSOs from IRAC colors. The SMOG field, which was published by Winston et al. (2019) using a modified version of the Gutermuth et al. (2009) color selection rules, provides an excellent testbed for such a comparison. Comparison between our SMOG candidates (1524 objects) and theirs (4648 objects) shows that our selection methodology is more restrictive; 97% of our YSO candidates were also classified as YSOs by Winston et al. (2019), while only 32% of their YSO candidates were classified as YSOs by us. Figure 26 (right) shows that objects from their list that not included by us tend to be either objects with bluer colors or objects fainter than most of our sample. The difference in magnitude distributions – ours peaking at $[4.5] = 13$ mag and theirs peaking at $[4.5] = 14.5$ mag – may be a limitation related to our training set which was dominated by objects from the shallower GLIMPSE survey areas (Section 2). In spatial distribution (Figure 26,

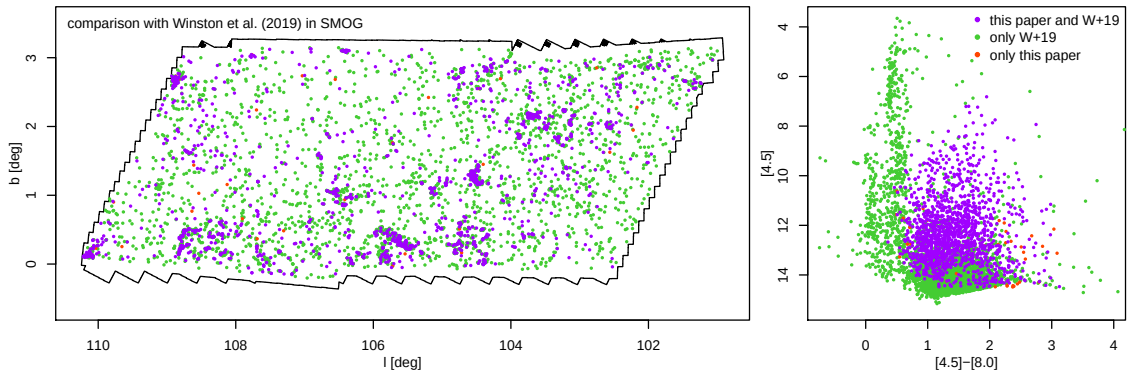


Figure 26. Comparison between our catalog and Winston et al. (2019) in the SMOG field, where Spitzer observations are deeper than the main GLIMPSE survey. YSOs in both catalogs are color-coded purple, objects only in Winston et al. (2019) are green, and a low number of objects only in our catalog are red. The left panel shows the spatial distribution and the right panel shows a color-magnitude diagram. These diagrams indicate that our criteria are more selective than Winston et al. (2019), and the sources we omit tend to either have bluer IR colors or be fainter. Consistency between the catalogs is greatest among the spatially clustered sources, but our catalog does not include many of the non-clustered objects found by Winston et al. (2019). However, formal assessment of the accuracy of either catalog would require additional information from follow-up spectroscopic observations.

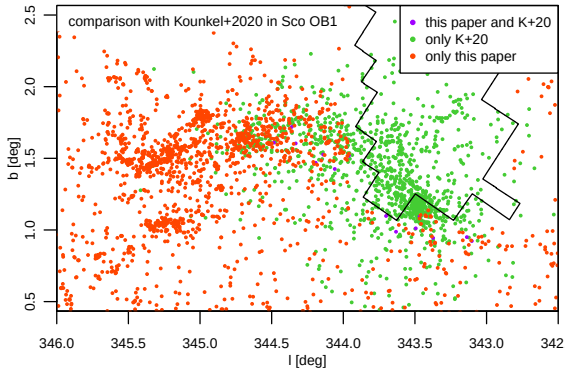


Figure 27. Comparison between our catalog and Kounkel et al. (2020) (age $<$ 10 Myr) in a field centered around the Sco OB1 association. There are few individual stars in common (purple points) between our infrared-excess selected YSOs (red points) and their astrometrically selected sample (green points). Nevertheless, these catalogs appear complementary because they trace different components of the same stellar association. The boundary of the GLIMPSE field (black lines) excludes a section of the sky in the upper right of this figure from our survey.

left) our candidate YSOs tend to be more spatially clustered than those from Winston et al. (2019).

The "Star Formation in the Outer Galaxy" (SFOG; Winston et al. 2020) YSO catalog was recently produced for the GLIMPSE 360 fields, observed during Spitzer's warm mission, meaning that Spitzer's 5.8 and 8.0 μm bands were unavailable. In Galactic coverage, this catalog is largely complementary to ours, but overlaps in the regions of SMOG and part of Cygnus X. The catalogs also overlap in Vela in Galactic longitude, but cover different ranges of Galactic latitude.

WISE covers similar wavelengths as Spitzer, but provides photometry for the whole sky. All-sky searches for YSOs in WISE data include Marton et al. (2016) and Marton et al. (2019), with the latter using cross-matches with Gaia. Below, we compare our catalog to the list of ∼130,000 candidate Class I–II objects from Marton et al. (2016); this paper also lists >600,000 candidate Class III sources, but we do not include these in our comparison because Class III sources are a minority in our catalog but make up the majority of the candidates from Marton et al. (2016). Within the footprint of our catalog, Marton et al. (2016) identify ∼75,000 Class I–II WISE sources, whereas we identify ∼110,000 Class I–II IRAC sources. It is unsurprising that Spitzer can identify more YSOs in the Galactic midplane due to IRAC's higher spatial resolution and WISE's greater susceptibility to detector saturation from bright nebulosity. Using a 2″ match radius, there are only ∼5000 sources in common between our catalog and theirs; visual inspection of the spatial distributions of the unmatched candidates reveal that we include more clustered YSO candidates (often more difficult to observe with WISE), while they include more spatially distributed candidates.

An effort to to identify intermediate-mass young stars (e.g., Herbig Ae/Be stars) via machine learning was made by Vioque et al. (2020), who identify 8,470 candidates using public optical and infrared catalogs. Their list, focused on the higher end of the initial mass function, includes many fewer stars than our catalog; how-

ever, within the spatial overlap area, most of their candidates were re-selected by us.

Another relevant catalog is provided by Kounkel et al. (2020) who identify groups of co-moving stars, including clusters, associations, moving groups, and stellar streams, in Gaia DR2 using a search radius of 3 kpc. Although these systems are not necessarily young, their catalog does include ∼35,000 members of groups with ages <10 Myr, many of which are located near the Galactic midplane. Only ∼300 objects are in common between our catalog and theirs, but this low fraction appears to be related to different selection biases, in particular their stringent Gaia quality cuts, which are only met by 4% of our YSO candidates. The objects in common are mostly assigned to groups with ages from Kounkel et al. (2020) between 4–10 Myr; a handful of objects with older ages may result from either errors in their age estimates or contaminants in our catalog. Visual examination suggests that the catalogs reveal complementary aspects of stellar associations, with Kounkel et al. (2020) mostly selecting diskless members and us the disk/envelope-bearing members. In Figure 27, we show Sco OB1 as an example where the combination of both lists provides a more complete picture of the association.

10. CONCLUSIONS

We present a catalog of 117,446 candidate YSOs (∼90,000 of which are new identifications) in the Galactic midplane from the GLIMPSE survey (Benjamin et al. 2003; Churchwell et al. 2009) and extensions of this survey observed during Spitzer's cryogenic mission. We classify objects obtained from the GLIMPSE I, II, 3D, Vela-Carina (Majewski et al. 2007; Zasowski et al. 2009), Cygnus X (Beerer et al. 2010), and SMOG (Winston et al. 2019) IRAC catalogs, using ancillary data from the near-IR 2MASS (Skrutskie et al. 2006), UKIDSS (Lawrence et al. 2007), and VVV (Minniti et al. 2010) surveys in the most comprehensive search for YSOs in the inner Galactic midplane to date. This catalog is largely restricted to the inner Galaxy and, thus, is complementary to YSO searches in the outer Galaxy (e.g., Winston et al. 2020).

Classification of candidates was entirely based on near-IR and IRAC photometry, and our random forest diagnostics confirm that the IRAC bands were the most important for classification. The IRAC catalogs contain many sources not detected by MIPS or WISE because IRAC, particularly when processed by the GLIMPSE pipeline, is more sensitive in regions of the Galaxy with high crowding and nebulosity. By focusing on IRAC, we are able to identify tens-of-thousands of new YSO candidates that we would not have been able to if we required additional bands. Depending on the science application, future studies that use our YSO list may wish to augment our catalog with YSOs selected using other wavelengths.

The spatial distribution of the candidates, as projected on the sky, is highly structured with cluster-like and filament-like patterns, but also includes a substantial non-clustered population. We have not used spatial information as an input to the classifier because the extent of spatial clustering of YSOs is still an open question and we wish to minimize the influence of selection effects on the observed spatial distributions. From the HDBSCAN algorithm, we identify ∼400 groups of YSOs and estimate their distances and proper motions from the mean astrometry of members detected by Gaia DR2.

The YSOs we identify in the Galactic midplane are mostly at distances $\gtrsim$1 kpc. Some YSO groups appear associated with the Orion, Sagittarius-Carina, and the Scutum-Centaurus arms of the Galaxy, but do not appear to closely trace the estimated arm centers found by other methods (e.g., Reid et al. 2019). Near the boundary between Galactic Quadrants I and IV a large collection of YSO groups are located at the approximate distance of the Sagittarius-Carina Arm. However, these groups are not aligned parallel to the arm but, instead, form a chevron-like shape.

From the portion of our catalog visible to the ZTF survey, our YSO candidates tend to be more variable than field stars in the same region of the Galaxy. Nearly half the stars measured have statistically significant ZTF variability. Visual examination of the sources with the highest variability amplitudes suggests that most of them have light curve morphologies that resemble those expected for YSOs, with large dipping or bursting features. This dataset provides a useful testbed for future work on statistical classification of YSO light curves.

Although the properties of our sample, including optical/infrared photometry, spatial clustering, and variability, are consistent with most of the candidates being YSOs, the level of contamination is difficult to constrain without follow up observation. Although most objects are optically faint, the large total number of YSO candidates means that there are plenty of objects bright enough to follow up with optical spectroscopy. For example, ∼66,000 YSO candidates have $G < 19$ mag, the faint limit for future large spectroscopic surveys such as WEAVE (Dalton et al. 2014) or 4MOST (de Jong et al. 2019), and more than 85% of them are newly proposed in this paper. Furthermore, in the IR, ∼2000 YSO candidates have $H < 11.5$ mag, bright enough for an instrument like APOGEE (Blanton et al. 2017; Cottle et al. 2018), approximately half of which are newly proposed in this paper. Our candidates have been selected with nearly uniform methodology, so they should provide a useful statistical sample for further studies.

APPENDIX

A. GLIMPSE FLAGS

We visually examined a sample of IRAC images from crowded, nebulous regions of the Galaxy to investigate what GLIMPSE flags, including the "close source flag" and the number of detections in each band, imply about the reliability of source detection. The close neighbors for YSO candidates can usually be seen in the 3.6 micron image, but rarely in the 8.0 micron images, possibly because they are fainter at this wavelength, since most neighbors lack IR excess. In such cases, it appears that the GLIMPSE pipeline has correctly identified both sources. Nearly all sources with two or more detections in a band look like a *bona fide* point sources in that band's images; however, some of the sources detected only once look like point sources while others do not. Thus, having fewer than 2 detections is an indicator that a source could be less reliable. Nevertheless, we do not filter out sources based on either the "close source flag" or the number of detections in each band because any such cuts would remove numerous objects that appear to be good YSOs. These GLIMPSE flags are included in Table 1 for any users of our YSO catalog who would like to make alternate choices for their own scientific applications.

B. RED NON-YSOS IN THE IRAC CATALOGS

B.1. *Evolved Stars*

Certain evolved stars, including dusty red giants, AGB stars, post-AGB stars, and red super giants (RSGs), have red IR SEDs due to dusty stellar winds (e.g., Marengo et al. 1997, 1999; Groenewegen 2012; Chun et al. 2015; Suh 2020), making such objects a significant category of contaminant in infrared YSO catalogs (e.g., Robitaille et al. 2008; Povich et al. 2013). Reiter et al. (2015) present a sample of AGB stars (including O-rich, S-rich, and C-rich stars) and RSG stars with JHK and IRAC photometry. In the near-IR, the $J-K$ colors of this sample ($J-K \gtrsim 0.9$) are consistent with the group of probable contaminates on the J vs. $J-K_s$ diagram that are brighter and redder than the typical YSOs (Figure 9, left panel). In IRAC color space, the distribution of the AGB+RGS sources partially overlap the distribution of YSO candidates. However, most of them have IRAC colors (e.g., $[3.6]-[4.5] \lesssim 0.5$, $[4.5]-[8.0] \lesssim 1$) bluer than the typical colors of YSOs, but similar in color to the bright sources classified as probable contaminants (e.g., Figure 9, right panel).

B.2. *Extragalactic Sources*

Extragalactic sources, including active and star-forming galaxies, can have mid-IR colors that mimic the IR excesses of YSOs (e.g., Stern et al. 2005; Jarrett et al. 2011). These sources can contribute a significant number of possible contaminants in some YSO searches (e.g., Harvey et al. 2007; Gutermuth et al. 2008). However, in the GLIMPSE survey, extragalactic contamination is expected to be lower, owing to shallower IRAC observations and high extinction near the Galactic midplane (e.g., Kang et al. 2009). Jarrett et al. (2011) provide a sample of IRAC sources in fields dominated by extragalactic sources. The galaxies in these fields tend to have $[3.6]-[4.5] \approx 0$–1.25 and $[4.5]-[8.0] \approx 1$–4. This distribution more closely resembles the distribution of probable contaminants in our sample than it does our YSO candidates (Figure 11, lower right). Their extragalactic sources mostly have $[3.6] \gtrsim 14$ mag, which is fainter than most of our candidate YSOs.

B.3. *Labeled Non-YSOs in the Training Set*

For the random forest classifier, the sample of non-YSOs in our training set is equal in importance to the sample of YSOs. As with the labeled YSOs, the labeled field objects are obtained from the set of NIR+IRAC sources that could not be fit by a reddened stellar photosphere (Section 3.1). Thus, they also represent objects with red IR colors.

We generate our list of "non-YSOs" using both sources within the boundaries of the MYStIX star-forming regions that were classified as non-YSOs and field stars in regions of the Galaxy where there is no star-formation activity. For the first category, we include all IRAC sources that lie within the MYStIX fields that show no indication of youth – requiring both rejection as a YSO based on its SED and non detection of X-ray emission.

We also include stars from rectangular regions of the sky in areas where there is no evidence for star-formation or the presence of YSOs. These regions, listed in Table 3, have been chosen to sample stars along different Galactic latitudes and longitudes and along lines of sight with different levels of extinction. Given the high numbers of stars within these fields, a random subsample of these stars are included in the training data for the classifier. Altogether, there are 14,019 labeled field objects for the 2MASS+IRAC sample, 5,320 for UKIDSS+IRAC, and 4,047 for VIRAC+IRAC.

C. TRAINING SETS AND IMPUTED COLORS

Figure 28 shows color-color/magnitude diagrams for the training set, containing both labeled YSOs (reddish points) and non-YSOs (bluish points) and observed (dark points) and imputed (light points) colors. Here, we show only the 2MASS+IRAC training data, but the general morphologies of the distributions are similar for target dataset, as well as for the training/target UKIDSS+IRAC and VIRAC+IRAC datasets. For each

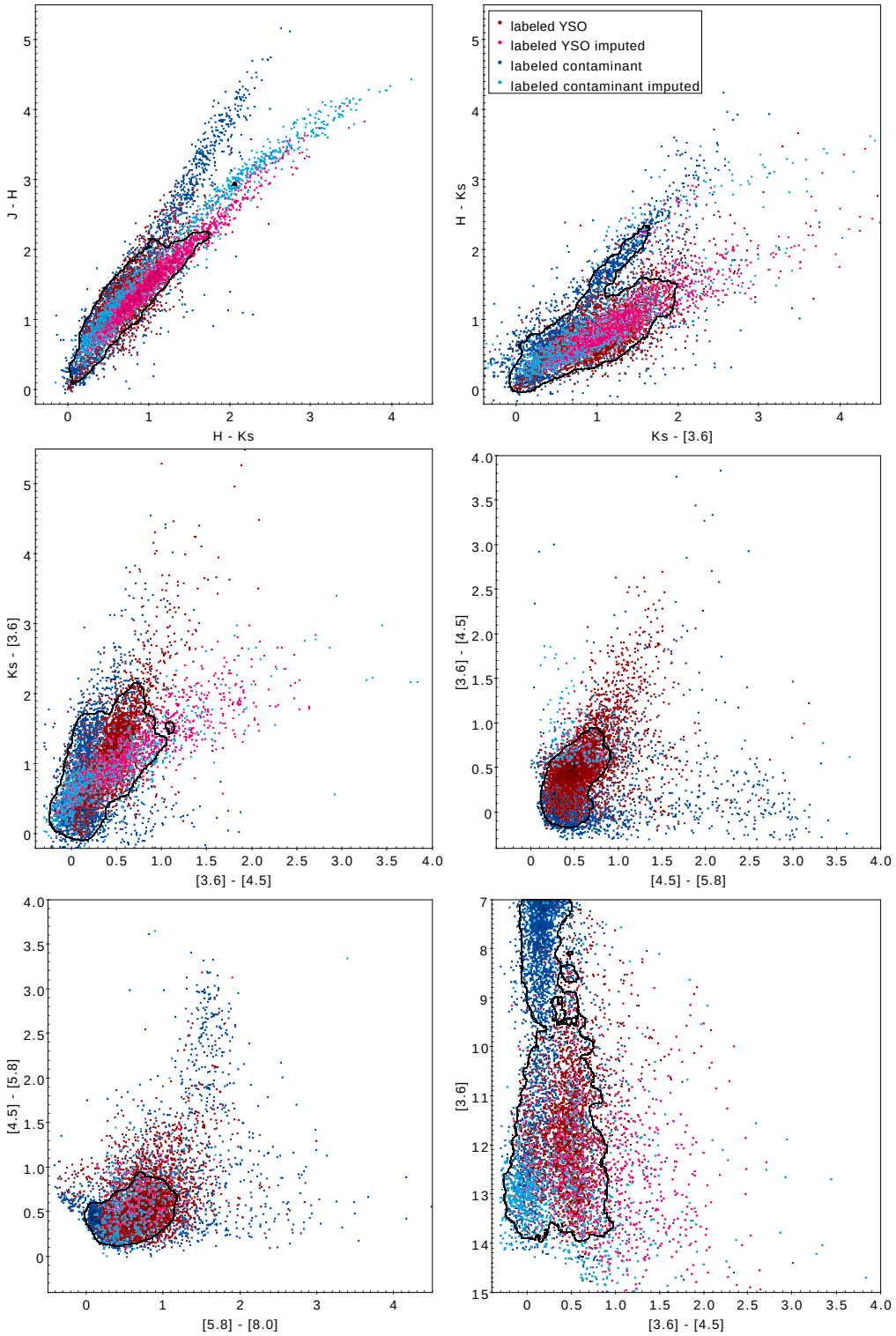


Figure 28. Scatter plots of labeled training data after copula imputation. Objects labeled "YSO" are indicated by reddish points and objects labeled "field" are indicated by bluish points. On each panel, the darker points have measurements of both colors, while points where one or both colors are imputed are marked with a lighter color as indicated by the legend.

NIR+IRAC combination, an identical copula was used for every data point regardless of label (YSO or non-YSO) or whether the data point belongs to the training or target set. Thus, any differences that emerge in the distributions of imputed data must emerge from the data itself.

In general the distribution of the imputed data lies within the distributions traced out by the observed data, with the $J-H$ vs. $H-K_s$ diagram being the main exception. In the JHK_s diagram, many sources are missing the $J-H$ color (presumably due to high extinction), and follow a locus with a slightly flatter slope than the objects for which both $J-H$ and $H-K_s$ have been measured. This behavior also appears when using UKIDSS or VIRAC JHK photometry, and it may suggest that there is an intrinsic difference in distributions for sources with and without measured $J-H$. The sources without measured $J-H$ also tend to have higher than average $H-K_s$ values, possibly influencing how the distribution of the imputed sources. Nevertheless, the analysis of the random forest classifier suggests that $J-H$ color is one of the less important features for producing a classification.

None of the other diagrams reveal such large deviations between observed and imputed data, but differences are visible in the distributions of YSOs and non-YSOs. On some diagrams there are regions with no imputed data (e.g., the lower left of the $[3.6]-[4.5]$ vs. $[4.5]-[5.8]$ diagram) because any missing data would have meant that the source would not be under consideration (see Section 3.1).

In the target set, there is a low number of outliers in regions of color space that are not well populated by sources from the training set, either by sources labeled "YSO" or "non-YSO," meaning that our classifier would not be able to generate reliable classifications for these objects. These are either rare objects that arise due to the large size ($\sim 5 \times 10^7$ sources) of the full IRAC photometric catalogs or are sources unlikely to have infrared excess but that we failed to remove with our procedures in Section 3.1. Given that we have no basis for classifying these objects with a RF, we are cautious and exclude them from our lists of candidate YSOs. We use the following criteria to define these outliers: $[3.6]-[4.5] < -0.3$, $[4.5]-[5.8] < 0$, $[4.5]-[8.0] < 0$, $[3.6]-[5.8] < 0$, $[3.6]-[5.8] > 6$, $[4.5]-[5.8] < 0$, or $[3.6]-[8.0] < 0$.

ACKNOWLEDGMENTS

We thank Robert Benjamin for useful discussions about GLIMPSE and spiral structure of the Galaxy, and Philip Lucas and Leigh Smith for assistance with the UKIDSS and VIRAC catalogs. This work is a result of the 6$^{\text{th}}$ COIN Residence Program (CRP#6; https://cosmostatistics-initiative.org/residence-programs/crp6) held in Chamonix, France in August 2019. COIN is financially supported by CNRS as part of its MOMENTUM programme over the 2018–2020 period. This work is based on observations made with the Spitzer Space Telescope, which is operated by the Jet Propulsion Laboratory, California Institute of Technology under a contract with NASA. This work has also made use of data from the European Space Agency mission Gaia, processed by the Gaia Data Processing and Analysis Consortium. Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the Gaia Multilateral Agreement. This work is also based in part on observations obtained with the Samuel Oschin 48-inch Telescope at the Palomar Observatory as part of the Zwicky Transient Facility project. ZTF is supported by the National Science Foundation under Grant No. AST-1440341 and a collaboration including Caltech, IPAC, the Weizmann Institute for Science, the Oskar Klein Center at Stockholm University, the University of Maryland, the University of Washington, Deutsches Elektronen-Synchrotron and Humboldt University, Los Alamos National Laboratories, the TANGO Consortium of Taiwan, the University of Wisconsin at Milwaukee, and Lawrence Berkeley National Laboratories. This research has made use of the NASA/IPAC Infrared Science Archive, which is funded by the National Aeronautics and Space Administration and operated by the California Institute of Technology. AKM acknowledges the support from the Portuguese Fundação para a Ciência e a Tecnologia (FCT) through grants SFRH/BPD/74697/2010, PTDC/FIS-AST/31546/2017 and from the Portuguese Strategic Programme UID/FIS/00099/2013 for CENTRA.

Facility: 2MASS, Gaia, Spitzer (IRAC, MIPS), UKIRT, VISTA/VIRCAM, WISE, ZTF, IRSA

Software: caret (Kuhn 2015), mclust (Scrucca et al. 2016), hdbscan (McInnes et al. 2017), modeest (Poncet 2019), mclust (Scrucca et al. 2016), PostgreSQL (PostgreSQL Global Development Group 2020), Python, R (R Core Team 2019), rjags (Plummer 2017, 2019), SAOImage DS9 (Joye & Mandel 2003), sbgcop (Hoff 2018), TOPCAT & STILTS (Taylor 2005)

REFERENCES

Allen, L., Megeath, S. T., Gutermuth, R., et al. 2007, in Protostars and Planets V, ed. B. Reipurth, D. Jewitt, & K. Keil (Tucson: University of Arizona Press), 361. https://arxiv.org/abs/astro-ph/0603096

Allen, L. E., Calvet, N., D'Alessio, P., et al. 2004, ApJS, 154, 363, doi: 10.1086/422715

Alves, J., Zucker, C., Goodman, A. A., et al. 2020, Nature, 578, 237, doi: 10.1038/s41586-019-1874-z

Anderson, L. D., Bania, T. M., Balser, D. S., et al. 2014, ApJS, 212, 1, doi: 10.1088/0067-0049/212/1/1

Andre, P., & Montmerle, T. 1994, ApJ, 420, 837, doi: 10.1086/173608

Andreani, P., Boselli, A., Ciesla, L., et al. 2018, A&A, 617, A33, doi: 10.1051/0004-6361/201832873

Ascenso, J. 2018, in The Birth of Star Clusters, ed. S. Stahler (Cham: Springer International Publishing), 1–37, doi: 10.1007/978-3-319-22801-3_1

Baddeley, A. 2017, Spatial Statistics, 22, 261 , doi: https://doi.org/10.1016/j.spasta.2017.03.001

Beerer, I. M., Koenig, X. P., Hora, J. L., et al. 2010, ApJ, 720, 679, doi: 10.1088/0004-637X/720/1/679

Bellm, E. C., Kulkarni, S. R., Graham, M. J., et al. 2019, PASP, 131, 018002, doi: 10.1088/1538-3873/aaecbe

Benjamin, R. A., Churchwell, E., Babler, B. L., et al. 2003, PASP, 115, 953, doi: 10.1086/376696

Blanton, M. R., Bershady, M. A., Abolfathi, B., et al. 2017, AJ, 154, 28, doi: 10.3847/1538-3881/aa7567

Bovy, J. 2017, MNRAS, 468, L63, doi: 10.1093/mnrasl/slx027

Breiman, L. 2001, Machine Learning, 45, 5, doi: 10.1023/A:1010933404324

Bressan, A., Marigo, P., Girardi, L., et al. 2012, MNRAS, 427, 127, doi: 10.1111/j.1365-2966.2012.21948.x

Bressert, E., Bastian, N., Gutermuth, R., et al. 2010, MNRAS, 409, L54, doi: 10.1111/j.1745-3933.2010.00946.x

Buckner, A. S. M., Khorrami, Z., González, M., et al. 2020, A&A, 636, A80, doi: 10.1051/0004-6361/201936935

Bufano, F., Leto, P., Carey, D., et al. 2018, MNRAS, 473, 3671, doi: 10.1093/mnras/stx2560

Campello, R. J., Moulavi, D., & Sander, J. 2013, in Pacific-Asia conference on knowledge discovery and data mining, Springer, 160–172

Candès, E. J., & Donoho, D. L. 2000, Curvelets – A Surprisingly Effective Nonadaptive Representation For Objects with Edges (TN, Nashville: Vanderbilt Univ. Press), 1–10

Cardelli, J. A., Clayton, G. C., & Mathis, J. S. 1989, ApJ, 345, 245, doi: 10.1086/167900

Carey, S. J., Noriega-Crespo, A., Mizuno, D. R., et al. 2009, PASP, 121, 76, doi: 10.1086/596581

Carpenter, J. M. 2000a, AJ, 120, 3139, doi: 10.1086/316845

—. 2000b, AJ, 120, 3139, doi: 10.1086/316845

Castelli, F., & Kurucz, R. L. 2003, in IAU Symposium, Vol. 210, Modelling of Stellar Atmospheres, ed. N. Piskunov, W. W. Weiss, & D. F. Gray, A20. https://arxiv.org/abs/astro-ph/0405087

Castro-Ginard, A., Jordi, C., Luri, X., et al. 2020, A&A, 635, A45, doi: 10.1051/0004-6361/201937386

Chiar, J. E., Ennico, K., Pendleton, Y. J., et al. 2007, ApJL, 666, L73, doi: 10.1086/521789

Chun, S.-H., Jung, M., Kang, M., Kim, J.-W., & Sohn, Y.-J. 2015, A&A, 578, A51, doi: 10.1051/0004-6361/201525849

Churchwell, E., Povich, M. S., Allen, D., et al. 2006, ApJ, 649, 759, doi: 10.1086/507015

Churchwell, E., Watson, D. F., Povich, M. S., et al. 2007, ApJ, 670, 428, doi: 10.1086/521646

Churchwell, E., Babler, B. L., Meade, M. R., et al. 2009, PASP, 121, 213, doi: 10.1086/597811

Cody, A. M., & Hillenbrand, L. A. 2018, AJ, 156, 71, doi: 10.3847/1538-3881/aacead

Contreras Peña, C., Lucas, P. W., Minniti, D., et al. 2017, MNRAS, 465, 3011, doi: 10.1093/mnras/stw2801

Cottle, J. N., Covey, K. R., Suárez, G., et al. 2018, ApJS, 236, 27, doi: 10.3847/1538-4365/aabada

Dalton, G., Trager, S., Abrams, D. C., et al. 2014, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9147, Ground-based and Airborne Instrumentation for Astronomy V, ed. S. K. Ramsay, I. S. McLean, & H. Takami, 91470L, doi: 10.1117/12.2055132

de Jong, R. S., Agertz, O., Berbel, A. A., et al. 2019, The Messenger, 175, 3, doi: 10.18727/0722-6691/5117

de Souza, R. S., Maio, U., Biffi, V., & Ciardi, B. 2014, MNRAS, 440, 240, doi: 10.1093/mnras/stu274

de Souza, R. S., Dantas, M. L. L., Costa-Duarte, M. V., et al. 2017, MNRAS, 472, 2808, doi: 10.1093/mnras/stx2156

Dewangan, L. K., & Ojha, D. K. 2013, MNRAS, 429, 1386, doi: 10.1093/mnras/sts430

Draine, B. T., & Li, A. 2007, ApJ, 657, 810, doi: 10.1086/511055

Ducourant, C., Teixeira, R., Krone-Martins, A., et al. 2017, A&A, 597, A90, doi: 10.1051/0004-6361/201527574

Elmegreen, B. G., & Scalo, J. 2004, ARA&A, 42, 211, doi: 10.1146/annurev.astro.41.011802.094859

Evans, D. W., Riello, M., De Angeli, F., et al. 2018, A&A, 616, A4, doi: 10.1051/0004-6361/201832756

Evans, Neal J., I., Dunham, M. M., Jørgensen, J. K., et al. 2009a, ApJS, 181, 321, doi: 10.1088/0067-0049/181/2/321

Evans, N., Calvet, N., Cieza, L., et al. 2009b, arXiv e-prints, arXiv:0901.1691. https://arxiv.org/abs/0901.1691

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. 2001, Cluster Analysis, A Hodder Arnold Publication (Wiley). https://books.google.com.br/books?id=htZzDGlCnQYC

Fang, M., Hillenbrand, L. A., Kim, J. S., et al. 2020, arXiv e-prints, arXiv:2009.11995. https://arxiv.org/abs/2009.11995

Fazio, G. G., Hora, J. L., Allen, L. E., et al. 2004, ApJS, 154, 10, doi: 10.1086/422843

Feigelson, E. D. 2018, in The Birth of Star Clusters, ed. S. Stahler (Cham: Springer International Publishing), 119–141, doi: 10.1007/978-3-319-22801-3_5

Feigelson, E. D., Townsley, L. K., Broos, P. S., et al. 2013, ApJS, 209, 26, doi: 10.1088/0067-0049/209/2/26

Forbrich, J., Tappe, A., Robitaille, T., et al. 2010, ApJ, 716, 1453, doi: 10.1088/0004-637X/716/2/1453

Furlan, E., Hartmann, L., Calvet, N., et al. 2006, ApJS, 165, 568, doi: 10.1086/505468

Furlan, E., McClure, M., Calvet, N., et al. 2008, ApJS, 176, 184, doi: 10.1086/527301

Furlan, E., Luhman, K. L., Espaillat, C., et al. 2011, ApJS, 195, 3, doi: 10.1088/0067-0049/195/1/3

Gaia Collaboration, Prusti, T., de Bruijne, J. H. J., et al. 2016, A&A, 595, A1, doi: 10.1051/0004-6361/201629272

Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2018, A&A, 616, A1, doi: 10.1051/0004-6361/201833051

Gieles, M., Moeckel, N., & Clarke, C. J. 2012, MNRAS, 426, L11, doi: 10.1111/j.1745-3933.2012.01312.x

Gouliermis, D. A. 2018, PASP, 130, 072001, doi: 10.1088/1538-3873/aac1fd

Graham, M. J., Kulkarni, S. R., Bellm, E. C., et al. 2019, PASP, 131, 078001, doi: 10.1088/1538-3873/ab006c

Greene, T. P., Wilking, B. A., Andre, P., Young, E. T., & Lada, C. J. 1994, ApJ, 434, 614, doi: 10.1086/174763

Greenwell, B. M. 2017, The R Journal, 9, 421, doi: 10.32614/RJ-2017-016

Groenewegen, M. A. T. 2012, A&A, 540, A32, doi: 10.1051/0004-6361/201118287

Gutermuth, R. A., & Heyer, M. 2015, AJ, 149, 64, doi: 10.1088/0004-6256/149/2/64

Gutermuth, R. A., Megeath, S. T., Myers, P. C., et al. 2009, ApJS, 184, 18, doi: 10.1088/0067-0049/184/1/18

Gutermuth, R. A., Myers, P. C., Megeath, S. T., et al. 2008, ApJ, 674, 336, doi: 10.1086/524722

Hartmann, L., Megeath, S. T., Allen, L., et al. 2005, ApJ, 629, 881, doi: 10.1086/431472

Harvey, P., Merín, B., Huard, T. L., et al. 2007, ApJ, 663, 1149, doi: 10.1086/518646

Herbig, G. H. 1954, ApJ, 119, 483, doi: 10.1086/145854

Herbst, W., Herbst, D. K., Grossman, E. J., & Weinstein, D. 1994, AJ, 108, 1906, doi: 10.1086/117204

Herczeg, G. J., Kuhn, M. A., Zhou, X., et al. 2019, ApJ, 878, 111, doi: 10.3847/1538-4357/ab1d67

Hilbe, J. M., de Souza, R. S., & Ishida, E. E. O. 2017, Bayesian Models for Astrophysical Data Using R, JAGS, Python, and Stan (Cambridge University Press), doi: 10.1017/CBO9781316459515

Ho, T. K. 1995, in Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1, ICDAR '95 (USA: IEEE Computer Society), 278

Hodgkin, S. T., Irwin, M. J., Hewett, P. C., & Warren, S. J. 2009, MNRAS, 394, 675, doi: 10.1111/j.1365-2966.2008.14387.x

Hodgkin, S. T., Wyrzykowski, L., Blagorodnova, N., & Koposov, S. 2013, Philosophical Transactions of the Royal Society of London Series A, 371, 20120239, doi: 10.1098/rsta.2012.0239

Hoff, P. 2018, sbgcop: Semiparametric Bayesian Gaussian Copula Estimation and Imputation. https://CRAN.R-project.org/package=sbgcop

Hoff, P. D. 2007, Ann. Appl. Stat., 1, 265, doi: 10.1214/07-AOAS107

Honaker, J., King, G., & Blackwell, M. 2011, Journal of Statistical Software, 45, 1

Hotelling, H. 1933, Journal of Educational Psychology, 24, 417

Indebetouw, R., Mathis, J. S., Babler, B. L., et al. 2005, ApJ, 619, 931, doi: 10.1086/426679

Ishida, E. E. O., & de Souza, R. S. 2013, MNRAS, 430, 509, doi: 10.1093/mnras/sts650

Jaeger, B. C., Tierney, N. J., & Simon, N. R. 2020, When to Impute? Imputation before and during cross-validation. https://arxiv.org/abs/2010.00718

Jarrett, T. H., Cohen, M., Masci, F., et al. 2011, ApJ, 735, 112, doi: 10.1088/0004-637X/735/2/112

Jayasinghe, T., Dixon, D., Povich, M. S., et al. 2019, MNRAS, 488, 1141, doi: 10.1093/mnras/stz1738

Joy, A. H. 1945, ApJ, 102, 168, doi: 10.1086/144749

Joye, W. A., & Mandel, E. 2003, in Astronomical Society of the Pacific Conference Series, Vol. 295, Astronomical Data Analysis Software and Systems XII, ed. H. E. Payne, R. I. Jedrzejewski, & R. N. Hook, 489

Kang, M., Bieging, J. H., Povich, M. S., & Lee, Y. 2009, ApJ, 706, 83, doi: 10.1088/0004-637X/706/1/83

Kauppinen, H., Seppanen, T., & Pietikainen, M. 1995, IEEE Transactions on Pattern Analysis and Machine Intelligence, 17, 201, doi: 10.1109/34.368168

Kobulnicky, H. A., Babler, B. L., Alexand er, M. J., et al. 2013, ApJS, 207, 9, doi: 10.1088/0067-0049/207/1/9

Koenig, X. P., & Leisawitz, D. T. 2014, ApJ, 791, 131, doi: 10.1088/0004-637X/791/2/131

Kounkel, M., & Covey, K. 2019, AJ, 158, 122, doi: 10.3847/1538-3881/ab339a

Kounkel, M., Covey, K., & Stassun, K. G. 2020, arXiv e-prints, arXiv:2004.07261. https://arxiv.org/abs/2004.07261

Krone-Martins, A., Graham, M. J., Stern, D., et al. 2019, arXiv:1912.08977

Kuhn, M. 2015, caret: Classification and Regression Training. http://ascl.net/1505.003

Kuhn, M. A., & Feigelson, E. D. 2017, in Handbook of Mixture Analysis, ed. S. Fruhwirth-Schnatter, G. Celeux, & C. Robert (New York: Chapman and Hall/CRC), 463–489, doi: 10.1201/9780429055911

Kuhn, M. A., Getman, K. V., & Feigelson, E. D. 2015, ApJ, 802, 60, doi: 10.1088/0004-637X/802/1/60

Kuhn, M. A., Hillenbrand, L. A., Carpenter, J. M., & Menendez, A. R. A. 2020, ApJ, 899, 128, doi: 10.3847/1538-4357/aba19a

Lada, C. J. 1987, in IAU Symposium, Vol. 115, Star Forming Regions, ed. M. Peimbert & J. Jugaku, 1

Lawrence, A., Warren, S. J., Almaini, O., et al. 2007, MNRAS, 379, 1599, doi: 10.1111/j.1365-2966.2007.12040.x

Leung, H. W., & Bovy, J. 2019, MNRAS, 489, 2079, doi: 10.1093/mnras/stz2245

Lin, C.-A., Kilbinger, M., & Pires, S. 2016, A&A, 593, A88, doi: 10.1051/0004-6361/201628565

Lindegren, L., Hernández, J., Bombrun, A., et al. 2018, A&A, 616, A2, doi: 10.1051/0004-6361/201832727

Loh, J. M. 2008, The Astrophysical Journal, 681, 726, doi: 10.1086/588631

Lucas, P. W., Hoare, M. G., Longmore, A., et al. 2008, MNRAS, 391, 136, doi: 10.1111/j.1365-2966.2008.13924.x

Luhman, K. L. 2018, AJ, 156, 271, doi: 10.3847/1538-3881/aae831

Lumsden, S. L., Hoare, M. G., Urquhart, J. S., et al. 2013, ApJS, 208, 11, doi: 10.1088/0067-0049/208/1/11

Lupton, R., Blanton, M. R., Fekete, G., et al. 2004, PASP, 116, 133, doi: 10.1086/382245

Majewski, S., Babler, B., Churchwell, E., et al. 2007, Galactic Structure and Star Formation in Vela-Carina, Spitzer Proposal

Mallick, K. K., Ojha, D. K., Tamura, M., et al. 2015, MNRAS, 447, 2307, doi: 10.1093/mnras/stu2584

Marengo, M., Busso, M., Silvestro, G., Persi, P., & Lagage, P. O. 1999, A&A, 348, 501

Marengo, M., Canil, G., Silvestro, G., et al. 1997, A&A, 322, 924. https://arxiv.org/abs/astro-ph/9607129

Marigo, P., Bressan, A., Nanni, A., Girardi, L., & Pumo, M. L. 2013, MNRAS, 434, 488, doi: 10.1093/mnras/stt1034

Marton, G., Tóth, L. V., Paladini, R., et al. 2016, MNRAS, 458, 3479, doi: 10.1093/mnras/stw398

Marton, G., Ábrahám, P., Szegedi-Elek, E., et al. 2019, MNRAS, 487, 2522, doi: 10.1093/mnras/stz1301

Masci, F. J., Laher, R. R., Rusholme, B., et al. 2019, PASP, 131, 018003, doi: 10.1088/1538-3873/aae8ac

McClure, M. 2009, ApJL, 693, L81, doi: 10.1088/0004-637X/693/2/L81

McInnes, L., Healy, J., & Astels, S. 2017, The Journal of Open Source Software, 2, 205, doi: 10.21105/joss.00205

McKee, C. F., & Ostriker, E. C. 2007, Annual Review of Astronomy and Astrophysics, 45, 565, doi: 10.1146/annurev.astro.45.051806.110602

Melchior, P., & Goulding, A. D. 2018, Astronomy and Computing, 25, 183, doi: 10.1016/j.ascom.2018.09.013

Melton, E. 2020, AJ, 159, 200, doi: 10.3847/1538-3881/ab72ac

Minniti, D., Lucas, P. W., Emerson, J. P., et al. 2010, NewA, 15, 433, doi: 10.1016/j.newast.2009.12.002

Morales, E. F. E., & Robitaille, T. P. 2017, A&A, 598, A136, doi: 10.1051/0004-6361/201628450

Nelsen, R. B. 2010, An Introduction to Copulas (Springer Publishing Company, Incorporated)

O'Donnell, J. E. 1994, ApJ, 422, 158, doi: 10.1086/173713

Oliveira, I., Pontoppidan, K. M., Merín, B., et al. 2010, ApJ, 714, 778, doi: 10.1088/0004-637X/714/1/778

Oliveira, J. M., van Loon, J. T., Sloan, G. C., et al. 2013, MNRAS, 428, 3001, doi: 10.1093/mnras/sts250

Oort, J. H. 1927, BAN, 3, 275

Pari, J., & Hora, J. L. 2020, PASP, 132, 054301, doi: 10.1088/1538-3873/ab7b39

Pearson, K. 1894, Philosophical Transactions of the Royal Society of London Series A, 185, 71, doi: 10.1098/rsta.1894.0003

Pearson, K. 1901, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2, 559, doi: 10.1080/14786440109462720

Pfalzner, S., Kaczmarek, T., & Olczak, C. 2012, A&A, 545, A122, doi: 10.1051/0004-6361/201219881

Plummer, M. 2017, Retrieved from sourceforge.net/projects/mcmc-jags/files/Manuals/4.x, 2

—. 2019, rjags: Bayesian Graphical Models using MCMC. https://CRAN.R-project.org/package=rjags

Poncet, P. 2019, modeest: Mode Estimation. https://CRAN.R-project.org/package=modeest

PostgreSQL Global Development Group. 2020, PostgreSQL 12.4. https://www.postgresql.org/

Povich, M. S., Townsley, L. K., Robitaille, T. P., et al. 2016, ApJ, 825, 125, doi: 10.3847/0004-637X/825/2/125

Povich, M. S., Churchwell, E., Bieging, J. H., et al. 2009, ApJ, 696, 1278, doi: 10.1088/0004-637X/696/2/1278

Povich, M. S., Smith, N., Majewski, S. R., et al. 2011, ApJS, 194, 14, doi: 10.1088/0067-0049/194/1/14

Povich, M. S., Kuhn, M. A., Getman, K. V., et al. 2013, The Astrophysical Journal Supplement Series, 209, 31, doi: 10.1088/0067-0049/209/2/31

R Core Team. 2019, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rebull, L. M., Guieu, S., Stauffer, J. R., et al. 2011, ApJS, 193, 25, doi: 10.1088/0067-0049/193/2/25

Rebull, L. M., Cody, A. M., Covey, K. R., et al. 2014, AJ, 148, 92, doi: 10.1088/0004-6256/148/5/92

Reid, M. J., Menten, K. M., Brunthaler, A., et al. 2019, ApJ, 885, 131, doi: 10.3847/1538-4357/ab4a11

Reiter, M., Marengo, M., Hora, J. L., & Fazio, G. G. 2015, MNRAS, 447, 3909, doi: 10.1093/mnras/stu2725

Rieke, G. H., & Lebofsky, M. J. 1985, ApJ, 288, 618, doi: 10.1086/162827

Ripley, B. D. 1976, Journal of Applied Probability, 13, 255–266, doi: 10.2307/3212829

Robitaille, T. P. 2017, Astronomy and Astrophysics, 600, A11, doi: 10.1051/0004-6361/201425486

Robitaille, T. P., Whitney, B. A., Indebetouw, R., & Wood, K. 2007, ApJS, 169, 328, doi: 10.1086/512039

Robitaille, T. P., Whitney, B. A., Indebetouw, R., Wood, K., & Denzmore, P. 2006, ApJS, 167, 256, doi: 10.1086/508424

Robitaille, T. P., Meade, M. R., Babler, B. L., et al. 2008, The Astronomical Journal, 136, 2413, doi: 10.1088/0004-6256/136/6/2413

Roche, P. F., & Aitken, D. K. 1984, MNRAS, 208, 481, doi: 10.1093/mnras/208.3.481

Rokach, L., & Maimon, O. 2014, Data Mining With Decision Trees: Theory and Applications, 2nd edn. (USA: World Scientific Publishing Co., Inc.)

Rosaria, Bonito, Hartigan, P., et al. 2018, arXiv e-prints, arXiv:1812.03135. https://arxiv.org/abs/1812.03135

Sagi, O., & Rokach, L. 2018, WIREs Data Mining and Knowledge Discovery, 8, e1249, doi: 10.1002/widm.1249

Samal, M. R., Pandey, A. K., Ojha, D. K., et al. 2010, ApJ, 714, 1015, doi: 10.1088/0004-637X/714/2/1015

Samal, M. R., Zavagno, A., Deharveng, L., et al. 2014, A&A, 566, A122, doi: 10.1051/0004-6361/201321794

Sasdelli, M., Ishida, E. E. O., Vilalta, R., et al. 2016, MNRAS, 461, 2044, doi: 10.1093/mnras/stw1228

Sato, M., Ichiki, K., & Takeuchi, T. T. 2011, PhRvD, 83, 023501, doi: 10.1103/PhysRevD.83.023501

Schwarz, G. 1978, Ann. Statist., 6, 461, doi: 10.1214/aos/1176344136

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. 2016, The R Journal, 8, 289. https://doi.org/10.32614/RJ-2016-021

Shao, Z., Jiang, B. W., Li, A., et al. 2018, MNRAS, 478, 3467, doi: 10.1093/mnras/sty1267

Shu, F. H., Adams, F. C., & Lizano, S. 1987, ARA&A, 25, 23, doi: 10.1146/annurev.aa.25.090187.000323

Simon, J. D., Bolatto, A. D., Whitney, B. A., et al. 2007, ApJ, 669, 327, doi: 10.1086/521544

Simpson, R. J., Povich, M. S., Kendrew, S., et al. 2012, MNRAS, 424, 2442, doi: 10.1111/j.1365-2966.2012.20770.x

Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, AJ, 131, 1163, doi: 10.1086/498708

Smith, L. C., Lucas, P. W., Kurtev, R., et al. 2018, MNRAS, 474, 1826, doi: 10.1093/mnras/stx2789

Soto, M., Barbá, R., Gunthardt, G., et al. 2013, A&A, 552, A101, doi: 10.1051/0004-6361/201220046

Starck, J. L., Donoho, D. L., & Candès, E. J. 2003, A&A, 398, 785, doi: 10.1051/0004-6361:20021571

Stauffer, J., Collier Cameron, A., Jardine, M., et al. 2017, AJ, 153, 152, doi: 10.3847/1538-3881/aa5eb9

Stern, D., Eisenhardt, P., Gorjian, V., et al. 2005, ApJ, 631, 163, doi: 10.1086/432523

Stolovy, S., Ramirez, S., Arendt, R. G., et al. 2006, in Journal of Physics Conference Series, Vol. 54, Journal of Physics Conference Series, 176–182, doi: 10.1088/1742-6596/54/1/030

Suh, K.-W. 2020, ApJ, 891, 43, doi: 10.3847/1538-4357/ab6609

Sung, H., Stauffer, J. R., & Bessell, M. S. 2009, AJ, 138, 1116, doi: 10.1088/0004-6256/138/4/1116

Taylor, M. B. 2005, in Astronomical Society of the Pacific Conference Series, Vol. 347, Astronomical Data Analysis Software and Systems XIV, ed. P. Shopbell, M. Britton, & R. Ebert, 29

Townsley, L. K., Broos, P. S., Corcoran, M. F., et al. 2011, ApJS, 194, 1, doi: 10.1088/0067-0049/194/1/1

van Breemen, J. M., Min, M., Chiar, J. E., et al. 2011, A&A, 526, A152, doi: 10.1051/0004-6361/200811142

van Buuren, S., & Groothuis-Oudshoorn, K. 2011, Journal of Statistical Software, Articles, 45, 1, doi: 10.18637/jss.v045.i03

van den Bergh, S. 1964, ApJS, 9, 65, doi: 10.1086/190097

van der Schaaf, A., & van Hateren, J. 1996, Vision Research, 36, 2759 , doi: https://doi.org/10.1016/0042-6989(96)00002-8

Venter, J. H. 1967, Ann. Math. Statist., 38, 1446, doi: 10.1214/aoms/1177698699

Vioque, M., Oudmaijer, R. D., Schreiner, M., et al. 2020, arXiv e-prints, arXiv:2005.01727. https://arxiv.org/abs/2005.01727

Watson, C., Povich, M. S., Churchwell, E. B., et al. 2008, ApJ, 681, 1341, doi: 10.1086/588005

Werner, M. W., Roellig, T. L., Low, F. J., et al. 2004, ApJS, 154, 1, doi: 10.1086/422992

Whitney, B. A., Robitaille, T. P., Bjorkman, J. E., et al. 2013, ApJS, 207, 30, doi: 10.1088/0067-0049/207/2/30

Williams, J. P., & Cieza, L. A. 2011, ARA&A, 49, 67, doi: 10.1146/annurev-astro-081710-102548

Winston, E., Hora, J., Gutermuth, R., & Tolls, V. 2019, ApJ, 880, 9, doi: 10.3847/1538-4357/ab27c8

Winston, E., Hora, J. L., & Tolls, V. 2020, AJ, 160, 68, doi: 10.3847/1538-3881/ab99c8

Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, AJ, 140, 1868, doi: 10.1088/0004-6256/140/6/1868

Xu, Y., Reid, M., Dame, T., et al. 2016, Science Advances, 2, e1600878, doi: 10.1126/sciadv.1600878

Xue, M., Jiang, B. W., Gao, J., et al. 2016, ApJS, 224, 23, doi: 10.3847/0067-0049/224/2/23

Yang, Y., Nie, F., Xu, D., et al. 2012, IEEE Transactions on Pattern Analysis and Machine Intelligence, 34, 723

Zari, E., Hashemi, H., Brown, A. G. A., Jardine, K., & de Zeeuw, P. T. 2018, A&A, 620, A172, doi: 10.1051/0004-6361/201834150

Zasowski, G., Majewski, S. R., Indebetouw, R., et al. 2009, ApJ, 707, 510, doi: 10.1088/0004-637X/707/1/510

Zavagno, A., Deharveng, L., Comerón, F., et al. 2006, A&A, 446, 171, doi: 10.1051/0004-6361:20053952

Zucker, C., Speagle, J. S., Schlafly, E. F., et al. 2020, A&A, 633, A51, doi: 10.1051/0004-6361/201936145

Table 1. Candidate YSOs

Column	Column ID	Description
1	SPICY	Candidate YSO designation
2	ra	ICRS right ascension coordinate in decimal degrees
3	dec	ICRS declination coordinate in decimal degrees
4	l	Galactic longitude
5	b	Galactic latitude
6	p1	YSO random forest score[a] from IRAC+2MASS photometry
7	p2	YSO random forest score[a] from IRAC+UKIDSS photometry
8	p3	YSO random forest score[a] from IRAC+VIRAC photometry
9	class	YSO class[b]
10	silicate	Flag for a possible strong silicate feature
11	pah	Flag for a possible strong PAH feature
12	alpha	Spectral index used for YSO class[c]
13	alpha_8	Spectral index derived from the $[4.5]-[8.0]$ color
14	alpha_24	Spectral index derived from the $[4.5]-[24]$ color
15	alpha_w4	Spectral index derived from the $[4.5]-W4$ color
16	env	Classification of the YSO environment[d] from the $3' \times 3'$ IRAC cutout
17	group	HDBSCAN group to which the star is assigned[e]
18	var	ZTF light curve variability flag[f]
19	nr	Number of good ZTF r-band observations used
20	r	ZTF mean magnitude in the r-band
21	sigmar	ZTF light curve $\sigma_{\rm var}$ standard deviation in the r-band
22	skewnessr	ZTF light curve skew in the r-band
GLIMPSE (and Extensions) Catalog Columns		
23	Spitzer	Spitzer source designation
24	mag3_6	Spitzer/IRAC channel 1 magnitude
25	e_mag3_6	Error on Spitzer/IRAC channel 1 magnitude
26	mag4_5	Spitzer/IRAC channel 2 magnitude
27	e_mag4_5	Error on Spitzer/IRAC channel 2 magnitude
28	mag5_8	Spitzer/IRAC channel 3 magnitude
29	e_mag5_8	Error on Spitzer/IRAC channel 3 magnitude
30	mag8_0	Spitzer/IRAC channel 4 magnitude
31	e_mag8_0	Error on Spitzer/IRAC channel 4 magnitude
32	csf	Close source flag
33	m3_6	number of detections in the 3.6 μm band
34	m4_5	number of detections in the 4.5 μm band
35	m5_8	number of detections in the 5.8 μm band
36	m8_0	number of detections in the 8.0 μm band
Cross-Matched Catalogs		
37	2MASS	2MASS source designation
38	UKIDSS	UKIDSS source designation
39	VIRAC	VIRAC DR1 source designation
40	GaiaDR2	Gaia DR2 source designation
41	MIPS	Spitzer/MIPS source designation
42	AllWISE	AllWISE source designation
43	ZTFDR3	ZTF DR3 source designation

Note— In addition to the quantities derived in this paper, for the convenience of the user, this table also provides select columns from the GLIMPSE (and extensions) catalogs.
(This table is available in its entirety in a machine-readable form in the online journal. The list of columns is shown here for guidance regarding its form and content.)

[a] Random forest scores range from 0 to 1, where higher scores indicate a greater chance that an object is a YSO. We include sources with scores >0.5 from one of the classifiers.

[b] YSO classes are "Class I," "FS" (flat SED), "Class II," and "Class III."

[c] This α combines the results from Columns 13–15 as described in Section 5.5.

[d] The environment classes are "EnvI" (no or minimal nebulosity), "EnvII" (mixed category), "EnvIII" (cloud-like environment).

[e] The list of groups is provided in Table 2.

[f] The variability classes are 1 (weak or statistically insignificant variability), 2 (moderate variability), 3 (high variability).

Table 2. YSO Groups from HDBSCAN

Column	Column ID	Description
1	group	Group designation
2	l0	Central Galactic longitude ℓ_0 [deg]
3	b0	Central Galactic latitude b_0 [deg]
4	plx	Mean parallax [mas]
5	e_plx	Error on mean parallax [mas]
6	pml	Mean proper motion in ℓ [mas yr^{-1}]
7	e_pml	Error on mean proper motion in ℓ [mas yr^{-1}]
8	pmb	Mean proper motion in b [mas yr^{-1}]
9	e_pmb	Error on mean proper motion in b [mas yr^{-1}]
10	n	Total number of constituents
11	nG	Number of constituents with 5-parameter Gaia astrometric solutions
12	flag	Flag for potential model problems

NOTE—Properties of YSO groups identified from the HDBSCAN algorithm. Median astrometric properties, including group parallax and proper motion, are inferred from the hierarchical Bayesian modeling of the Gaia DR2 astrometry. The group parallaxes and proper motions in this table are in the Gaia DR2 system, with no correction for zero-point offsets. We report formal (MAD) uncertainties from our model added in quadrature to the ± 0.04 mas and ± 0.07 mas yr^{-1} spatially correlated systematic errors on DR2 zero points (Lindegren et al. 2018). Groups are flagged if potential problems could affect interpretation of the Bayesian model as described in Section 7.1.

Table 3. Regions Near the Galactic Midplane without Significant YSO Populations

$\ell_{\rm min}$ (deg)	$\ell_{\rm max}$ (deg)	$b_{\rm min}$ (deg)	$b_{\rm max}$ (deg)
275.63	277.85	−1.52	0.51
328.95	331.37	−2.97	−1.17
349.02	349.40	−2.22	−0.09
358.32	2.08	0.81	3.87
1.25	1.44	−0.33	−0.08
0.12	2.28	−4.43	−3.11
0.12	6.43	−3.12	−1.88
23.89	26.47	−3.01	−1.22
28.84	31.46	−2.97	−1.30
32.06	32.73	−1.14	−0.23
43.22	44.37	0.28	1.13
47.79	48.46	0.20	0.86

NOTE—This table provides the lower left and upper right boundaries of the rectangular regions from which we obtained our labeled list of "field" objects. Rectangles are drawn with lines of constant Galactic ℓ and b.

Trumpler, R. J. (1930). "Preliminary results on the distances, dimensions and space distribution of open star clusters." In: *Lick Observatory Bulletin* 420, pp. 154–188.

Lin, C. C. and F. H. Shu (1964). "On the Spiral Structure of Disk Galaxies." In: *ApJ* 140, p. 646.

Toomre, A. (1964). "On the gravitational stability of a disk of stars." In: *ApJ* 139, pp. 1217–1238.

Hinton, G. (1989). "Connectionist learning procedures." In: *Artificial Intelligence* 40.1–3, pp. 185–234.

Friel, E. D. (1995). "The Old Open Clusters Of The Milky Way." In: *ARA&A* 33, pp. 381–414.

Ester, M. et al. (1996). "A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. KDD'96. Portland, Oregon: AAAI Press, pp. 226–231.

Perryman, M. A. C. et al. (1997). "The Hipparcos Catalogue." In: *A&A* 500, pp. 501–504.

Dias, W. S. et al. (2002). "New catalogue of optically visible open clusters and candidates." In: *A&A* 389, pp. 871–873.

Friel, E. D. et al. (2002). "Metallicities of Old Open Clusters." In: *AJ* 124.5, pp. 2693–2720.

Lada, C. J. and E. A. Lada (2003). "Embedded Clusters in Molecular Clouds." In: *ARA&A* 41, pp. 57–115.

Robin, A. C. et al. (2003). "A synthetic view on structure and evolution of the Milky Way." In: *A&A* 409, pp. 523–540.

Werner, M. W. et al. (2004). "The Spitzer Space Telescope Mission." In: *ApJS* 154.1, pp. 1–9.

Dias, W. S. and J. R. D. Lépine (2005). "Direct Determination of the Spiral Pattern Rotation Speed of the Galaxy." In: *ApJ* 629.2, pp. 825–831.

Skrutskie, M. F. et al. (2006). "The Two Micron All Sky Survey (2MASS)." In: *AJ* 131, pp. 1163–1183.

Lawrence, A. et al. (2007). "The UKIRT Infrared Deep Sky Survey (UKIDSS)." In: *MNRAS* 379.4, pp. 1599–1617.

Dewdney, P. E. et al. (2009). "The Square Kilometre Array." In: *IEEE Proceedings* 97.8, pp. 1482–1496.

Dobbs, C. L. and J. E. Pringle (2010). "Age distributions of star clusters in spiral and barred galaxies as a test for theories of spiral structure." In: *MNRAS* 409.1, pp. 396–404.

Zaharia, M. et al. (2012). "Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing." In: *9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)*. San Jose, CA: USENIX Association, pp. 15–28.

Friel, E. D. (2013). "Open Clusters and Their Role in the Galaxy." In: *Planets, Stars and Stellar Systems. Volume 5: Galactic Structure and Stellar Populations*. Ed. by T. D. Oswalt and G. Gilmore. Vol. 5, p. 347.

Frinchaboy, P. M. et al. (2013). "The Open Cluster Chemical Analysis and Mapping Survey: Local Galactic Metallicity Gradient with APOGEE Using SDSS DR10." In: *ApJ* 777.1, L1, p. L1.

Kharchenko, N. V. et al. (2013). "Global survey of star clusters in the Milky Way. II. The catalogue of basic parameters." In: *A&A* 558, A53, A53.

Randich, S., G. Gilmore, and Gaia-ESO Consortium (2013). "The Gaia-ESO Large Public Spectroscopic Survey." In: *The Messenger* 154, pp. 47–49.

Krone-Martins, A. and A. Moitinho (2014). "UPMASK: unsupervised photometric membership assignment in stellar clusters." In: *A&A* 561, A57, A57.

Reid, M. J. et al. (2014). "Trigonometric Parallaxes of High Mass Star Forming Regions: The Structure and Kinematics of the Milky Way." In: *ApJ* 783, 130, p. 130.

Junqueira, T. C. et al. (2015). "A new method for estimating the pattern speed of spiral structure in the Milky Way." In: *MNRAS* 449.3, pp. 2336–2344.

Michalik, D., L. Lindegren, and D. Hobbs (2015). "The Tycho-Gaia astrometric solution - How to get 2.5 million parallaxes with less than one year of Gaia data." In: *A&A* 574, A115.

Netopil, M., E. Paunzen, and G. Carraro (2015). "A comparative study on the reliability of open cluster parameters." In: *A&A* 582, A19, A19.

Casamiquela, L. et al. (2016). "The OCCASO survey: presentation and radial velocities of 12 Milky Way open clusters." In: *MNRAS* 458.3, pp. 3150–3167.

Gaia Collaboration et al. (2016a). "Gaia Data Release 1. Summary of the astrometric, photometric, and survey properties." In: *A&A* 595, A2, A2.

Gaia Collaboration et al. (2016b). "The Gaia mission." In: *A&A* 595, A1, A1.

Lindegren, L. et al. (2016). "Gaia Data Release 1. Astrometry: one billion positions, two million proper motions and parallaxes." In: *A&A* 595, A4, A4.

Netopil, M. et al. (2016). "On the metallicity of open clusters. III. Homogenised sample." In: *A&A* 585, A150, A150.

van der Marel, R. P. and J. Sahlmann (2016). "First Gaia Local Group Dynamics: Magellanic Clouds Proper Motion and Rotation." In: *ApJ* 832.2, L23, p. L23.

Blanton, M. R. et al. (2017). "Sloan Digital Sky Survey IV: Mapping the Milky Way, Nearby Galaxies, and the Distant Universe." In: *AJ* 154.1, 28, p. 28.

Bovy, J. (2017). "Galactic rotation in Gaia DR1." In: *MNRAS* 468.1, pp. L63–L67.

Casamiquela, L. et al. (2017). "OCCASO - II. Physical parameters and Fe abundances of red clump stars in 18 open clusters." In: *MNRAS* 470.4, pp. 4363–4381.

Gaia Collaboration et al. (2017). "Gaia Data Release 1 - Open cluster astrometry: performance, limitations, and future prospects." In: *A&A* 601, A19.

Helmi, A. et al. (2017). "A box full of chocolates: The rich structure of the nearby stellar halo revealed by Gaia and RAVE." In: *A&A* 598, A58, A58.

Tejedor, E. et al. (2017). "PyCOMPSs: Parallel computational workflows in Python." In: *The International Journal of High Performance Computing Applications* 31.1, pp. 66–82.

Andrae, R. et al. (2018). "Gaia Data Release 2. First stellar parameters from Apsis." In: *A&A* 616, A8, A8.

Antoja, T. et al. (2018). "A dynamically young and perturbed Milky Way disk." In: *Nature* 561.7723, pp. 360–362.

Buder, S. et al. (2018). "The GALAH Survey: second data release." In: *MNRAS* 478.4, pp. 4513–4552.

Cantat-Gaudin, T. et al. (2018). "A Gaia DR2 view of the open cluster population in the Milky Way." In: *A&A* 618, A93, A93.

Castro-Ginard, A. et al. (2018). "A new method for unveiling open clusters in Gaia. New nearby open clusters confirmed by DR2." In: *A&A* 618, A59, A59.

Gaia Collaboration et al. (2018). "Gaia Data Release 2. Summary of the contents and survey properties." In: *A&A* 616, A1, A1.

Helmi, A. et al. (2018). "The merger that led to the formation of the Milky Way's inner stellar halo and thick disk." In: *Nature* 563.7729, pp. 85–88.

Hunt, J. A. S. et al. (2018). "Transient spiral structure and the disc velocity substructure in Gaia DR2." In: *MNRAS* 481.3, pp. 3794–3803.

Luri, X. et al. (2018). "Gaia Data Release 2. Using Gaia parallaxes." In: *Astronomy and Astrophysics* 616, A9, A9.

Mor, R. et al. (2018). "BGM FASt: Besançon Galaxy Model for big data. Simultaneous inference of the IMF, SFH, and density in the solar neighbourhood." In: *A&A* 620, A79, A79.

Quillen, A. C. et al. (2018). "Spiral arm crossings inferred from ridges in Gaia stellar velocity distributions." In: *MNRAS* 480.3, pp. 3132–3139.

Shabani, F. et al. (2018). "Search for star cluster age gradients across spiral arms of three LEGUS disc galaxies." In: *MNRAS* 478.3, pp. 3590–3604.

Smith, L. C. et al. (2018). "VIRAC: the VVV Infrared Astrometric Catalogue." In: *MNRAS* 474.2, pp. 1826–1849.

Zari, E. et al. (2018). "3D mapping of young stars in the solar neighbourhood with Gaia DR2." In: *A&A* 620, A172, A172.

Ahumada, R. et al. (2019). "The Sixteenth Data Release of the Sloan Digital Sky Surveys: First Release from the APOGEE-2 Southern Survey and Full Release of eBOSS Spectra." In: *arXiv e-prints*, arXiv:1912.02905, arXiv:1912.02905.

Álvarez Cid-Fuentes, J. et al. (2019). "dislib: Large-scale High Performance Machine Learning in Python." In: *Proceedings of the 15th International Conference of eScience*, pp. 96–105.

Anders, F. et al. (2019). "Photo-astrometric distances, extinctions, and astrophysical parameters for Gaia DR2 stars brighter than G = 18." In: *A&A* 628, A94, A94.

Cantat-Gaudin, T. et al. (2019a). "Expanding associations in the Vela-Puppis region. 3D structure and kinematics of the young population." In: *A&A* 626, A17, A17.

Cantat-Gaudin, T. et al. (2019b). "Gaia DR2 unravels incompleteness of nearby cluster population: new open clusters in the direction of Perseus." In: *A&A* 624, A126, A126.

Castro-Ginard, A. et al. (2019). "Hunting for open clusters in *Gaia* DR2: the Galactic anticentre." In: *A&A* 627, A35, A35.

Galli, P. A. B. et al. (2019). "Structure and kinematics of the Taurus star-forming region from Gaia-DR2 and VLBI astrometry." In: *A&A* 630, A137, A137.

Ivezić, Ž. et al. (2019). "LSST: From Science Drivers to Reference Design and Anticipated Data Products." In: *ApJ* 873.2, 111, p. 111.

Kounkel, M. and K. Covey (2019). "Untangling the Galaxy. I. Local Structure and Star Formation History of the Milky Way." In: *AJ* 158.3, 122, p. 122.

Leung, H. W. and J. Bovy (2019). "Simultaneous calibration of spectrophotometric distances and the Gaia DR2 parallax zero-point offset with deep learning." In: *MNRAS* 489.2, pp. 2079–2096.

Lim, B. et al. (2019). "A Gaia view of the two OB associations Cygnus OB2 and Carina OB1: the signature of their formation process." In: *MNRAS* 490.1, pp. 440–454.

Liu, L. and X. Pang (2019). "A Catalog of Newly Identified Star Clusters in Gaia DR2." In: *ApJS* 245.2, 32, p. 32.

Mor, R. et al. (2019). "Gaia DR2 reveals a star formation burst in the disc 2-3 Gyr ago." In: *A&A* 624, L1, p. L1.

Reid, M. J. et al. (2019). "Trigonometric Parallaxes of High-mass Star-forming Regions: Our View of the Milky Way." In: *ApJ* 885.2, 131, p. 131.

Romero-Gómez, M. et al. (2019). "Gaia kinematics reveal a complex lopsided and twisted Galactic disc warp." In: *Astronomy and Astrophysics* 627, A150, A150.

Röser, S., E. Schilbach, and B. Goldman (2019a). "Hyades tidal tails revealed by Gaia DR2." In: *A&A* 621, L2, p. L2.

Röser, S. and E. Schilbach (2019b). "Praesepe (NGC 2632) and its tidal tails." In: *A&A* 627, A4, A4.

Sim, G. et al. (2019). "Open cluster survey within 1 kpc by the Gaia DR2." In: *arXiv e-prints*, arXiv:1907.06872, arXiv:1907.06872.

Tang, S.-Y. et al. (2019). "Discovery of Tidal Tails in Disrupting Open Clusters: Coma Berenices and a Neighbor Stellar Group." In: *ApJ* 877.1, 12, p. 12.

Yeh, F. C. et al. (2019). "Ruprecht 147: A Paradigm of Dissolving Star Cluster." In: *AJ* 157.3, 115, p. 115.

Zečević, P. et al. (2019). "AXS: A Framework for Fast Astronomical Data Processing Based on Apache Spark." In: *AJ* 158.1, 37, p. 37.

Cantat-Gaudin, T. and F. Anders (2020a). "Clusters and mirages: cataloguing stellar aggregates in the Milky Way." In: *A&A* 633, A99, A99.

Cantat-Gaudin, T. et al. (2020b). "Painting a portrait of the Galactic disc with its stellar clusters." In: *A&A* 640, A1, A1.

Castro-Ginard, A. et al. (2020). "Hunting for open clusters in Gaia DR2: 582 new open clusters in the Galactic disc." In: *A&A* 635, A45, A45.

Gaia Collaboration et al. (2020a). "Gaia Early Data Release 3: Summary of the contents and survey properties." In: *arXiv e-prints*, arXiv:2012.01533, arXiv:2012.01533.

Gaia Collaboration et al. (2020b). "Gaia Early Data Release 3: The Gaia Catalogue of Nearby Stars." In: *arXiv e-prints*, arXiv:2012.02061, arXiv:2012.02061.

Garabato Míguez, D. (2020). "Análisis no supervisado de observaciones atípicas en la misión espacial *Gaia*; optimización mediante procesamiento distribuido e integración en APSIS." PhD thesis. Universidade da Coruña.

Hunt, E. L. and S. Reffert (2020). "Improving the open cluster census. I. Comparison of clustering algorithms applied to Gaia DR2 data." In: *arXiv e-prints*, arXiv:2012.04267, arXiv:2012.04267.

Kounkel, M., K. Covey, and K. G. Stassun (2020). "Untangling the Galaxy. II. Structure within 3 kpc." In: *AJ* 160.6, 279, p. 279.

Kovaleva, D. A. et al. (2020). "Collinder 135 and UBC 7: A physical pair of open clusters." In: *A&A* 642, L4, p. L4.

Kuhn, M. A. et al. (2020). "SPICY: The Spitzer/IRAC Candidate YSO Catalog for the Inner Galactic Midplane." In: *arXiv e-prints*, arXiv:2011.12961, arXiv:2011.12961.

Mor, R., X. Luri, and GAIA UB Team (2020). "Expanding Big Data mining for Astronomy." In: *Contributions to the XIV.o Scientific Meeting (virtual) of the Spanish Astronomical Society*, p. 235.

Necib, L. et al. (2020). "Evidence for a vast prograde stellar stream in the solar vicinity." In: *Nature Astronomy* 4, pp. 1078–1083.

Ostdiek, B. et al. (2020). "Cataloging accreted stars within Gaia DR2 using deep learning." In: *A&A* 636, A75, A75.

Piatti, A. E. (2020). "Binary star sequence in the outskirts of the disrupting Galactic open cluster UBC 274." In: *A&A* 639, A55, A55.

Tarricq, Y. et al. (2020). "3D kinematics and age distribution of the Open Cluster population." In: *arXiv e-prints*, arXiv:2012.04017.

Zhang, Y. et al. (2020). "Diagnosing the Stellar Population and Tidal Structure of the Blanco 1 Star Cluster." In: *ApJ* 889.2, 99, p. 99.

Anders, F. et al. (2021). "The star cluster age function in the Galactic disc with Gaia DR2. Fewer old clusters and a low cluster formation efficiency." In: *A&A* 645, L2, p. L2.

The bibliography printed above refers to citations in the Introduction, in individual chapter introductions and in Summary of Results, Discussions and Conclusions. Citations in each individual publication can be found listed within the corresponding publication.

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede and Ivo Pletikosić. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

https://bitbucket.org/amiede/classicthesis/

Happy users of `classicthesis` usually send a real postcard to the author, a collection of postcards received so far is featured here:

http://postcards.miede.de/

Thank you very much for your feedback and contribution.

Final Version as of March 5, 2021 (classicthesis v4.6).