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

A. Castro-Ginard, C. Jordi, X. Luri, J. Álvarez Cid-Fuentes, L. Casamiquela, F. Anders, T. Cantat-Gaudin, M. Monguió, L. Balaguer-Núñez, S. Solà, et al.

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

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, $G_{\rm BP}$, and $G_{\rm RP}$), 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.

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,

[3] https://pytorch.org/

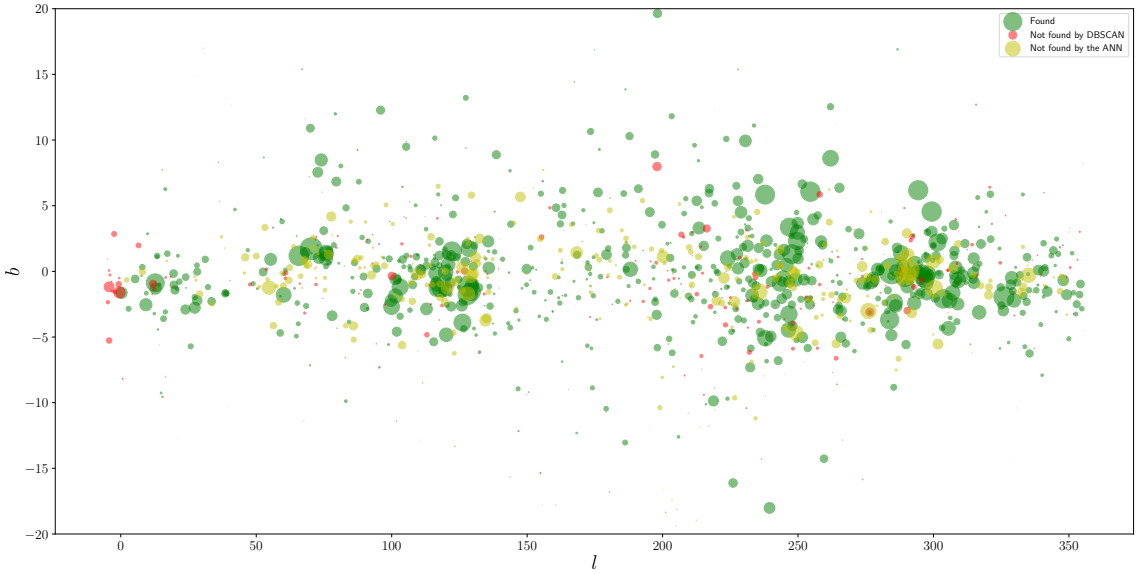

Fig. 1. Distribution in Galactic coordinates (l vs. b) of the OCs catalogued in Cantat-Gaudin et al. (2018). Green dots represent OCs that our method recovers, red dots are OCs not found by DBSCAN, and yellow dots are OCs which are found by DBSCAN but for which the CMD is not recognised by our ANN. The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

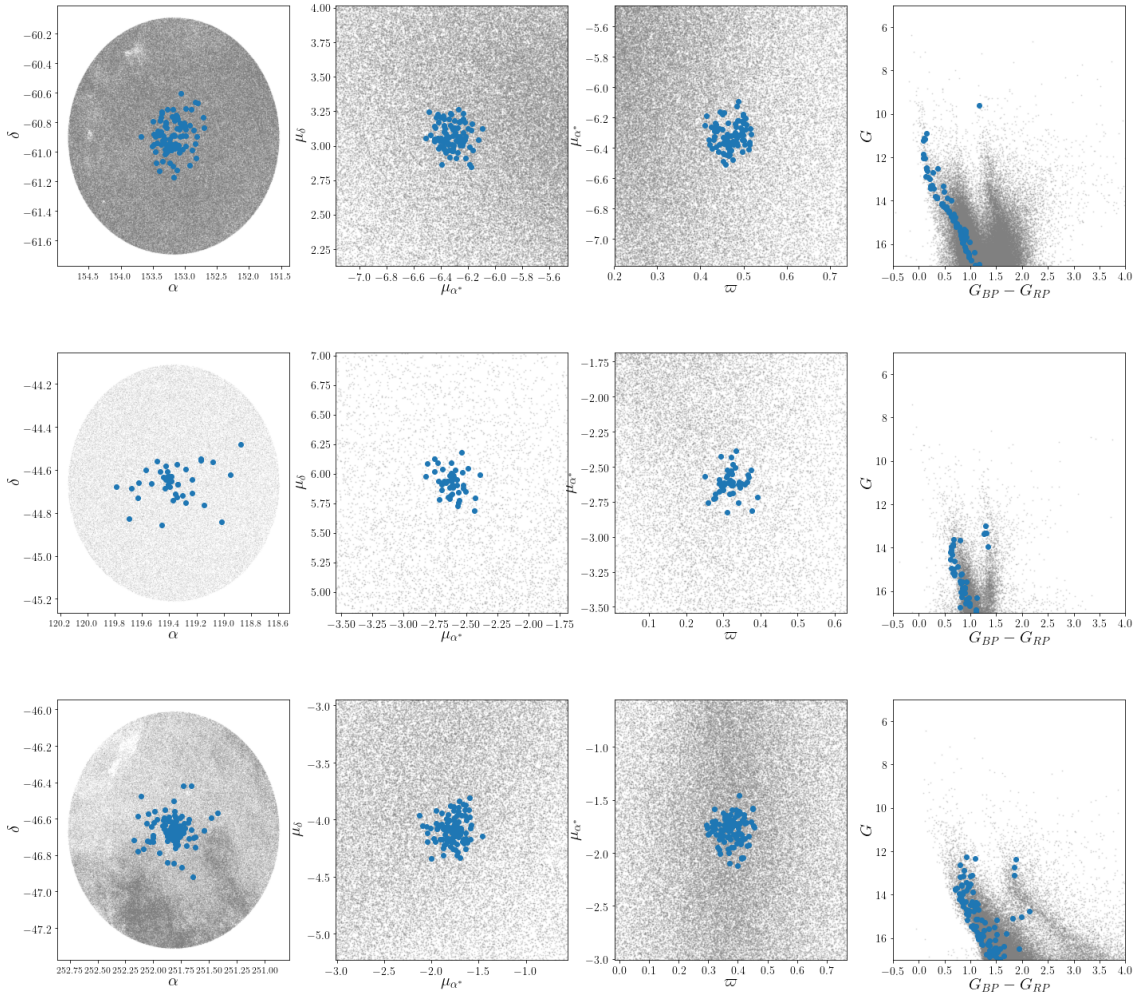

Fig. 2. Examples of class A (*top row*), class B (*middle row*), and class C (*bottom row*) clusters. The columns represent, *from left to right*, a distribution of the member stars (in blue) and field stars (grey) for: i) position in (α, δ), ii) proper motions in $(\mu_{\alpha^*}, \mu_\delta)$, iii) 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

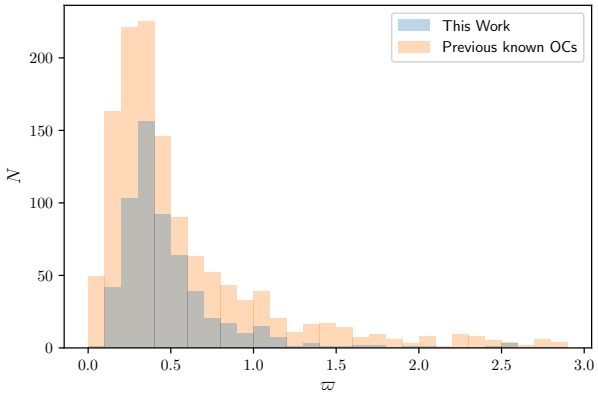

Fig. 3. Parallax histogram of the new OCs (light blue) and OCs known previous to this study (light orange), i.e. CG18, CG19, Cantat-Gaudin et al. (2018), and Cantat-Gaudin et al. (2019a).

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

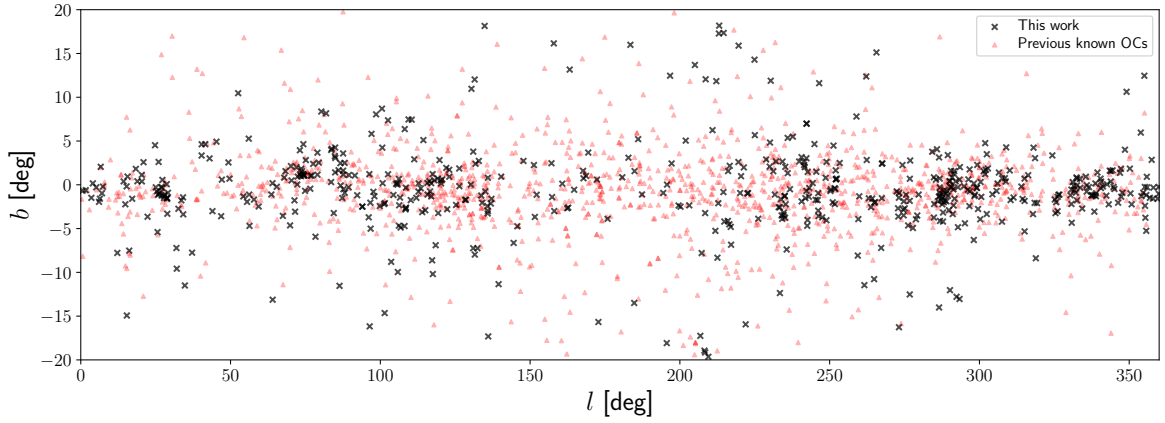

Fig. 4. Distribution of the OC census in l vs. b. Black crosses represent new OCs while red triangles represent OCs in CG18, CG19, Cantat-Gaudin et al. (2018), and Cantat-Gaudin et al. (2019a).

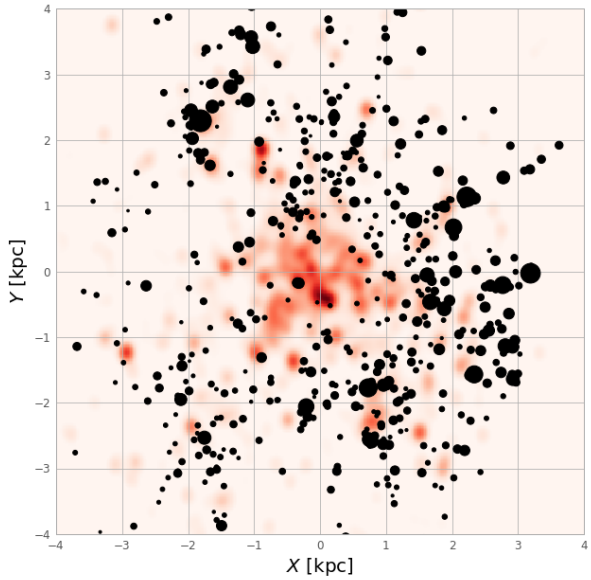

Fig. 5. Distribution of the OCs projected in the $X - Y$ plane. Previously known OCs (CG18,CG19, Cantat-Gaudin et al. 2018, 2019a) are shown as a density map in red. Newly found OCs reported here are shown as black dots, where the size is proportional to the number of members of each cluster.

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

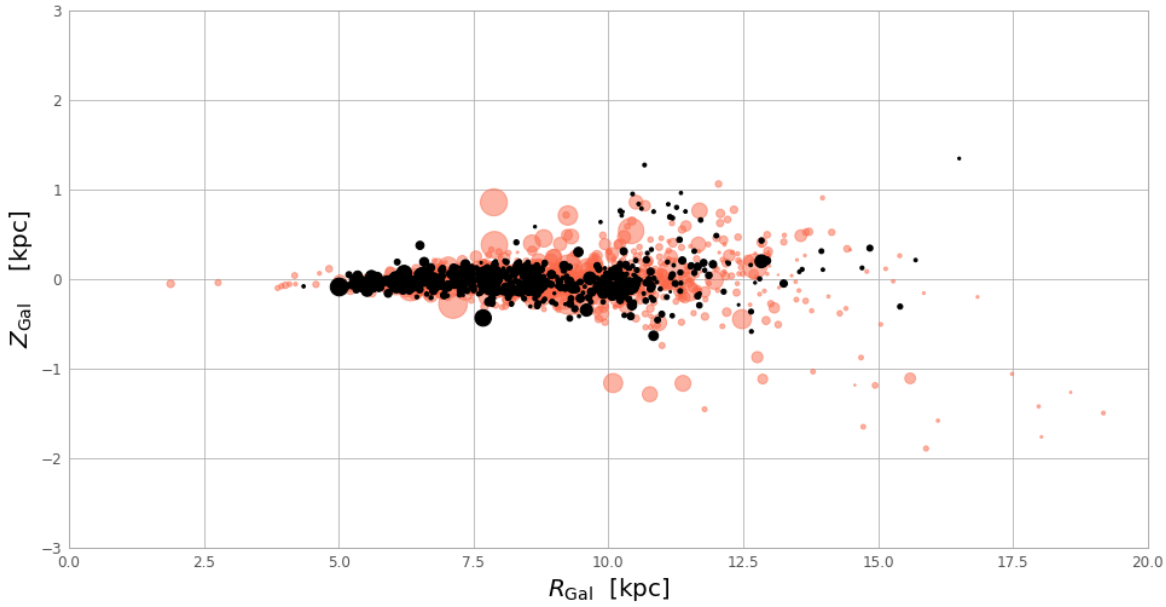

Fig. 6. Distribution of the OCs in $R - Z$ in Galacto-centric coordinates. Previously known OCs (CG18,CG19, Cantat-Gaudin et al. 2018, 2019a) are shown as red dots while newly found OCs are shown in black dots; the sizes of the dots are proportional to the number of members of each cluster.

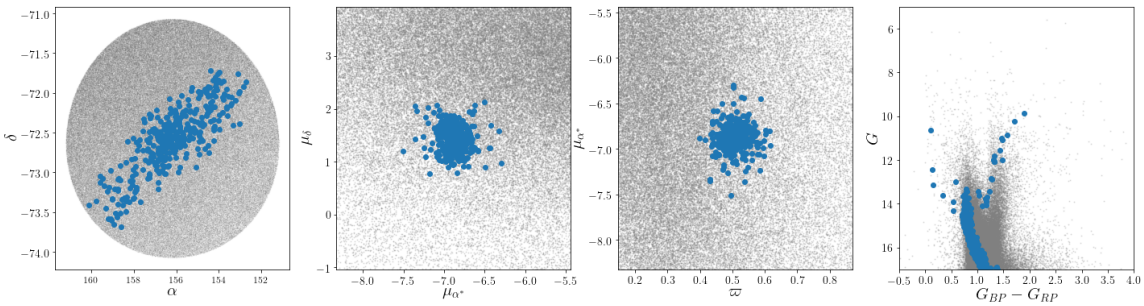

Fig. 7. Distribution of the member stars of UBC 274 (blue points) in comparison with field stars (grey points). The leftmost plot is a distribution in position (α, δ). The inner left plot shows the proper motion vector diagram while the inner right plot includes the parallax (ϖ, μ_{α^*}). The rightmost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

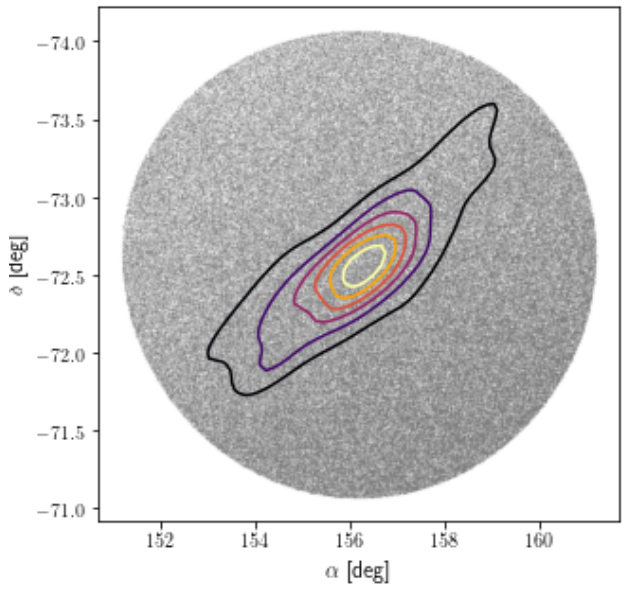

Fig. 8. Density contours for the members in cluster UBC 274, and field stars (grey points). UBC 274 shows an elongated shape in its outskirts.

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_{BP}−G_{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.

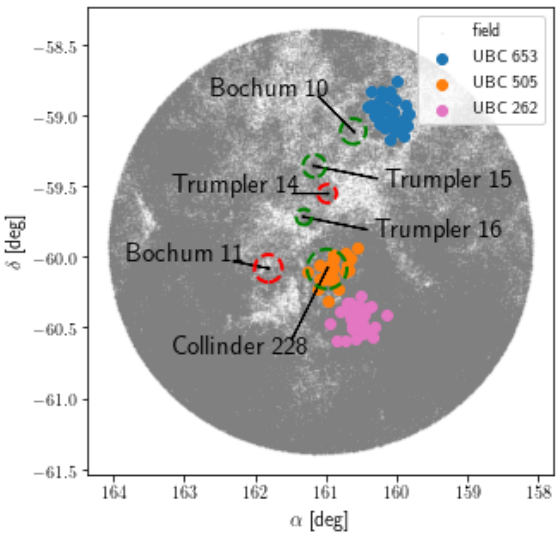

Fig. 9. Region around the Carina Nebula. Grey points represent field stars, while points in blue, orange, and pink represent UBC 653, UBC 505, and UBC 262 respectively. The dashed circle represents the locations of the OCs Cantat-Gaudin et al. (2018), which are related to the Carina Nebula. Dashed green 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