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

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.

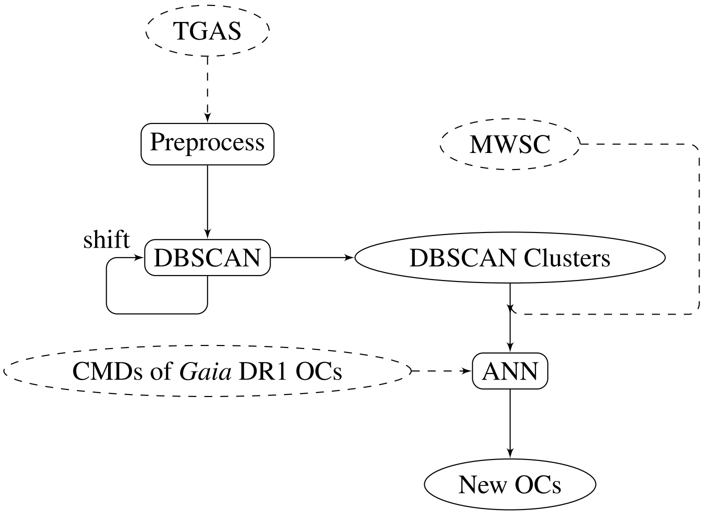

Fig. 1. Flow chart of the method applied to find OCs. Solid boxes represent code, solid ellipses represent generated catalogues, and dashed ellipses represent external catalogues.

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

[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).

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.

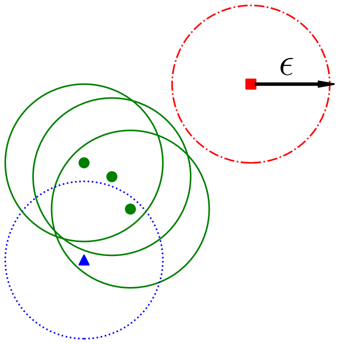

Fig. 2. Schematic representation of a DBSCAN cluster with *minPts* = 3. Points in green represent core points, each point has *minPts* points in its (green solid) hypersphere. The blue triangle represents a member point, it does not have *minPts* in its (blue dashed) hypersphere but it is reached by 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.

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:

- 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.

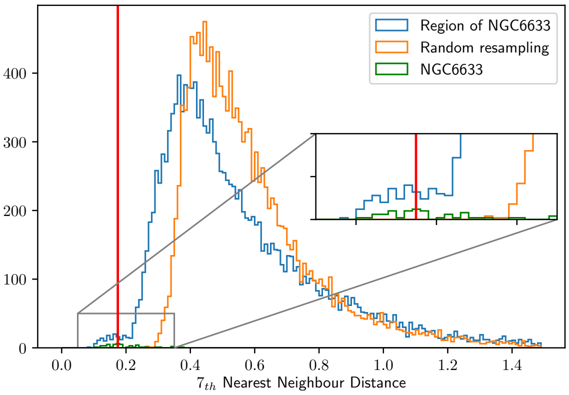

Fig. 3. Histogram of the $7_{\rm th}$-NNDs of the region around the cluster NGC 6633. The blue line shows the $7_{\rm th}$-NND histogram of all the stars in that sky region in TGAS. Orange line shows the $7_{\rm th}$-NND histogram of one realization of a random resample. Green line shows the $7_{\rm th}$-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 $minPts = 8$.

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 $v_i = \omega_{i1}x_1 + \omega_{i2}x_2 + \cdots + \omega_{in}x_n$, and responds according to an activation function, in our case we use a hyperbolic tangent activation function

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

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.

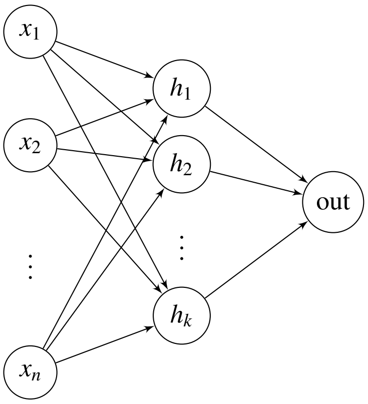

Fig. 4. Schematic representation of a multilayer perceptron with one hidden layer. The x_i values represent the input data. The h_i labels represent neurons in the hidden layer.

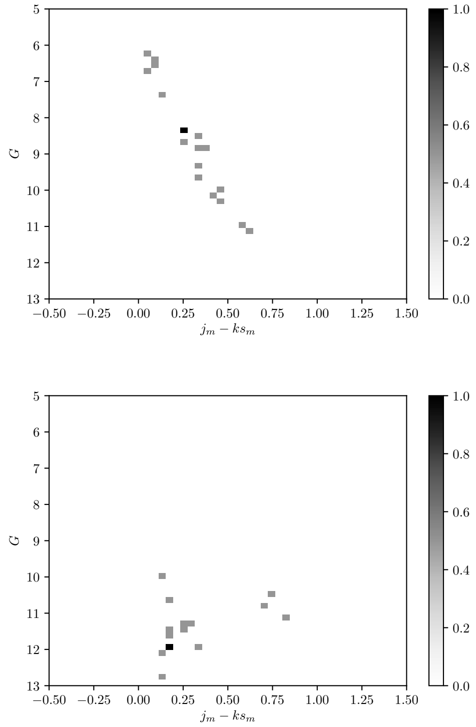

Fig. 5. Examples of training data for the ANN classificator. The upper plot corresponds to a density map of a CMD of a subset of the members of Coma Berenices. The lower plot is the density map of a CMD of a cluster found by DBSCAN that we labelled as noise. In both cases, the colours represent the value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

2.3.3. 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_{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.

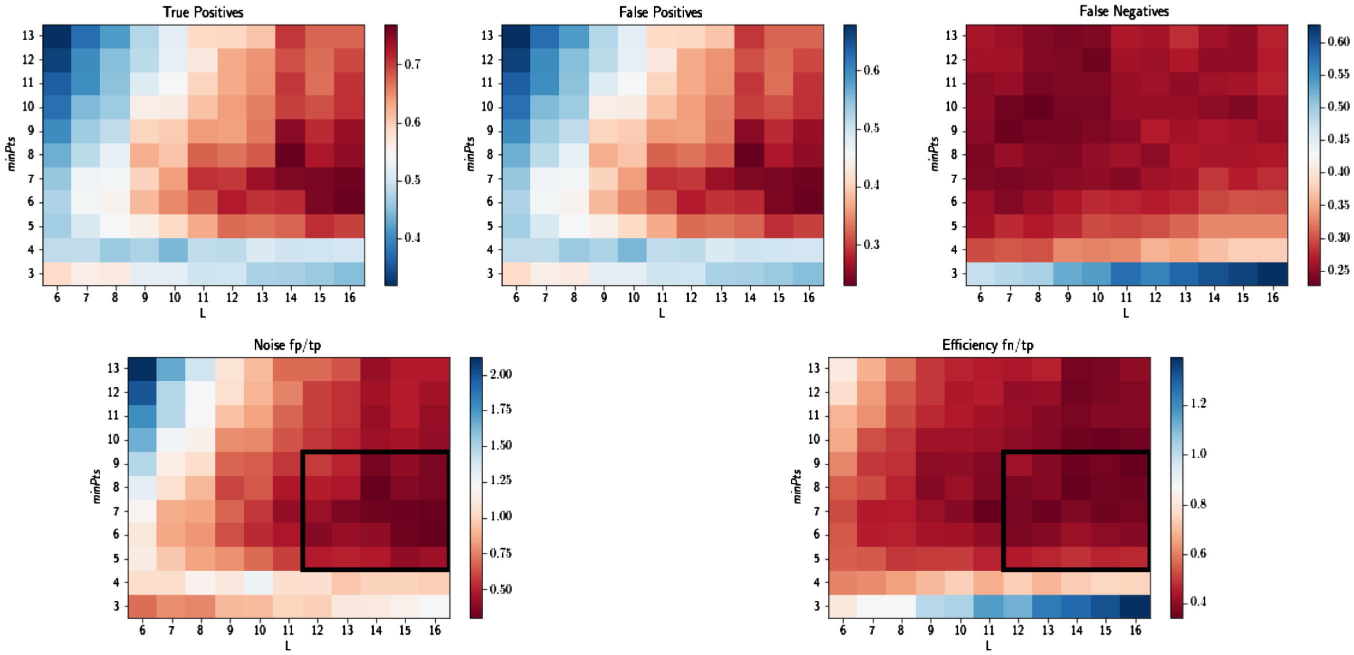

Fig. 6. Performance of the algorithm with a different set of parameters ($L, minPts$) tested with simulated data. *Top panels*: true positive (*left*), false positive *(middle*), and false negative (*right*) rates. We highlight the inversion of the colour bar in the true positive rate to always represent 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 ($N_{\text{found}} < 5$) within the explored parameters ($L, minPts$). The criteria followed in order to sort the list of candidates is reasonable; 100% of the clusters with $N_{\text{found}} \geq 5$ are confirmed, while for $N_{\text{found}} < 5$, 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).

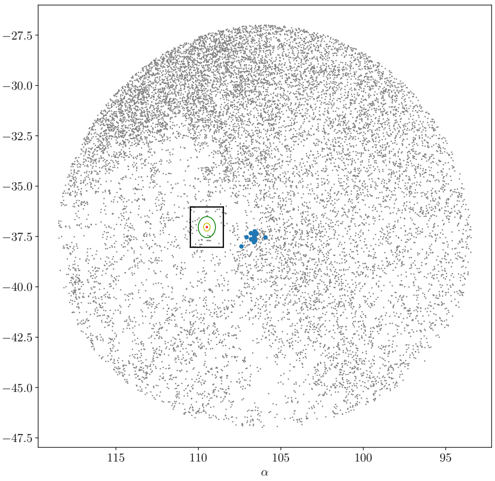

Fig. 7. Cone search of 10 deg centred in UBC7 in the TGAS data with more than 120 photometric observations. Blue dots represent members of UBC7. The red, yellow, and green circles represent the r_0, r_1 and r_2 radius in the MWSC catalogue for Collinder 135. The black box is the 2 deg × 2 deg zone where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

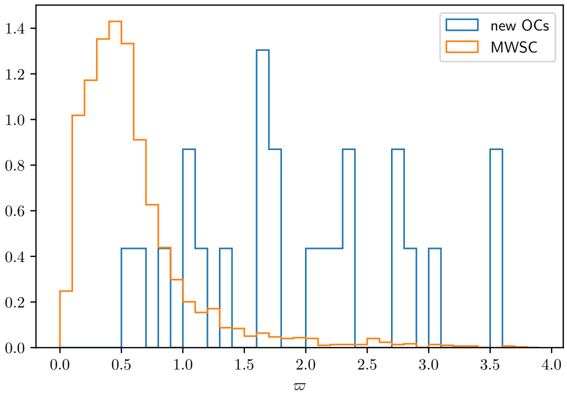

Fig. 8. Normalized parallax distribution of the found OCs (blue) and the ones listed in MWSC (orange). The newly detected OCs are closer than most of the catalogued clusters in MWSC.

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.

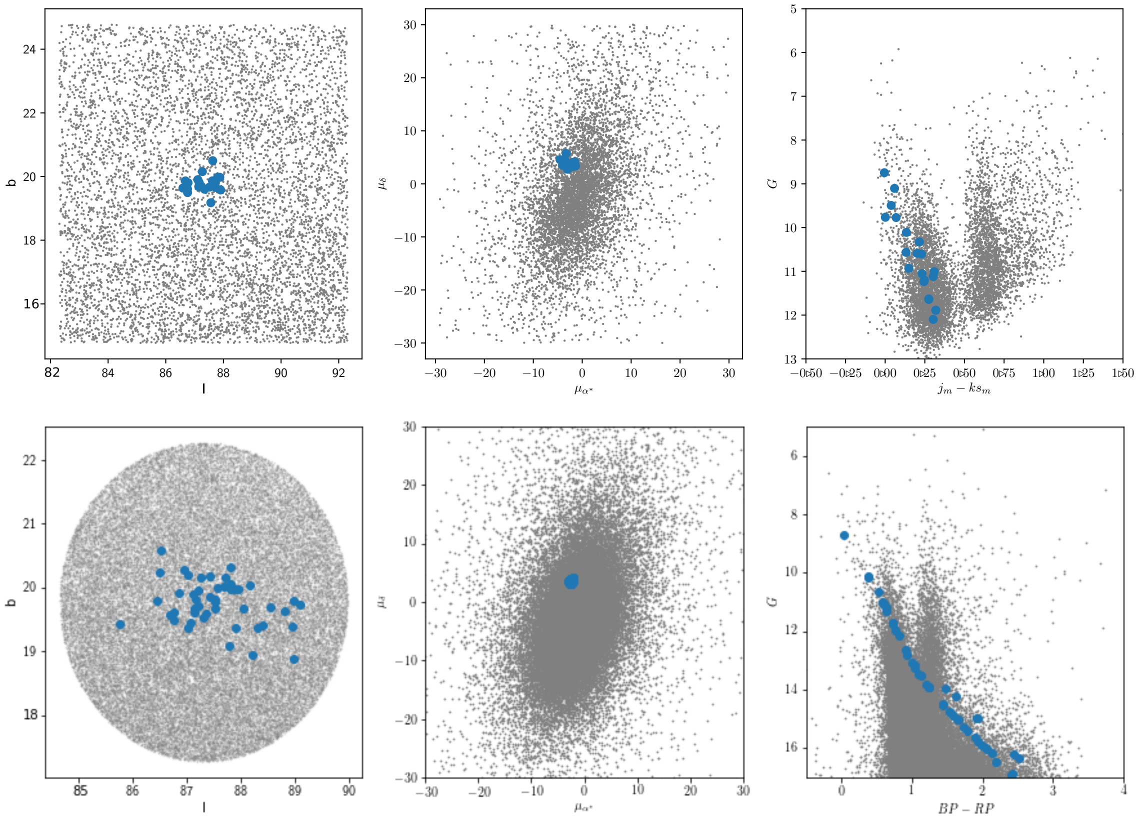

Fig. 9. Visualisation of UBC1 from Table 1. *Top panels left plot*: position of the member stars (blue) along with field stars (grey) in a 10 deg × 10 deg area in TGAS data. *Middle plot*: same stars in the proper motion space. *Right plot*: CMD of the stars in the field using photometry from *Gaia* 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).

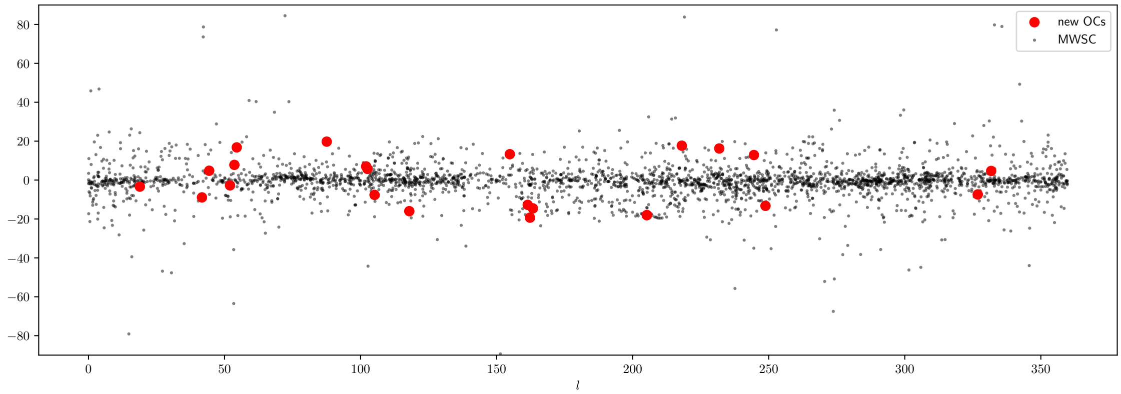

Fig. 10. Spatial distribution in (l, b) of the found OCs (red) together with the ones listed in MWSC (black). The confirmed OCs tend to be at latitudes $|b| > 5$ deg.

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

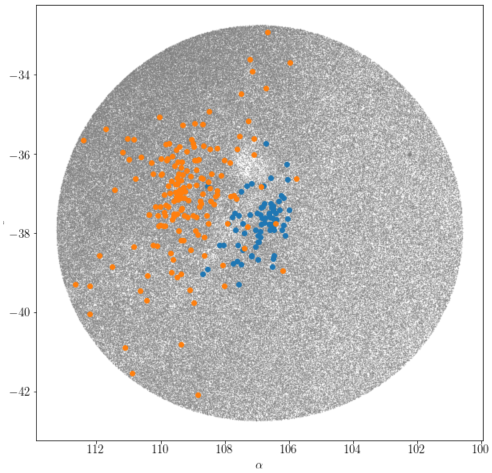

Fig. 11. Cone search of 5 deg in the area of UBC7 (blue) and Collinder 135 (orange). The grey dots correspond to the stars brighter than $G = 17$ mag with more than 120 photometric observations in *Gaia* DR2 data. We have checked that the lower stellar density between the two clusters only appears for 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, $\mu_{\alpha*}$, 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 $N_{\mathrm{found}} > 5$. 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

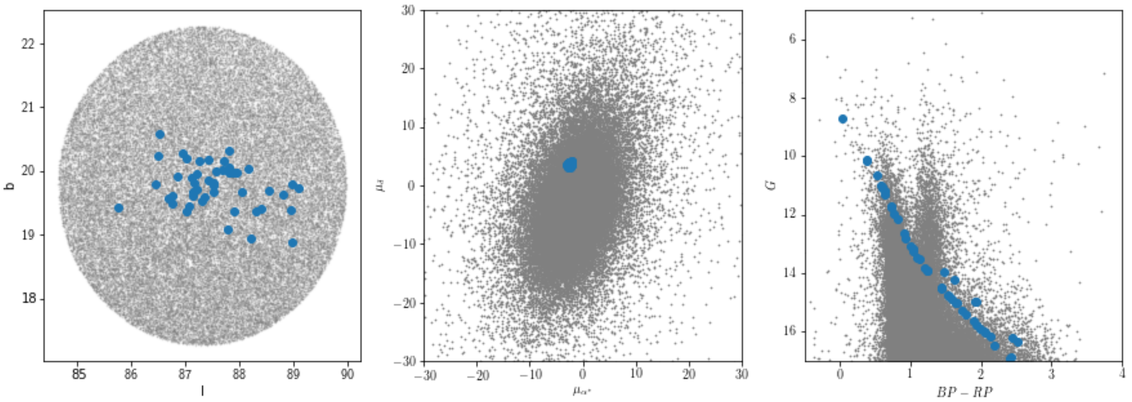

Fig. A.1. Member stars (blue) together with field stars (grey) for UBC1 in (l, b) (*left panel*) and in proper motion space (*middle panel*). The CMD shows the sequence of the identified members (outlining an empirical isochrone) (*right panel*).

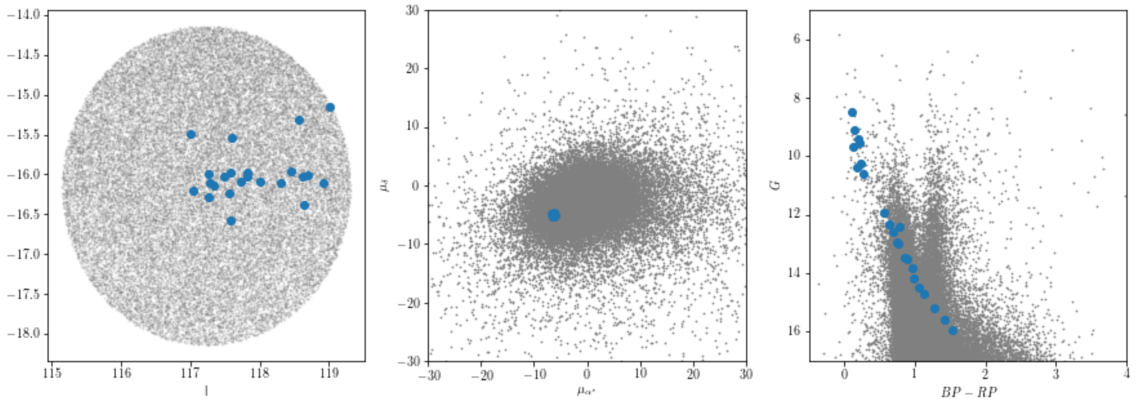

Fig. A.2. As in Fig. A.1 but for UBC2.

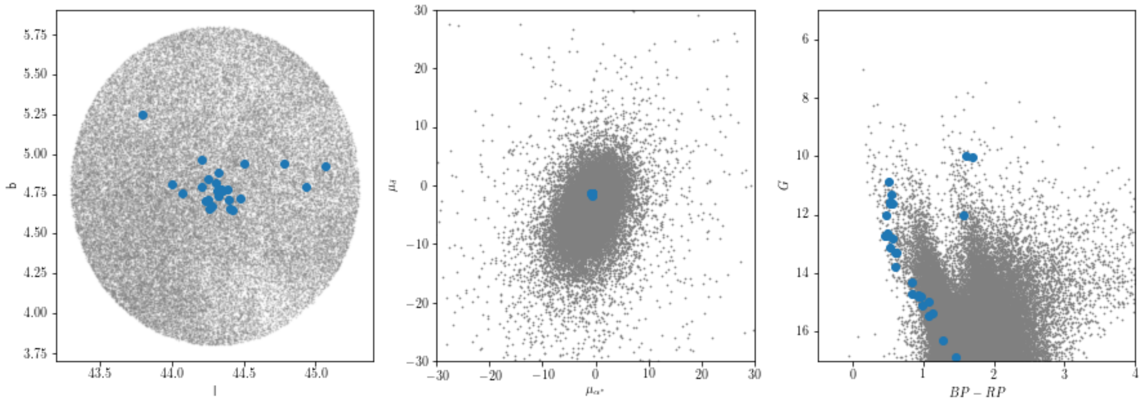

Fig. A.3. As in Fig. A.1 but for UBC3.

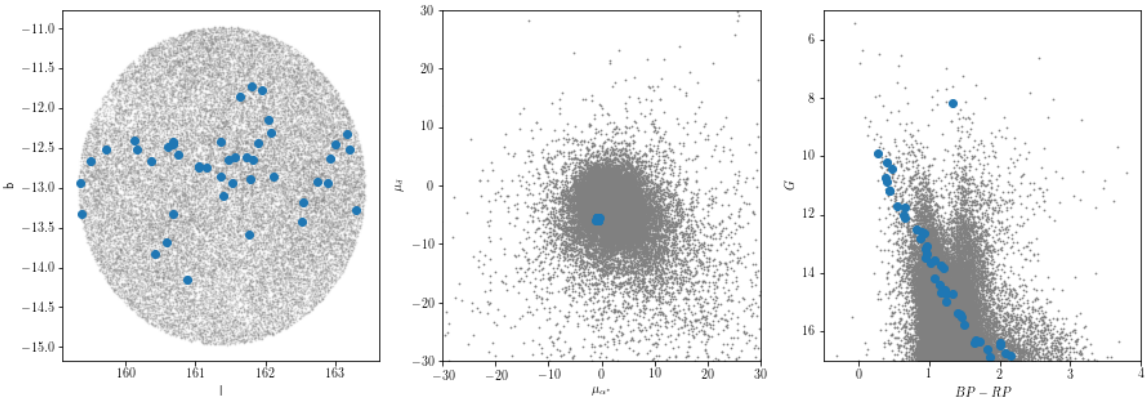

Fig. A.4. As in Fig. A.1 but for UBC4.

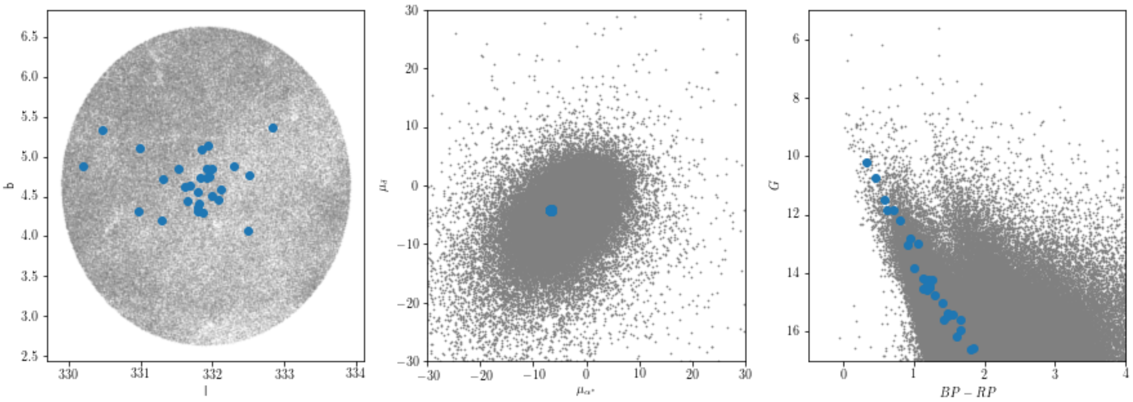

Fig. A.5. As in Fig. A.1 but for UBC5.

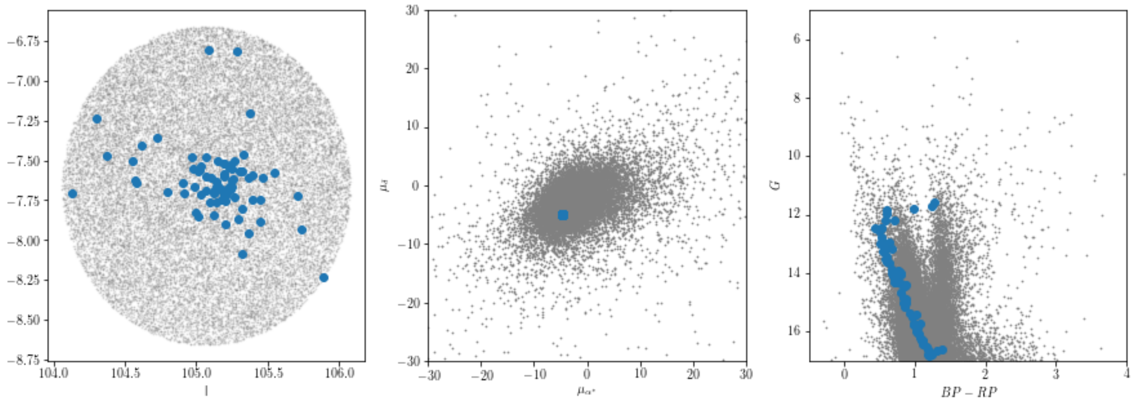

Fig. A.6. As in Fig. A.1 but for UBC6.

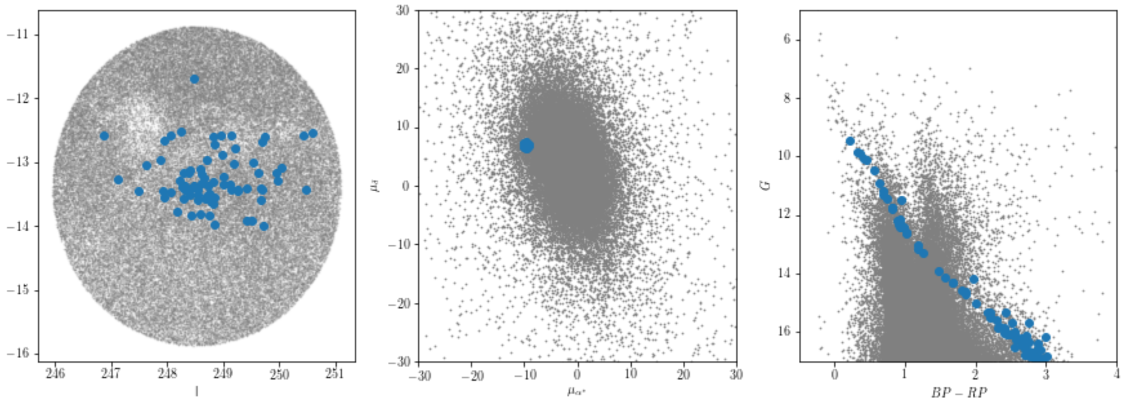

Fig. A.7. As in Fig. A.1 but for UBC7.

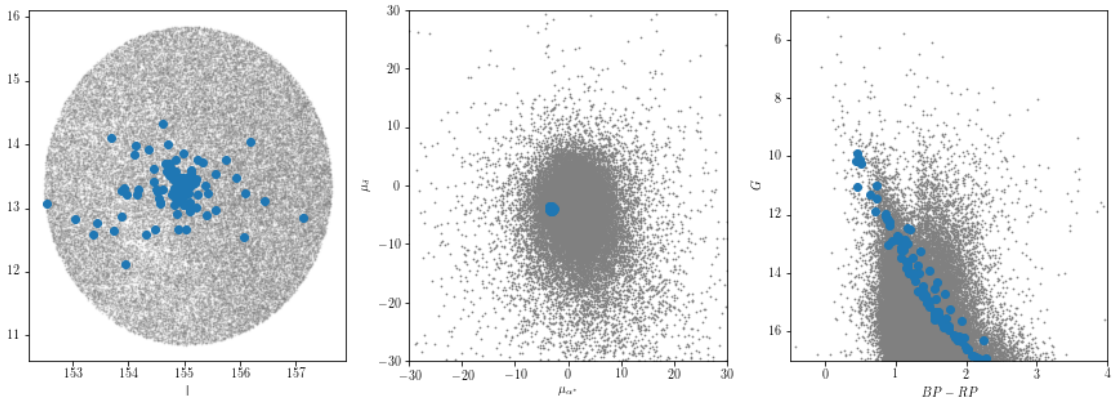

Fig. A.8. As in Fig. A.1 but for UBC8.

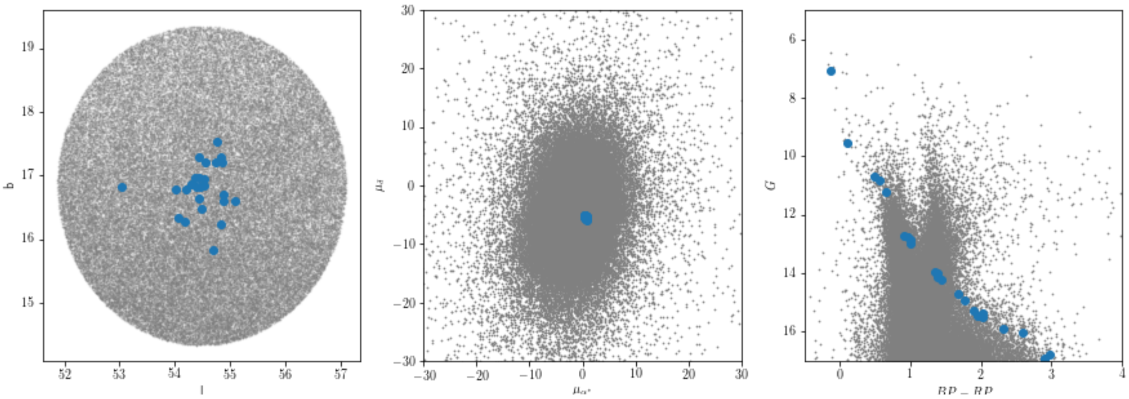

Fig. A.9. As in Fig. A.1 but for UBC9.

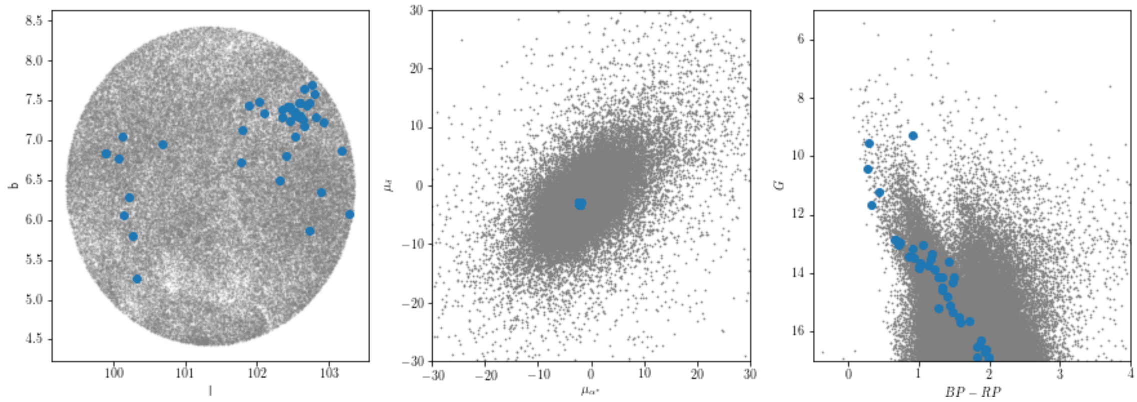

Fig. A.10. As in Fig. A.1 but for UBC10a.

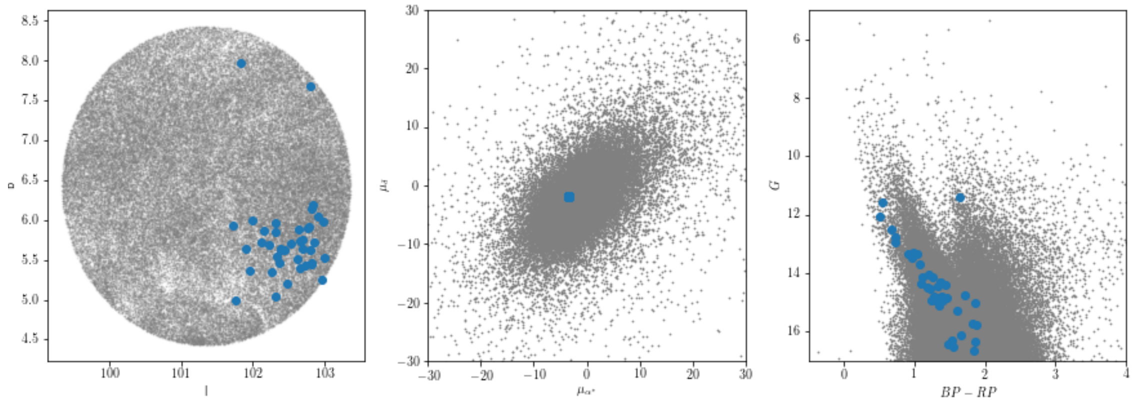

Fig. A.11. As in Fig. A.1 but for UBC10b.

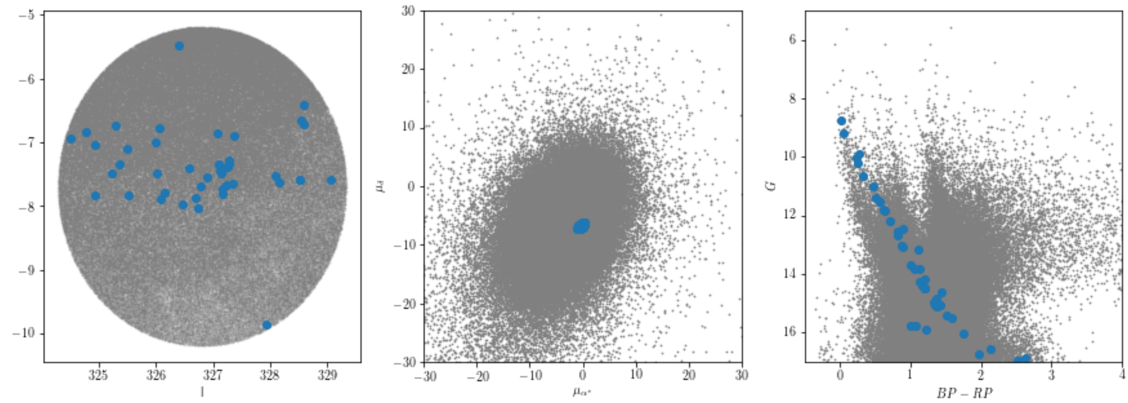

Fig. A.12. As in Fig. A.1 but for UBC11.

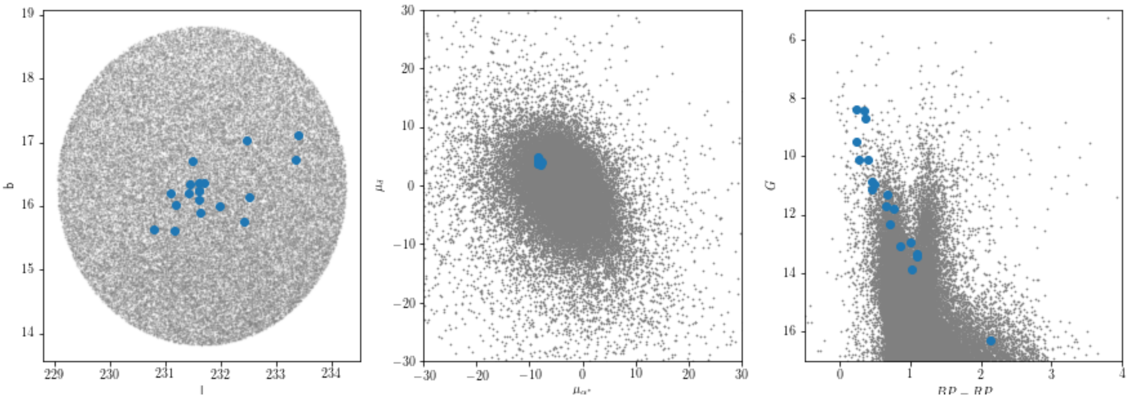

Fig. A.13. As in Fig. A.1 but for UBC12.

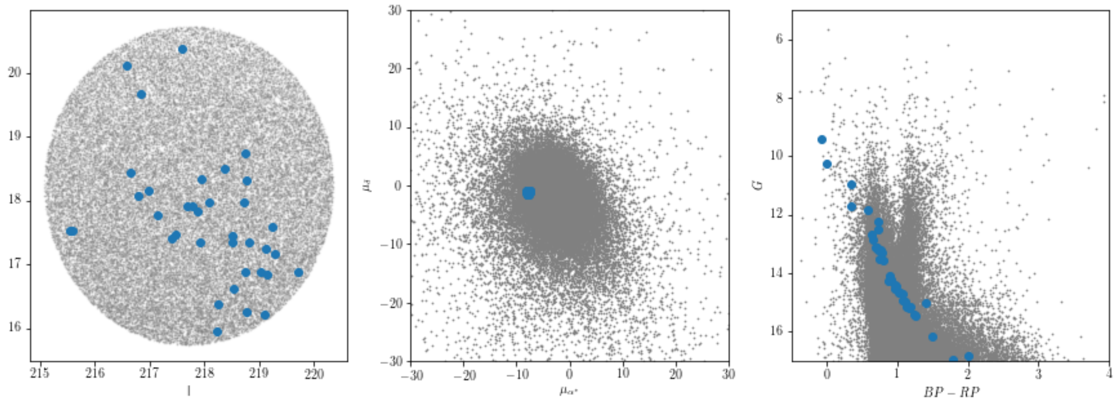

Fig. A.14. As in Fig. A.1 but for UBC13.

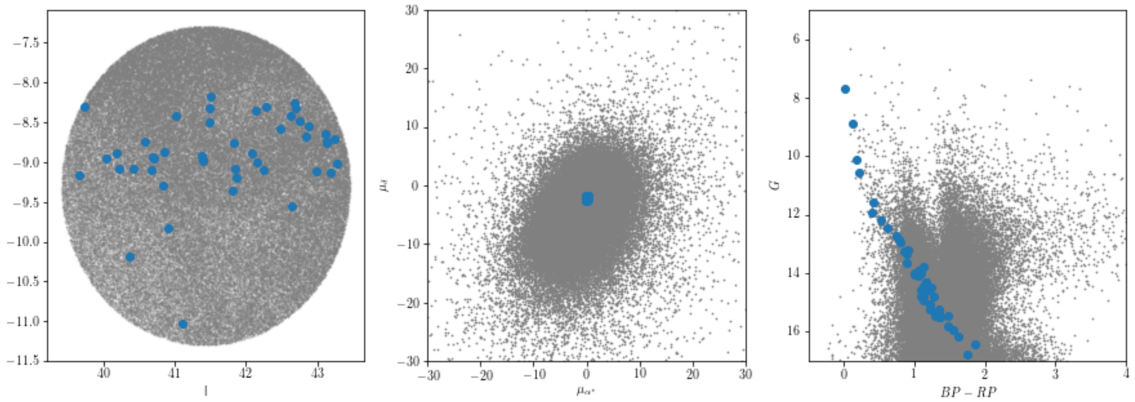

Fig. A.15. As in Fig. A.1 but for UBC14.

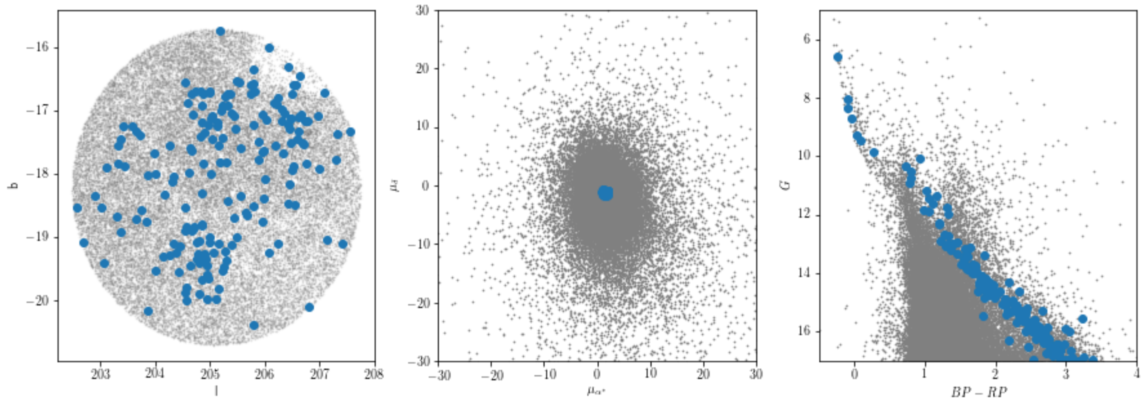

Fig. A.16. As in Fig. A.1 but for UBC17a.

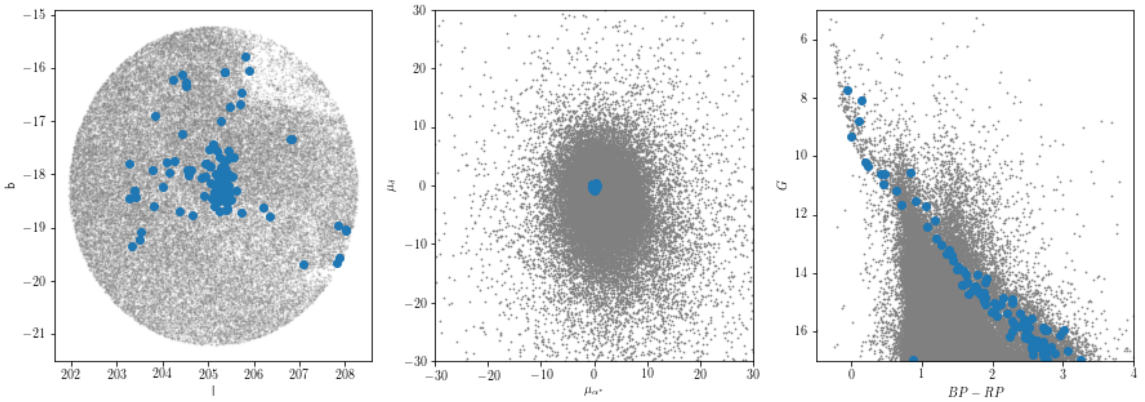

Fig. A.17. As in Fig. A.1 but for UBC17b.

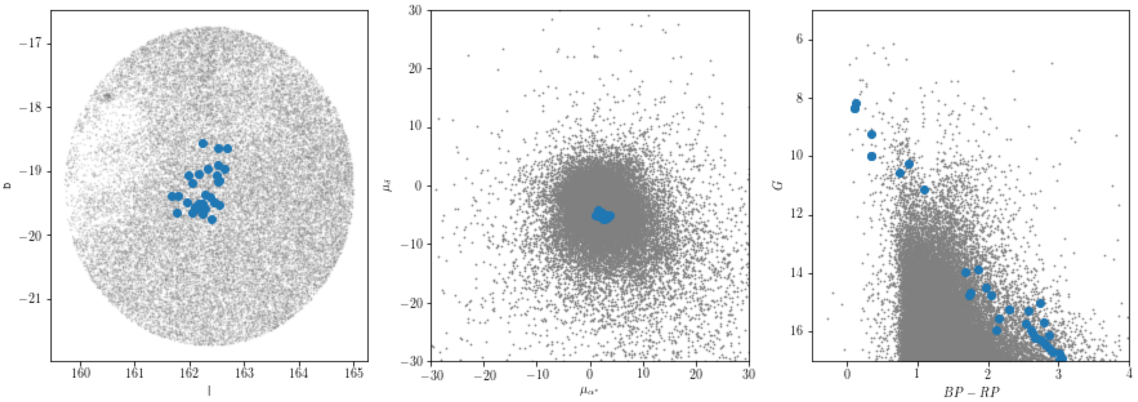

Fig. A.18. As in Fig. A.1 but for UBC19.

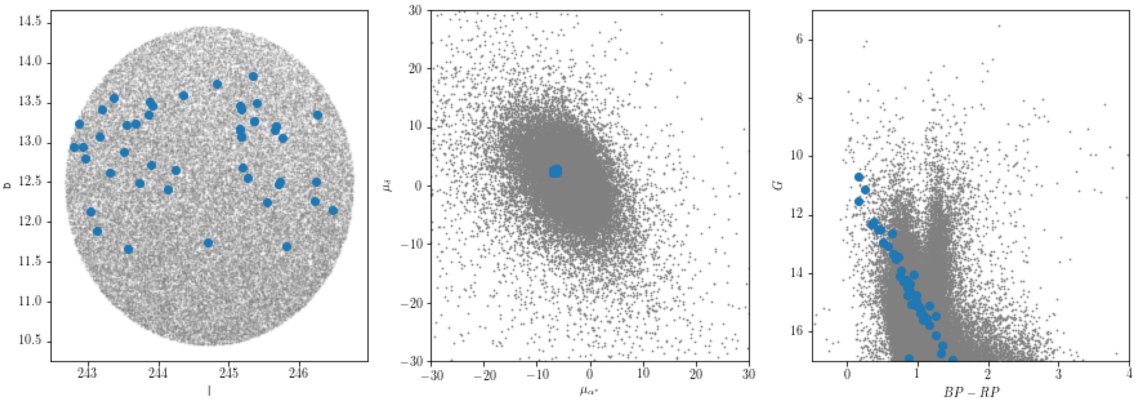

Fig. A.19. As in Fig. A.1 but for UBC21.

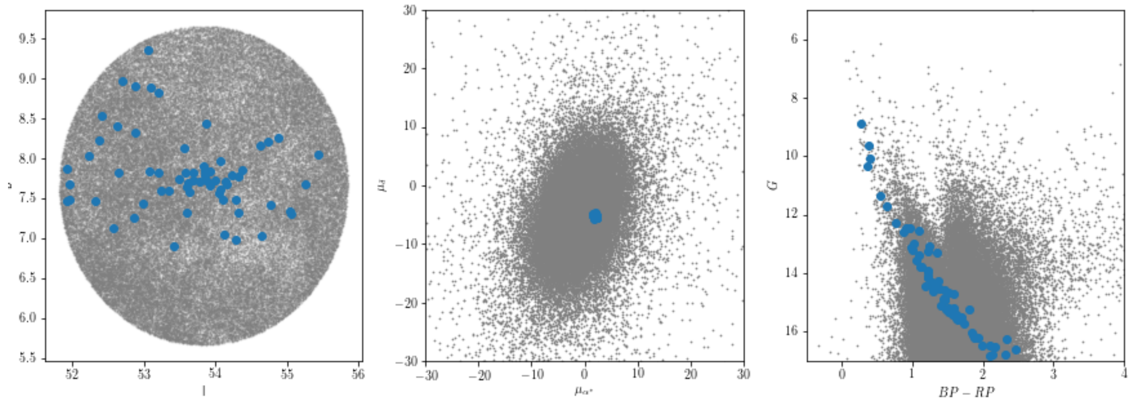

Fig. A.20. As in Fig. A.1 but for UBC26.

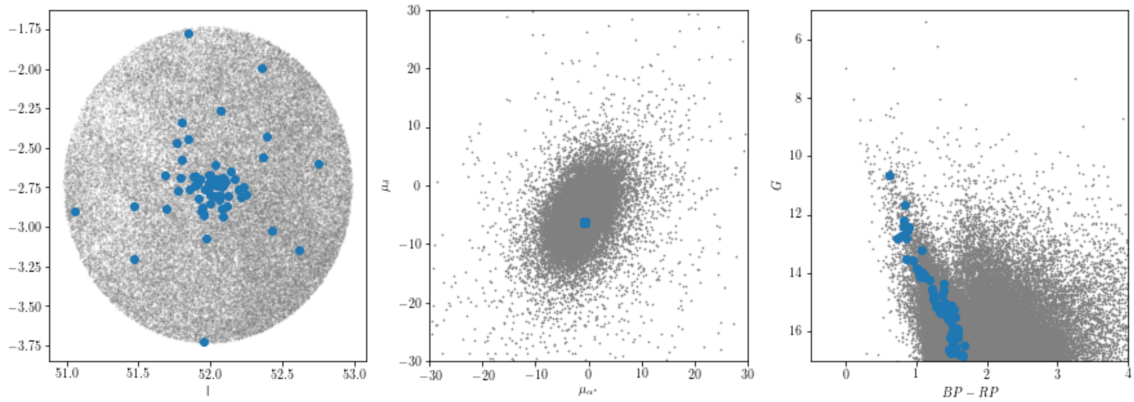

Fig. A.21. As in Fig. A.1 but for UBC27.

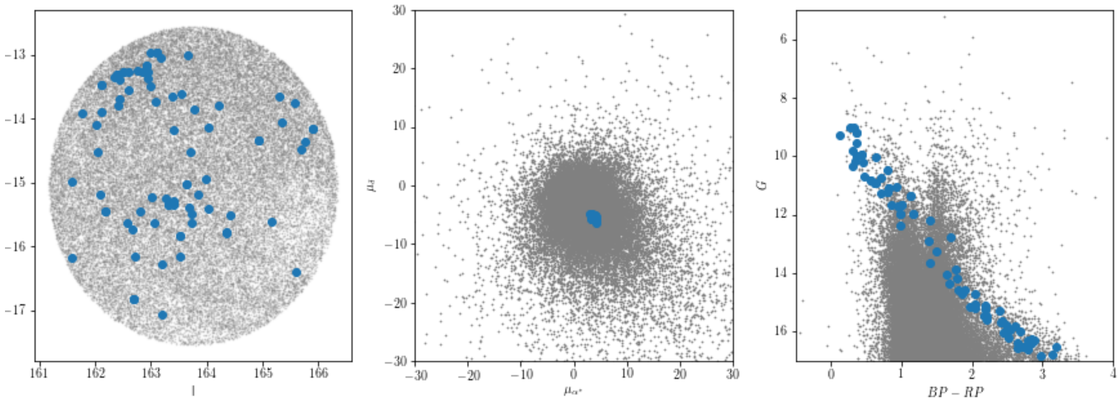

Fig. A.22. As in Fig. A.1 but for UBC31.

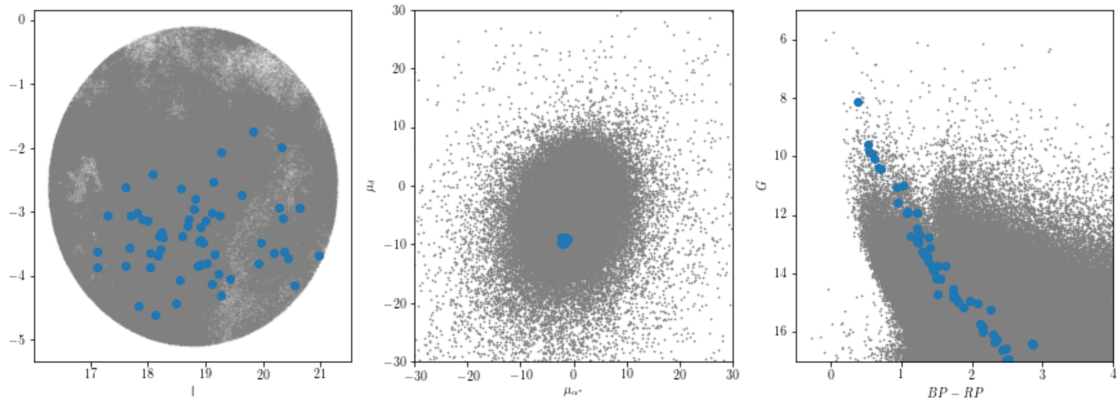

Fig. A.23. As in Fig. A.1 but for UBC32.