MNRAS 000, 1–13 (2023) Preprint 8 August 2023 Compiled using MNRAS LATEX style file v3.0

The Unified Cluster Catalogue: towards a comprehensive and homogeneous database of stellar clusters

Gabriel I. Perren,[1,3]★ María S. Pera,[2,3] Hugo D. Navone[2,3] and Rubén A. Vázquez[1,4]
[1]Instituto de Astrofísica de La Plata, IALP (CONICET-UNLP), 1900 La Plata, Argentina
[2]Instituto de Física de Rosario, IFIR (CONICET-UNR), 2000 Rosario, Argentina
[3]Facultad de Ciencias Exactas, Ingeniería y Agrimensura (UNR), 2000 Rosario, Argentina
[4]Facultad de Ciencias Astronómicas y Geofísicas (UNLP), 1900 La Plata, Argentina

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

We present the Unified Cluster Catalogue, the largest catalogue of stellar clusters with almost 14000 objects listed to date. In this initial version only Milky Way open clusters are present, but other objects will be included in the future. Each cluster is processed with a new probability membership algorithm designed to incorporate each star's coordinates, parallax, proper motions, and their uncertainties into the probability assignment process. We employ Gaia DR3 data up to a G magnitude of 20, resulting in more than a million probable members identified. The catalogue is accompanied by a public web site aimed at facilitating the search and data exploration of stellar clusters.

Key words: (Galaxy:) open clusters and associations: general – catalogues – methods: data analysis

1 INTRODUCTION

Open clusters (OCs) are groups of gravitationally bound coeval stars with a wide range of masses, that formed from the same molecular cloud. Their orbits generally locate them close to the Galactic disk, although there are a few examples of OCs with large vertical distances. Having originated from the same cloud, their member stars share a chemical composition and age, and are located spanning a somewhat compact position in space. The study of OCs is of fundamental importance for a number of key aspects in astrophysics research, including the process of stellar evolution as well as the dynamics, structure, formation, and chemical evolution of the Galaxy (Friel 1995).

Catalogues of OCs are crucial not only to organize these objects into publicly available systematic databases, but most importantly to assist researchers on the task of discovering new potential objects and, almost equally important, discard false detections. Catalogued bonafide OCs are also the most natural training dataset against which we can test and eventually improve new algorithms for stellar clusters detection. There have been efforts to provide catalogues of OCs to the astrophysical research community at least since the late 18th century, even if that was not the primary objective of the compilation. The first of such well known catalogues to include OCs is the Messier Catalogue (Messier 1774) with less than 30 OCs listed. This work was quickly followed by Herschel's Catalogue of One Thousand New Nebulae and Clusters of Stars (Herschel 1786), culminating a century later with Dreyer's New General Catalogue of Nebulae and Clusters of Stars (NGC, Dreyer 1888). The NGC lists less than 650 OCs, which is more than twenty times the number of objects present in Messier's catalogue.

After that first rapid growth, the pace with which OCs were discovered and catalogued slowed down. The next big catalogue, again published almost a century later, was the Base Données Amas (Mermilliod 1995) which listed a little over 1100 OCs, based on the previous compilation by Lynga (1987). This work is the foundation of the WEBDA catalogue,[1] a heavily used resource in the analysis of OCs which currently lists almost 1800 objects.

In the following decade, with the advent of large public databases containing from hundreds of thousands to millions of stars — such as Hipparcos and Tycho (Perryman et al. 1997; Høg et al. 1997) and 2MASS (Skrutskie et al. 2006) — the task of detecting new candidate OCs became substantially more attainable. This is particularly true for those objects that are faint, obscured by dust, or not in the vicinity of the solar system. Catalogues such as the Milky Way Star Clusters (Kharchenko et al. 2012) or those presented by Loktin & Popova (2017) or Bica et al. (2019) increased the number of listed OCs to more than 3000 in a few decades.

Finally in present time, the release of the database for the Gaia survey (Gaia Collaboration et al. 2016) with over a billion observed stars, translated into an enormous quantity of new candidate OCs reported in the literature. Clustering algorithms such as HDBSCAN (Campello et al. 2013) are used for the automatic identification of overdensities, improving dramatically the detection sensitivity compared to manual methods. The latest published catalogue is the one presented in Hunt & Reffert (2023) with ~7000 listed OCs, ~2000 of which are new candidates.

In this work we compiled a total of almost 14000 OCs, with about half of that number being new candidates discovered just in the past two years. As seen in Fig. 1 the growth in catalogued OCs in the last two centuries is almost exponential and, taking the last few years as

★ E-mail: gabrielperren@gmail.com

[1] https://webda.physics.muni.cz/webda.html

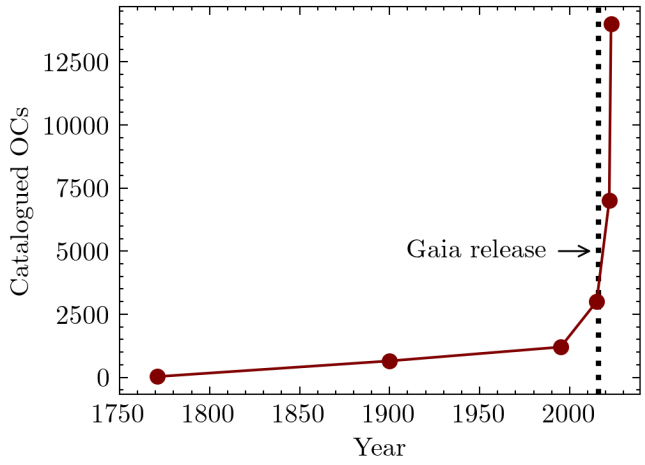

Figure 1. Approximate number of catalogued OCs in the literature since the late 1700s to the present day (including this work). The date of the release of the Gaia's survey data is marked as a dotted line.

a trend, it shows no sign of slowing down. Estimates for the total number of OCs in the Galaxy locate the upper limit over 10^5, which means that there remain still a lot of objects waiting to be found. The problem with these massive studies where new candidate OCs are presented by the thousands is simple: there are too many of them to be properly and individually analysed. Depending on the parameters used for the clustering algorithms employed, apparent groupings of non-physically related stars can easily disguise themselves as OCs. Furthermore, the scatter of information among dozens of articles and the influx of new articles published just months apart means that authors are often times not able to check their results against other catalogues; recent or not. This translates into duplication issues which, combined with the aforementioned problem, can have non trivial effects when these objects are used in massive studies (for example, when analysing the Galactic structure).

Our aim is to provide the community with a service that can be used to alleviate these issues. We combine — to the best of our knowledge— all the recent catalogues of OCs in the literature, and use them to generate a single unified catalogue of ~14000 OCs. We call this the Unified Cluster Catalogue (UCC hereafter), which can be accessed via the site `ucc.ar`. We apply over each listed OC a new tool to assign membership probabilities called `fastMP` ("fast Membership Probability").[2] The code was designed to work on the Gaia survey data, resulting so far in more than 1 million stars identified as probable members of the catalogued OCs. This gives us the largest homogeneously processed database of OC members to date. The data collected for each OC in the UCC catalogue, along with their estimated members, is publicly available in the aforementioned site. This site will be periodically updated as new candidate OCs are presented to the community.

This article is structured as follows. In Sect. 2 we introduce the stellar cluster databases employed to generate this initial version of the combined catalogue. Section 3 presents the new membership algorithm employed in the study of all the catalogued clusters. A comparison of our membership results with those recently published is performed in Sect. 4, along with the presentation of the public web site where these results will be hosted, updated, and expanded in the future. Finally, our conclusions are highlighted in Sect. 5.

[2] `fastMP`:`https://github.com/Gabriel-p/fastMP`

Table 1. Catalogues used in this article to generate the combined final catalogue with 13684 entries. Columns ID and N are the denomination used for new candidate OCs presented in those works, and the total number of OCs taken from them, respectively. Abbreviations used throughout this article for two of the works are also shown.

Reference	ID	N
Kharchenko et al. (2012)	–	2854
Loktin & Popova (2017)	LP	1050
Castro-Ginard et al. (2018)	UBC	23
Bica et al. (2019)	Bica	3555
Castro-Ginard et al. (2019)	UBC	53
Sim et al. (2019)	UPK	207
Liu & Pang (2019)	FoF	76
Ferreira et al. (2019)	UFMG	3
Castro-Ginard et al. (2020)	UBC	570
Ferreira et al. (2020)	UFMG	25
Cantat-Gaudin et al. (2020, CANTAT20)	–	2017
Hao et al. (2020)	HXWHB	16
Ferreira et al. (2021)	UFMG	34
He et al. (2021)	HXHWL	74
Dias et al. (2021)	–	1742
Hunt & Reffert (2021)	PHOC	41
Casado (2021)	Casado	20
Jaehnig et al. (2021)	XDOCC	11
Santos-Silva et al. (2021)	CMa	5
Tarricq et al. (2022)	–	467
Castro-Ginard et al. (2022)	UBC	628
He et al. (2022a)	CWNU	541
He et al. (2022b)	CWNU	836
Hao et al. (2022)	OC	703
Li et al. (2022)	LISC	61
He et al. (2023b)	CWNU	1656
Hunt & Reffert (2023, HUNT23)	HSC	6272
Qin et al. (2023)	OCSN	101
Li & Mao (2023)	LISC	35
Chi et al. (2023b)	CWWL	46
Chi et al. (2023a)	LISC-III	82
Chi et al. (2023c)	CWWDL	1179
Total OCs in combined catalogues		24983
Total unique OCs after cross-matching		13684

2 DATABASES

We have gathered, to the best of our knowledge, all the recent OCs listed in the literature up to present day. These were extracted from 32 databases that were published in the last 11 years. The total number of entries is 24983, cross-matched to a final catalogue of 13684 unique OCs. A few small databases were not added individually as they are included entirely in Hunt & Reffert (2023): Zari et al. (2018), Bastian (2019), Tian (2020), Qin et al. (2021), Anders et al. (2022), Casado & Hendy (2023). Table 1 lists these works along with the number of OCs that were left after extensive cleaning and sanitizing. To avoid cluttering the article we mention the details in Appendix A, and restrict this section to a general discussion of the process.

Unlike other works such as the recent HUNT23, we do not attempt to cross-match the OCs listed in different databases using positions and/or astrometry. Instead, we standardize the names of all the catalogued OCs and combine their parameters when present in multi-

ple catalogues. This name-based matching means that if an OC is listed with more than one denomination in a given database, this will be carried out and added to the rest of the catalogues where it is found. Although we do not attempt to find duplicate OCs via positions and/or astrometry cross-matching, we do use this data to flag the most obvious duplicate entries. For example, the CWWDL 14677 candidate OC presented in Chi et al. (2023c), has coordinates and astrometry values estimated as: RA=0.9526°, DEC=-30.002°, pmRA=4.222 mas yr^{-1}, pmDE=18.721 mas yr^{-1}, plx=2.61 mas. This is marked in the UCC as a duplicate of the well known Blanco 1 OC which has almost identical coordinates and astrometry values in the literature: RA=0.9149°, DEC=-29.958°, pmRA=4.215 mas yr^{-1}, pmDE=18.724 mas yr^{-1}, and plx=2.59 mas. There are many entries with similar issues. Note that although cases such as this one are flagged as possible duplicates, they are not removed from the final catalogue. The reasoning behind this is that the authors presented these objects as new findings, and we maintain this denomination. Even if the duplication is very clear, we believe that this decision can help future research in identifying structures that were wrongly tagged as unique OCs, and avoid repeating the same mistake. It can also aid in the detection and analysis of binary cluster systems. We will show in Sect. 4 how each OC is assigned a probability of being a duplicate of others in the catalogue, based on a simple set of rules. This classification, even if not rigorous, is a very reasonable first step in a deeper analysis about duplication of candidates OCs in the literature.

In Fig. 2 we show the distribution of the 13684 OC candidates catalogued in the UCC so far, segregated by distance range. These distances are estimated as the inverse of the listed parallaxes. It is worth noting that many of the OCs in older catalogues, such as Kharchenko et al. (2012) and Bica et al. (2019), have no associated astrometry but can have distances estimated as part of the fundamental parameters analysis process (usually along with age and extinction). As seen in the top left plot, these catalogued OCs with no astrometry represent almost a quarter of the database. Expectedly, those OCs catalogued closer than ~1.5 kpc (right top plot) are listed in many more databases than the rest (as shown by the proportional sizes of the circles).

The UCC includes candidate OCs found through analysis of infrared photometry such as the FSR (Froebrich et al. 2007), Ryu (Ryu & Lee 2018), and VVV (Barbá et al. 2015) objects. These have not been included in any of the recent large scale analysis like CANTAT20 and HUNT23 because they are considered to be too obscure for Gaia photometry. A large portion of the OCs with no astrometry seen in the top left panel of Fig. 2 are precisely these objects. Our method, as we will show in Sect 3, does not depend on the ability of a clustering algorithm to detect a faint and small overdensity, which it will most likely not be able to accomplish. Thus, we can include these OCs here and report a first approximation of their mean positions in proper motions and parallax.

In this initial version of the UCC we did not include information on candidate OCs identified to be probable asterisms, as that presented for example in Cantat-Gaudin & Anders (2020). This data is much more scattered and difficult to collect but we plan on including it in future updates to the catalogue. This fact notwithstanding, we do assign two different quality parameters to each object which are very useful in the process of identifying probable non-clusters. This will be discussed in more detail in Sect 4.

Recently Kounkel et al. (2020) presented a list of more than 8000 moving groups. Many of these are very extended and small groups which can hardly be classified as OCs. For this initial version of the UCC we remove these groups from the catalogues where they are included, i.e. He et al. (2022b) and HUNT23.

While we were preparing this manuscript a new work by He et al. was presented with ~2000 new candidate OCs, whose associated data is still not fully available He et al. (2023a). Although it is not included in the initial version of the UCC presented here, it will most likely already be included by the time this article is published (assuming the data is eventually made public).

Finally, we made use of the latest release of the Gaia survey data (DR3) (Gaia Collaboration et al. 2022; Babusiaux et al. 2022) imposing a single cut on a maximum magnitude of G=20 mag. No other filters were applied on this data which was employed, as we will detail in Sect. 3, to process the entire catalogue extracting the most likely members for each listed OC.

3 MEMBERSHIP METHOD

Once the unified catalogue is generated, our next objective is to compile an homogeneous database of likely member stars for each OC in the UCC. Usually a handful of tools are employed for this task, referred to as clustering algorithms. Three of the most often used tools in the stellar cluster literature are Friends-of-Friends (Huchra & Geller 1982, FoF), DBSCAN (Ester et al. 1996) and HDBSCAN. Examples of recent articles mentioned in Table 1 employing these algorithms are Liu & Pang (2019), He et al. (2023b), and HUNT23; for FoF, DBSCAN, and HDBSCAN respectively. Even more specialized tools such as UPMASK (Krone-Martins & Moitinho 2014) or our own recently developed pyUPMASK (Pera et al. 2021, a generalized Python-based version of UPMASK) also depend at their cores on these clustering algorithms (k-Means in the case of UPMASK; pyUPMASK is able to work with about a dozen different such methods).

Although these tools have been used many times in the literature, we believe they are afflicted by some important shortcomings that need to be addressed. The first one is the processing time. For such a large amount of data as the one we are handling, it is very convenient that the codes are efficient. For example, it took HUNT23 eight days of runtime on 48 CPU cores to process the Gaia DR3 database using HDBSCAN. This large requirements can easily become an obstacle in the analysis. Second, and tightly related to the first problem mentioned previously, these algorithms do not take uncertainties into account. One could of course incorporate the uncertainties associated to the input data through some type of bootstrapping mechanism (Efron 1979), but again the first problem arises. If a single processing run of the data is markedly time consuming, thousands of runs are virtually impossible. Finally, the definition of what exactly constitutes a cluster is glossed over in most (if not all) of the works that deal with stellar membership estimation. This is mainly because there is no standard definition across different clustering algorithms, and each one employs a different one. Furthermore, when applying these algorithms, the selection of cluster members depends on a number of parameters unique to the method being used. The values for these parameters are not trivial to set, and their choices can not be easily fundamented other than because they give "reasonable" results.

The ideal scenario is one where we are given a database that contains for each star both its mass and its coordinates in the full phase space, i.e.: three positional and three momentum variables. This would allow us to define a cluster in a way that is physically

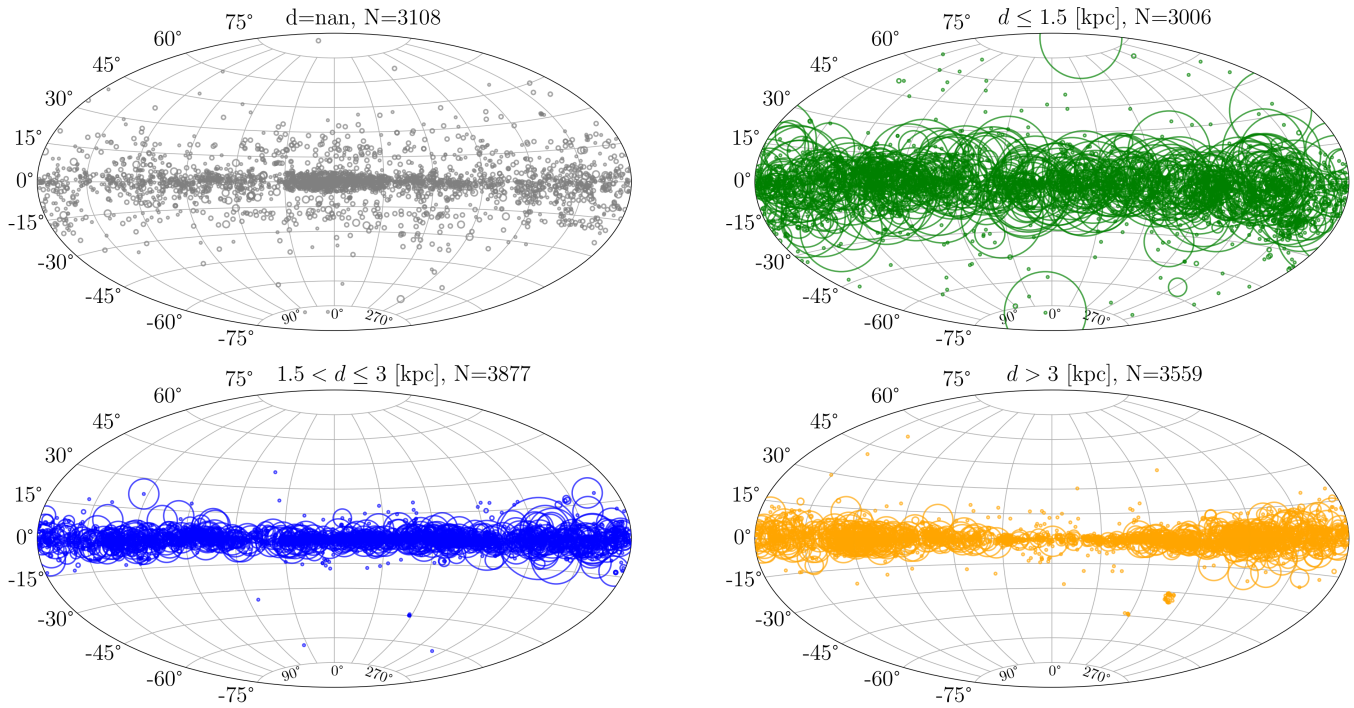

Figure 2. Map of the full list of catalogued OCs in this work, in galactic coordinates, by range of catalogued distance. From top left to bottom right: OCs catalogued with no distance (grey), OCs with $d \leq 1.5$ kpc (green), OCs in the range $1.5 < d \leq 3$ kpc (blue), and OCs with $d > 3$ kpc (orange). Sizes are proportional to the number of databases from Table 1 that include the candidate OC.

proper, taking the gravitational potential of the Galaxy into account. Of course, this is not the case. We do not have available information about masses and even the phase space is not complete, since we lack radial velocities in the large majority of the cases[3]. What we have for each observed stars is thus a 5-dimensional data vector made up of coordinates (equatorial, galactic), parallax, and proper motions. With this at our disposal, we propose the following definition of cluster:

Given an integer value $m > 0$ and a point c in an n-dimensional space, a "cluster" is defined as the collection of m elements with the smallest n-dimensional Euclidean distance to c.

The advantage of explicitly stating a definition of a cluster is that we no longer depend on different clustering methods or their extraneous parameters. Assuming the centre and the estimated number of members are given, this definition will always yield the same set of selected stars as probable members; approximately at least, since uncertainties do play a role here as we will see below.

Taking into account the aforementioned shortcomings of general clustering algorithms, we decided to develop a new tool to estimate membership probabilities called `fastMP` (acronym for *fast Membership Probabilities*), based on the above definition of a cluster. We named the code `fastMP` due to its processing speed. In our tests, it takes less than 3 seconds to analyse an average OC in an 11 years old 4-cores CPU. This means that the entire UCC — that is almost 14000 OCs so far — can be processed in approximately 11 hours on a very modest CPU; which includes the time required to incorporate the uncertainties into the process. In Fig. 3 we show the arrangement of the basic blocks in the `fastMP` algorithm. As can be seen, it is a rather simple process that mainly depends on the two basic parameters mentioned in the definition of cluster: its centre and the number of members.

We briefly describe each block in the following subsections, noting that the code is fully open source and released with a GPL v3 general public license[4], meaning that it can be easily tested, modified, and audited.

3.1 Centre estimation

To determine the most likely centre c for the OC under analysis, the code uses a three step process. First, it searches for the region of maximum density in proper motions. Only the 2-dimensional space of proper motions is employed here, since this is generally where the overdensity is most visible and stands out against the surrounding contaminating stars in the observed field. The position of the overdensity is found in an iterative process that starts with the full set of data, and gradually "zooms in" until a convergence criteria is reached. The second step selects a subset of stars with the closest distance to this centre coordinates in proper motions space. In the third step, the final centre value is obtained using a k-nearest neighbours algorithm to select the point with the largest density in the 5-dimensional space (coordinates, parallax, proper motions).

[3] Currently Gaia contains radial velocity data for less than 2% of the observed stars, see: https://www.cosmos.esa.int/web/gaia/dr3.

[4] https://www.gnu.org/copyleft/gpl.html

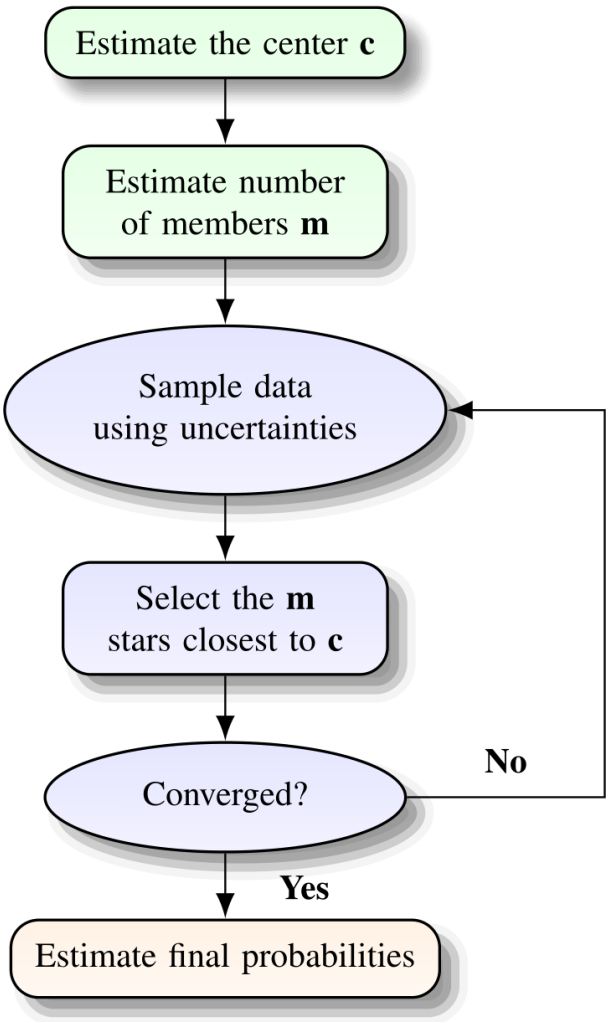

Figure 3. Basic flow chart of the algorithm used by our `fastMP` membership estimation tool.

3.2 Number of members estimation

Once centre point is estimated, the total number m of stars that can be considered members of the OC is obtained through Ripley's K function (Ripley 1976, 1979). This function is used to asses how close a group of points is to a random uniform distribution. We refer the reader to our previous article where we presented pyUPMASK (Pera et al. 2021), where we introduced the concept in much more detail. Subsets of stars are selected in rings, moving outward from the centre values estimated in proper motions and parallax. If these stars are considered to be far enough from a random uniform distribution in coordinates space, they are kept as probable members of the OC. In this block we can also reject stars that are more likely to belong to other clusters in the frame, to more accurately estimate the true number of members for the OC under analysis.

Since it is run only once, Ripley's K function can be replaced by any other method to estimate the size of a cluster (for example some of the already mentioned clustering algorithms) without much impact on the performance. It can even be skipped entirely by feeding this number to `fastMP`, estimated manually or by some external process, as explained in the final paragraph of this section.

3.3 Sampling, selection, convergence

This is the bootstrap block where we incorporate the data uncertainties into the final membership probabilities. Its basic function is to sample the observed data using its uncertainties, select those m stars closest to the centre c, and finally repeat this process until a convergence criterion is reached. We define as a stopping condition the run when the total number of stars with probability greater than 50% has stabilized for several of runs. The bootstrap block is run usually a few hundred times before convergence is achieved.

3.4 Membership probabilities

To estimate the membership probability for each star, the algorithm simply divides the number of times a given star was selected in the bootstrap block by the number of runs required until convergence. In the case where no star was selected by the bootstrap block, probabilities are assigned based on the 5-dimensional distances to the centre. These are lower quality probabilities since they do not make use of the uncertainties of the data. This last resort option is only used for candidate OCs that are either too faint to be perceptible, or for objects that simply have no detectable stars in a region of density large enough to be considered stellar clusters.

Finally, it is worth noting that an important advantage of `fastMP` over tools commonly used like UPMASK, pyUPMASK, HDBSCAN, etc., is that it can run in supervised mode. In this context, we make the distinction of supervised versus unsupervised in the sense that the algorithms mentioned above work with no prior information about the cluster(s) being analysed; we call this "unsupervised". In contrast, `fastMP` allows information about the cluster to be passed along with the input data; we call this "supervised". Such information can be the centre of the cluster, its total number of members, or both. If any of these values are fed to the code after estimating them manually or via an external method, then its corresponding block (either centre or number of members estimation)is skipped. This is of great help, particularly for OCs that are very faint or sparse and can not easily be picked up by the usual clustering algorithms. It is also a feature that allows the code to analyse systems that are very close together, either in the positional space or a combination of this and the proper motions and/or parallax dimensions, by fixing its centre and/or number of constituent members. In the following section we show that `fastMP` has an excellent performance when compared to recent works that generate lists of members for OCs, like CANTAT20 and HUNT23.

4 RESULTS

This section is divided as follows. In Sect. 4.1 we analyse the issue of duplicated entries across catalogues. Sect. 4.2 compares the results of our membership estimation with those from recent large catalogues. Sect. 4.3 discusses a possible classification of the candidate OCs as real physical objects to separate them from artefacts derived from the application of different clustering algorithms. Finally, Sect.4.4 presents a brief overview of the online service where this catalogue is hosted and a few of the issues that will be improved in the future.

4.1 Duplicates

One of the main problems with today's state of research in the area of OCs is, as mentioned earlier, the large number of articles being pub-

lished on the subject. This should not be a concern a priori, but with articles appearing every few months presenting new candidates by the thousands, keeping track of the latest proposed objects becomes a not so simple task. This is evidenced by the large number of potential duplications that can be found when cross-matching the most recent databases with older ones.

In this work we do not attempt to merge and/or discard candidates as duplicates, as this is not a trivial assessment to make. The UCC only flags OCs that have the potential of being duplicates of others, following a parallax-based decision rule. This rule checks the distance from a given OC to all the others in the catalogue in three separate components: coordinates (arcmin), parallax (mas), and proper motions (mas yr^{-1}). If these distances are smaller than a given threshold, they are converted into a probability using a linear relation; else they are assigned a probability of zero.[5] The relations depend on the distance to the OC (estimated from its catalogued parallax) and can be seen in the block below.

```
if parallax >= 4
    xy_r, plx_r, pm_r = 20, 0.5, 1
else 3 <= parallax < 4
    xy_r, plx_r, pm_r = 15, 0.25, 0.75
else 2 <= parallax < 3
    xy_r, plx_r, pm_r = 10, 0.2, 0.5
else 1.5 <= parallax < 2
    xy_r, plx_r, pm_r = 7.5, 0.15, 0.35
else 1 <= parallax < 1.5
    xy_r, plx_r, pm_r = 5, 0.1, 0.25
else .5 <= parallax < 1
    xy_r, plx_r, pm_r = 2.5, 0.075, 0.2
else parallax < .5
    xy_r, plx_r, pm_r = 2, 0.05, 0.15
else parallax < .25
    xy_r, plx_r, pm_r = 1.5, 0.025, 0.1
else parallax is nan
    xy_r, pm_r = 2.5, 0.2
```

Here, `xy_r`, `plx_r` and `pm_r` are the parallax-based thresholds for each component (in arcmin, mas, and mas yr^{-1}; respectively). For example, if a candidate OC has a catalogued parallax of 0.75 mas, then `xy_r, plx_r, pm_r = 2.5, 0.075, 0.2`. This means that any other OC with distances smaller than those thresholds in any (or all) of the components, will have a non zero probability of being a duplicate; where a larger probability is associated to smaller distances. The reason to split the threshold in parallax ranges is that the farther away the OC the smaller its parallax, coordinates radius, and mean proper motions will tend to be. These thresholds and parallax ranges are of course entirely arbitrary, but we have observed very reasonable results using them. Employing this rule, the aforementioned OC candidate CWDDL 14677 is flagged with a 65% probability of being a duplicate of Blanco 1; if the catalogued positions are used. If we use the values for the main coordinates, proper motions, and parallax obtained from the most likely members found by `fastMP`, this probability increases to 84%.

In Fig. 4 we show the percentage of entries flagged as probable duplicates for five of the latest articles mentioned in Table 1. We selected as probable duplicates those entries with an assigned probability larger than 50%. The CWWDL clusters from Chi et al. (2023c) stand out with almost 40% of its 1179 candidate OCs flagged as probable duplicates of entries in previous (older) catalogues. This is a rather large value that translates to more than 400 of the candidate OCs listed in this catalogue.

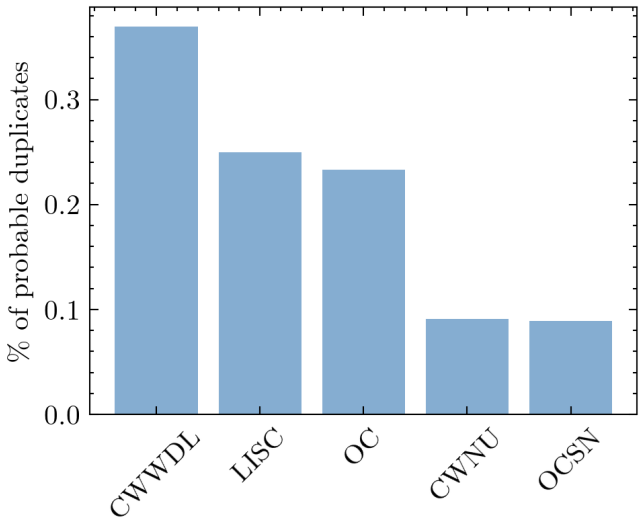

Figure 4. Percentage of probable duplicates for five recent databases of candidate OCs presented in the literature, characterized by their IDs. See Table 1 to match the corresponding article(s) to each ID.

Table 2. Examples of two pairs of OCs flagged as duplicates by our parallax-based rule with probabilities of 50% (1st and 2nd columns) and 90% (3rd and 4th columns).

	P≈50%		P≈90%	
	CWDDL 578	UBC 395	LISC 3279	UBC 361
RA	345.743	345.655	290.000	290.016
DEC	57.231	57.206	15.146	15.157
Plx	0.412	0.409	0.630	0.632
pmRA	-2.844	-2.869	-1.669	-1.701
pmDE	-2.542	-2.591	-5.225	-5.232

An example of two candidate OCs flagged as possible duplicates, with a probability value of 50% and 90%, are CWWDL 578 Chi et al. (2023c) and LISC 3279 (Li et al. 2022). These objects were associated by our parallax-based rule to two UBC OCs, presented in Castro-Ginard et al. (2020). The five dimensional positions catalogued values for these four entries are shown in Table 2. Although we understand that such a simple rule based on distances in the coordinates, parallax, and proper motion spaces is not a replacement for a proper detailed analysis on duplication, it can certainly serve as a reasonable starting point. Studies on systems of binary clusters can base their initial assessment on these probabilities as well. There are in total more than 2000 entries in the UUC flagged as being a duplicate, or flagged as having one or more duplicates, which amount to ≈15% of the entire catalogue.

[5] If the distances in all three components between the OC and another object are zero, then the probability of these two being duplicates of each other is 1. If all three distances are beyond the maximum limits shown in the parallax-based rules, then the probability of duplication is zero. Distance values beyond zero and these limits are converted linearly in the probability range (0, 1).

4.2 Membership analysis

More than one million estimated members are stored in this initial version of the UCC; this is almost half a million more than those in the HUNT23 catalogue (after removing globular cluster and moving groups), and approximately five times those in CANTAT20. This last database is expected to contain less entries not only because it is smaller regarding the number of OCs listed, but also because it only reaches G=18 mag, whereas UCC and HUNT23 go two magnitudes further down.

In the top plot of Fig. 5 we show the distribution of estimated members for these three catalogues versus Gaia's G magnitude. The centre plot shows the same distributions but normalized by the total number of entries in each catalogue. Finally, the bottom plot shows the percentage of members in CANTAT20 and HUNT23 matched to members estimated by the UCC. While CANTAT20 displays an overall match in the range ~75-80% for the entire magnitude span, up to its maximum of G=18 mag, the match with HUNT23 hovers around 70-75% up to G≈17 mag after which it begins to drop. For the largest magnitude in UCC and HUNT23, G=20 mag, the match percentage is ~35%. This can be explained mainly by two processes. In one hand, the large sensitivity of the HDBSCAN algorithm often causes it to return false positives, as reported by HUNT23. This can translate to an overconfidence in the number of members assigned to each candidate OC. On the other hand, the methods employed in CANTAT20 and HUNT23, UPMASK and HDBSCAN respectively, do not make use of the uncertainties of Gaia's data whereas `fastMP` does. Incorporating uncertainties into the membership probabilities estimation affects stars in the lower mass region the most, as these are the ones with the largest errors. The `fastMP` code also uses a more cautious approach when estimating the total number of stars associated to a given OC, based on Ripley's K function as mentioned in Sect. 3.2. The centre plot in Fig. 5 clearly shows these effects, where normalizing by the number of OCs in each catalogue makes the UCC's distribution dip below those of CANTAT20 and HUNT23. The UCC has a combined ~90 members per OC, HUNT23 has around ~110 members per OC, and CANTAT20 is close to ~95 members per OC.

4.3 Classification

Determining what constitutes a true physical cluster of stars is not an easy task when dealing with OCs. Whereas globular clusters are made of hundreds of thousands of member stars, OCs are much smaller. In the previous section we showed that, for the three largest catalogues published recently including the UCC, a reasonable general average for the number of members in an OC is of the order of just 100 stars; with a heavy skew towards smaller values. A proper physical analysis of whether a few dozen stars are gravitationally bound requires information that we currently lack, particularly the individual masses (or a reliable total mass estimation) and precise velocities.

This fact notwithstanding, there are still methods we can use to approximate a full dynamical analysis to characterize candidate OCs as more or less likely to be real. Two of these methods, or quality cuts, were proposed by Cantat-Gaudin & Anders (2020). The first one is based on the fact that we can set an upper limit to the internal velocity dispersion of OCs, beyond which objects should either be globular clusters or unbound groups. Conservatively, this limit is 5 km s^{-1}, but it is relaxed to 1 mas yr^{-1} for candidate OCs beyond ~1000 pc for which uncertainties in the proper motions tend to dominate the measured values. We set the break between both limits at ~1.5 kpc, the distance where the velocity dispersion lines mentioned

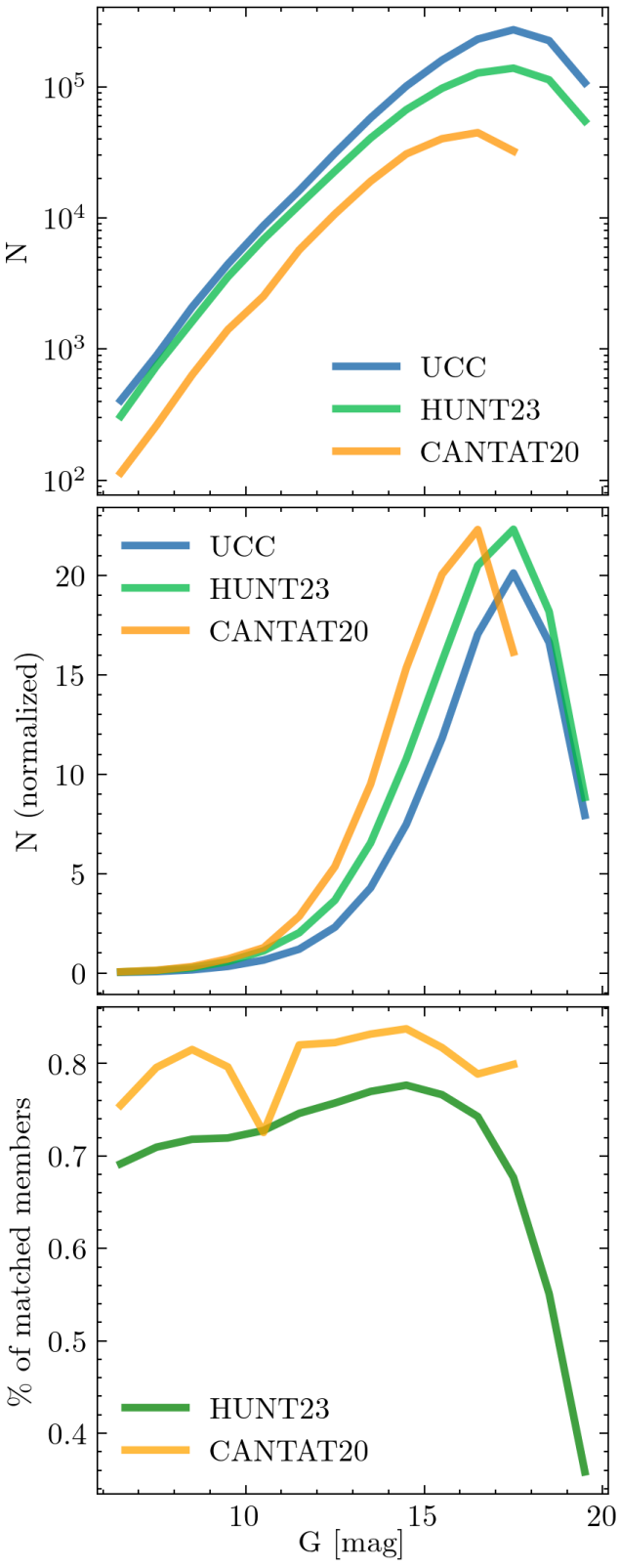

Figure 5. Top: total number of identified members versus magnitude, for the catalogues HUNT23, CANTAT20, and the UCC. Center: same as above but normalized by the total number of entries in each catalogue. Bottom: percentage of members in HUNT23 and CANTAT20 matched with members identified by the UCC, versus magnitude.

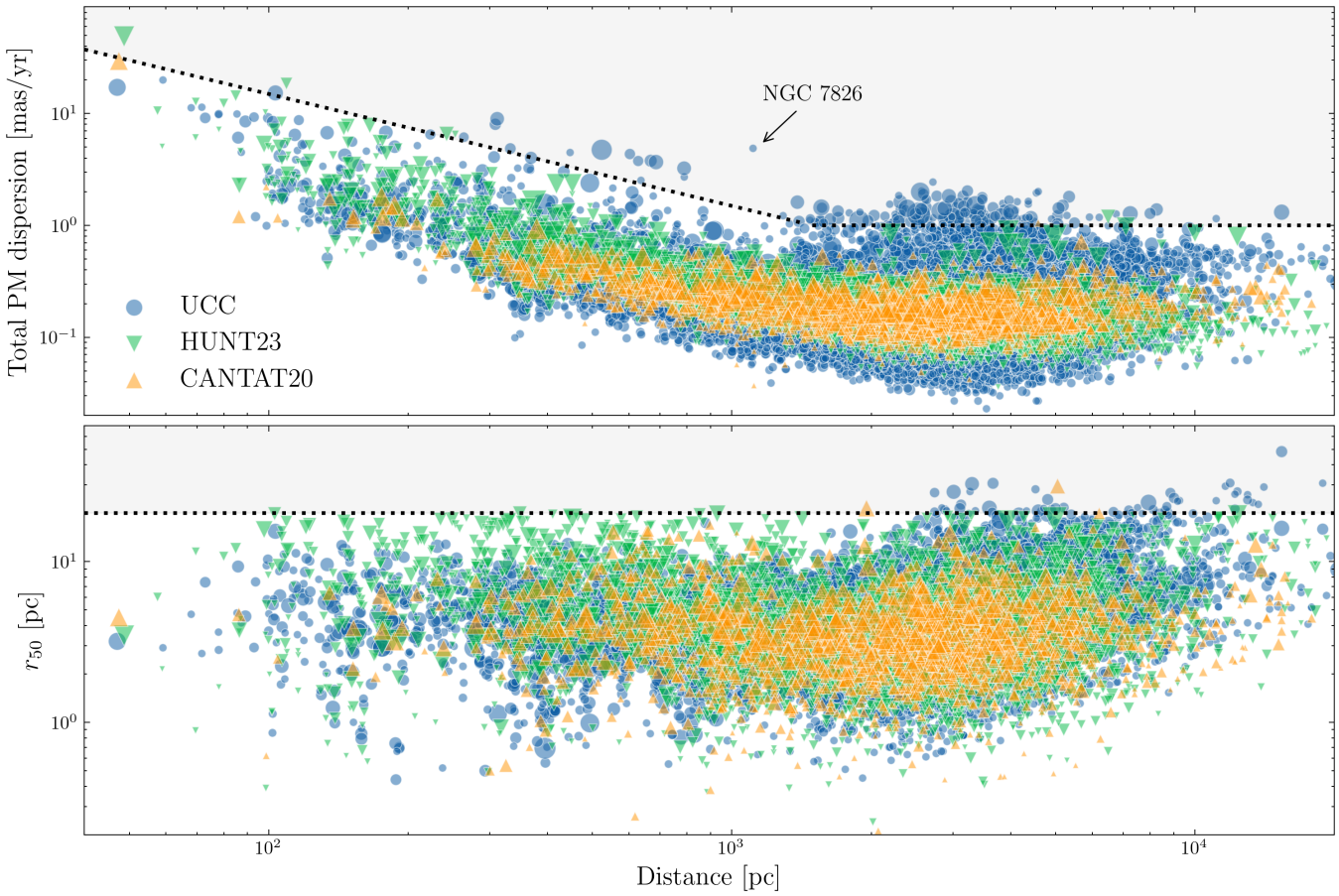

Figure 6. Distribution of total proper motion dispersion (top) and radius that contains half the members (bottom) versus distance, for the candidate OCs listed in CANTAT20 (orange), HUNT23 (green), and the UCC (blue). Sizes are proportional to each candidate's associated number of members. The grey region in both plots is the "not true OC region" defined by the rules proposed by Cantat-Gaudin & Anders (2020).

above intersect. The second method is a measure of the internal spatial dispersion of an OC. In Cantat-Gaudin & Anders (2020) it is stated that, for the majority of bonafide OCs, the maximum dimension that contains half the members can be set to ∼15 pc; HUNT23 relaxes this condition slightly to 20 pc, which we also adopt. In Fig. 6 we show the result of both these approaches to approximate a "true OC region" on the data collected in the UCC, along with that presented in CANTAT20 and HUNT23. As can be seen, the region occupied for the three catalogues is very similar with the UCC showing a broader distribution in total proper motions dispersion beyond ∼1000 pc. Nonetheless, the large majority of candidate OCs are well below the proposed cuts in both plots. Those objects with large proper motion dispersions are mostly candidates listed only in the Kharchenko et al. (2012) catalogue that were, to the best of our knowledge, never properly analysed. There are also ∼30 OCs from Ryu & Lee (2018) (infrared candidates) whose membership estimation is poor.

An example of an object that is located beyond the proper motions dispersion quality cut is NGC 7826, identified in the top plot of Fig. 6. This object has been discarded as a true OC in Kos et al. (2018) and listed as an asterism in Cantat-Gaudin & Anders (2020), but we include it nonetheless as it is listed in the Loktin & Popova (2017) catalogue. Expectedly, the proper motions distribution of its most likely associated stars also exclude it from the "true OC region" in this work.

We can also define other metrics to classify candidate OCs into more or less likely true physical objects, employing their estimated members' data. The first metric we developed is a density-based classification, C_{dens}, and the second one is a photometry-based classification, C_{phot}. These are similar in conception to the CST score and CMD class defined in HUNT23, respectively. The density-based metric compares the distribution of member stars to the distribution of close-by field stars, in the 5-dimensional space of coordinates, proper motions and parallax. The reasonable expectation is that neighbouring cluster members should have smaller average distances than neighbouring field stars. The photometry data is processed separately because member stars of an OC in a colour-magnitude diagram (CMD) do not cluster together around a centre value, as they do in the remaining data dimensions. Instead, they are distributed following an elongated path across the evolutionary sequence. For this reason we do not employ a closest-neighbour density based method, but one based on the likelihood of the members' sequence of being equivalent to a random sequence drawn from field stars. To quantify this we make use of the same function written for our `ASteCA` package (Perren et al. 2015), based on the Poissonian distribution likelihood defined in Tremmel et al. (2013). Fig 7 shows the distribution of these two metrics, both normalized in the [0, 1] range where 1 means most likely to be a collection of related cluster member stars in either metric. The colour is assigned according to the vertical distance to the quality cut in the total proper motions dispersion diagram shown in

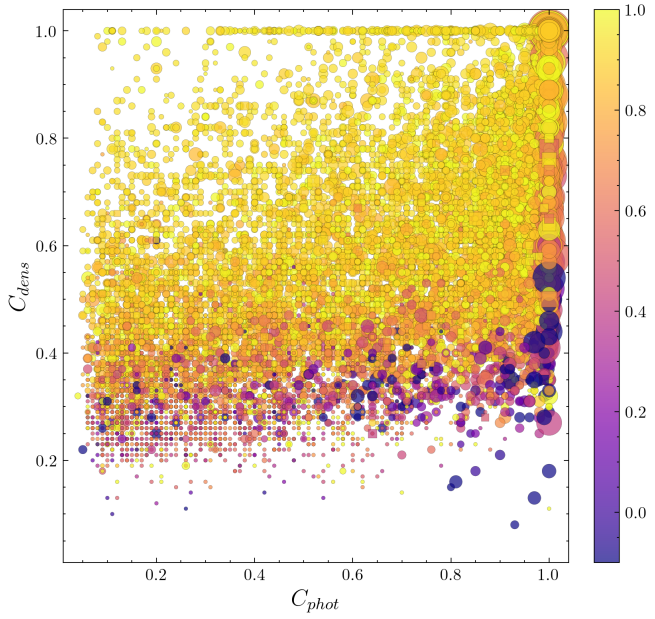

Figure 7. Five dimensional density C_{dens} versus photometry based metric C_{lkl} for the candidate OCs listed in the UCC. Colour is related to the vertical distance to the total proper motion quality cut; the colourbar is clipped at -0.1 and 1 to improve visibility. Size is related to the spatial extension of its members. Squares are associated to candidates beyond the 20 pc spatial dispersion quality cut.

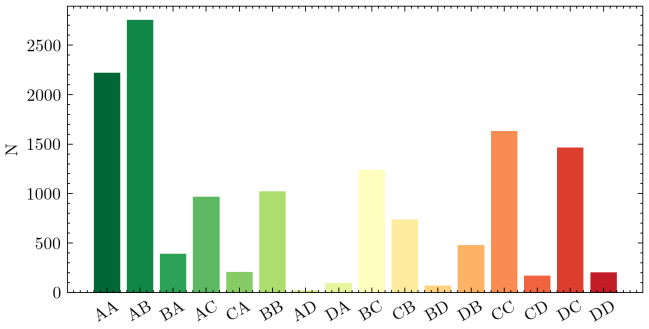

Figure 8. Distribution of the quality classes obtained combining the C_{phot} and C_{dens} values for each OC as described in the text.

the top plot of Fig. 6. Objects that are beyond these limit (i.e., larger proper motions dispersion than that allowed for an OC) have values below zero and are drawn in purple. Sizes follow the spatial extension of the estimated members. A clear tendency can be seen where candidate OCs with larger C_{dens} values also display large C_{phot} values, which is expected. There is also a visible scatter around the 1:1 identity relation, meaning these two metrics are not entirely correlated. This is desirable, otherwise both methods would be returning the same information. Objects with large total proper motion dispersions are mostly correlated with low C_{dens} values but tend to span the entire range of C_{phot} values.

We take these two classification metrics and combine them into a single quality class for each candidate OC, to provide a quick glance of the characteristics of the estimated members for the catalogued objects. First we split the [0, 1] range for both metrics into 4 equal length segments (from 0 to 0.25, from 0.25 to 0.5, etc.) and assign a letter to each from D for the [0, 0.25] segment, to A for the [0.75, 1] segment. These two letters, one for each metric, are combined to generate a single class out of 16 possible combinations. The letter that corresponds to the C_{phot} value is positioned first, followed by the letter obtained for the C_{dens} value for that object. The better quality clusters are thus assigned AA classes whereas the lowest quality ones DD classes. The complete distribution of classes is shown in Fig. 8. The three better quality classes (AA, AB, and BA), group more than 5300 objects (~40% of the catalogue), while the lowest three classes (CD, DC, and DD) contain almost 2000 candidate OCs or ~15% of the UCC catalogue. The classes AD and DA are two of the least populated, meaning that it is not likely to find an object with a large value in C_{phot} and a low value in C_{dens} or vice versa; as one would expect.

In Fig. 9 we show examples of four OCs listed in the UCC with combined classes of AA, BB, CC, and DD. The difference is clear between the better AA and worse DD classes, both in the more dense spatial, proper motions, and parallax distributions; as well as the sharper definition of the cluster sequence in the CMD.

These quality cuts and classification methods are unfortunately not enough to completely replace a proper analysis of a candidate OC. We can of course use them as guides to identify problematic cases, but their values still depend heavily on the precision of the membership estimation process. If this process did a poor job, which it will undoubtedly do for the more complicated cases, then these results will be compromised. More and better quality for the data of observed stars will hopefully improve this scenario in the near future.

4.4 Overview of the service

To facilitate cross-catalogues identification of entries, each candidate OC in the UCC is assigned a unique name following the International Astronomical Union's specifications. The naming convention is formatted as *UCC GLLL.LsBB.B*; where *UCC* is the catalogue's name, *G* indicates that we are using galactic coordinates, *LLL.L* is the (truncated) longitude for the object, *s* is the sign of its latitude, and *BB.B* is the (truncated) latitude. If two objects share coordinates, a lowercase letter from a to z is added to the end of the name.

The spatial position of each entry in the catalogue is displayed making use of Aladin's visualization tool,[6] showing the coordinates, proper motions, parallax, and radial velocity values associated to it, when present from the literature. Links to search the cluster's main name in the SAO/NASA Astrophysics Data System (ADS)[7] as well as a region search in the Strasbourg astronomical Data Center (CDS)[8] are provided. A Python notebook hosted in Google's Colaboratory service[9] is made available, to allow the user interactive exploration of the Gaia survey data of each candidate's estimated members. In this initial version the UCC only contains membership data obtained through our own `fastMP` tool. In future updates we will add the members estimated by other works, such as those from CANTAT20 and HUNT23, as well as the most recent catalogues.

Fundamental parameters such as distance, extinction, age, and metallicity are also shown when available, taken from as many databases as possible. This will also be expanded in future updates of the UCC, as more values from the literature are incorporated. For

[6] Aladin: `http://aladin.cds.unistra.fr/`
[7] SAO/NASA ADS: `https://ui.adsabs.harvard.edu/`
[8] CDS: `http://cdsportal.u-strasbg.fr/`
[9] Colaboratory: `https://colab.research.google.com/`

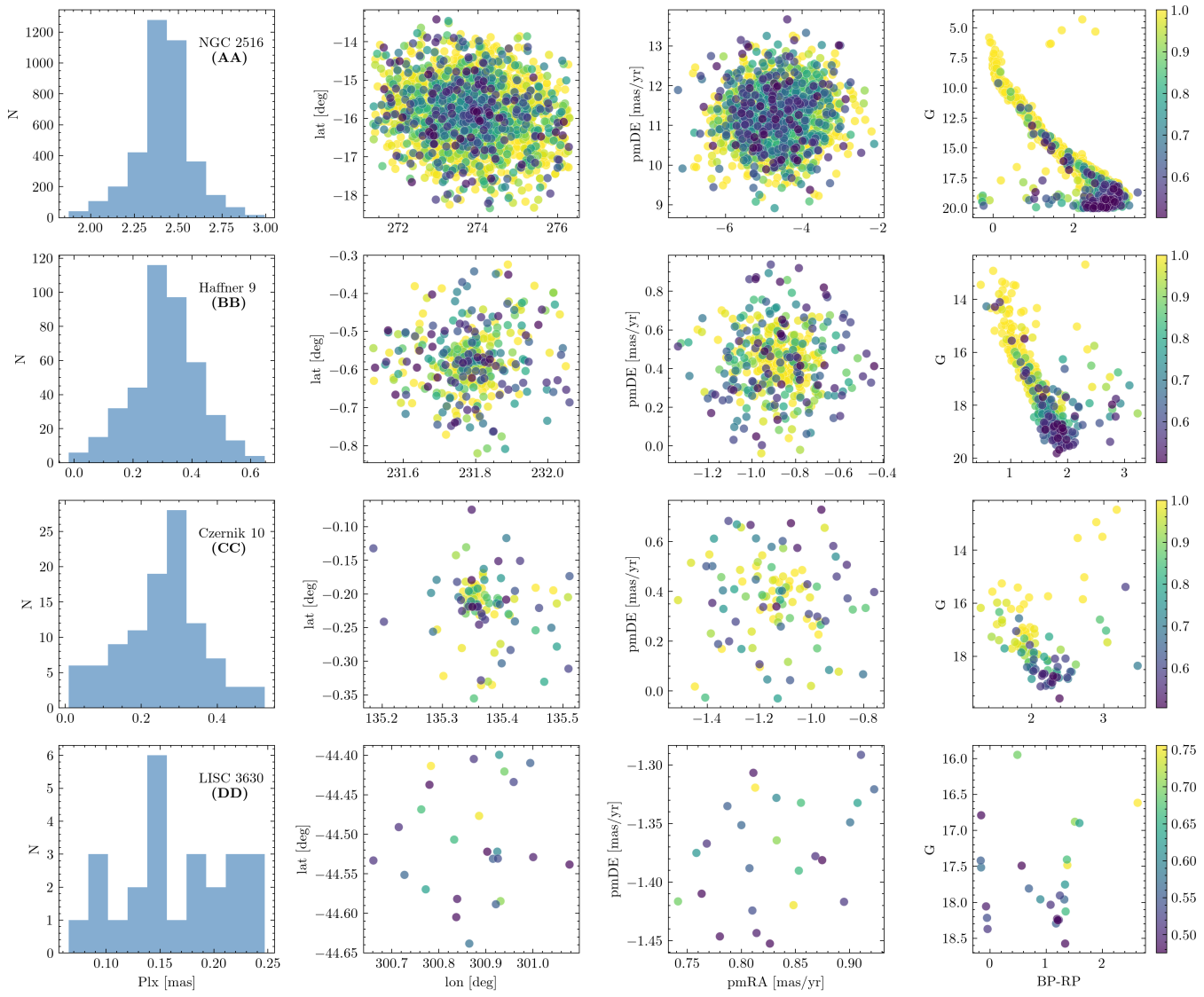

Figure 9. Examples of listed OCs with combined classes AA, BB, CC, and DD from top to bottom rows, respectively. Plots show from left to right: parallax distribution, spatial dispersion, proper motions, and CMD; respectively. Colourbars are associated to the membership probabilities assigned by `fastMP`.

each candidate OC its nearby entries and possible duplicates are also shown, along with the probability of being a duplicate obtained as detailed in Sect. 4.1.

In an upcoming version of the UCC we will add individual notes whenever available. These are much more difficult to acquire as they tend to be scattered throughout the literature instead of compiled in databases. For example Cantat-Gaudin & Anders (2020) contains in its appendix a list of OCs classified as asterisms; information that is lost if only the parameters of the candidate are displayed in the catalogue.

5 CONCLUSIONS

We presented the Unified Cluster Catalogue or UCC, along with its accompanying tool for membership probabilities estimation, the `fastMP` code. This is the largest catalogue of open clusters to date and it will be regularly updated as new databases are made public in the literature. The UCC is accessible through its own dedicated web site at `https://ucc.ar` where each OC is displayed along with its fundamental parameters (gathered from the literature when available), allowing the user to interactively explore the data online through Python notebooks hosted by the Google Colaboratory service. In its initial version the UCC lists almost 14000 unique candidate OCs whit a combined number of proposed homogeneously obtained member stars larger than 1 million, or ~90 member stars per OC.

Replacing the trained eye of a researcher even for the initial assessment of what constitutes a true OC in a completely generalized approach, is no easy task. With new candidate OCs being presented in the literature by the hundredths or thousands every few months, a systematic method for addressing this issue becomes more and more pressing. Our classification parameters were created to help with this task, but a close visual inspection is still required; particularly for the more complicated objects. We expect the Unified Custer Catalogue and its associated online service to be a useful resource for the astro-

physical research community, and welcome all suggestions to expand and/or improve it in the future.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Emily Hunt for her assistance with the processing of the HUNT23 database. This work has made use of data from the European Space Agency (ESA) mission *Gaia* (`https://www.cosmos.esa.int/gaia`), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, `https://www.cosmos.esa.int/web/gaia/dpac/consortium`). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement. This research has made use of the WEBDA database, operated at the Department of Theoretical Physics and Astrophysics of the Masaryk University. This research has made use of the VizieR catalog access tool, operated at CDS, Strasbourg, France (Ochsenbein et al. 2000). This research has made use of "Aladin sky atlas" developed at CDS, Strasbourg Observatory, France (Bonnarel et al. 2000; Boch & Fernique 2014; Baumann et al. 2022). This research has made use of NASA's Astrophysics Data System. This research made use of the Python language (van Rossum 1995) and the following packages: NumPy[10] (Van Der Walt et al. 2011); SciPy[11] (Jones et al. 2001); matplotlib[12] (Hunter et al. 2007); scikit-learn[13] (Pedregosa et al. 2011); `ASteCA`[14] (Perren et al. 2015). This work made use of Astropy:[15] a community-developed core Python package and an ecosystem of tools and resources for astronomy (Astropy Collaboration et al. 2013, 2018, 2022).

DATA AVAILABILITY

The data underlying this article are available in the repositories associated to the Unified Cluster Catalogue, accessible at `https://github.com/ucc23`. The code employed to process the data and generate the images in this article con be found in the repositories for the `fastMP` code at `https://github.com/Gabriel-p/fastMP`, and in the repository for the article itself at `https://github.com/gabriel-p-artcls/GDR3_members`. Any missing data file and/or code file can be requested to the corresponding author and we will gladly make it available.

REFERENCES

Anders F., Castro-Ginard A., Casado J., Jordi C., Balaguer-Núñez L., 2022, Research Notes of the AAS, 6, 58
Astropy Collaboration et al., 2013, A&A, 558, A33
Astropy Collaboration et al., 2018, AJ, 156, 123
Astropy Collaboration et al., 2022, apj, 935, 167
Babusiaux C., et al., 2022, arXiv e-prints, p. arXiv:2206.05989
Barbá R. H., et al., 2015, A&A, 581, A120
Bastian U., 2019, A&A, 630, L8
Baumann M., Boch T., Pineau F.-X., Fernique P., Bot C., Allen M., 2022, in Ruiz J. E., Pierfedereci F., Teuben P., eds, Astronomical Society of the Pacific Conference Series Vol. 532, Astronomical Society of the Pacific Conference Series. p. 7
Bica E., Pavani D. B., Bonatto C. J., Lima E. F., 2019, AJ, 157, 12
Boch T., Fernique P., 2014, in Manset N., Forshay P., eds, Astronomical Society of the Pacific Conference Series Vol. 485, Astronomical Data Analysis Software and Systems XXIII. p. 277
Bonnarel F., et al., 2000, AAPS, 143, 33
Campello R. J. G. B., Moulavi D., Sander J., 2013, in Pei J., Tseng V. S., Cao L., Motoda H., Xu G., eds, Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 160–172
Cantat-Gaudin T., Anders F., 2020, A&A, 633, A99
Cantat-Gaudin T., et al., 2020, A&A, 640, A1
Casado J., 2021, Research in Astronomy and Astrophysics, 21, 117
Casado J., Hendy Y., 2023, Monthly Notices of the Royal Astronomical Society, 521, 1399
Castro-Ginard A., Jordi C., Luri X., Julbe F., Morvan M., Balaguer-Núñez L., Cantat-Gaudin T., 2018, A&A, 618, A59
Castro-Ginard A., Jordi C., Luri X., Cantat-Gaudin T., Balaguer-Núñez L., 2019, A&A, 627, A35
Castro-Ginard A., et al., 2020, A&A, 635, A45
Castro-Ginard A., et al., 2022, A&A, 661, A118
Chi H., Wang F., Li Z., 2023a, Research in Astronomy and Astrophysics, 23, 065008
Chi H., Wei S., Wang F., Li Z., 2023b, ApJS, 265, 20
Chi H., Wang F., Wang W., Deng H., Li Z., 2023c, ApJS, 266, 36
Dias W. S., Monteiro H., Moitinho A., Lépine J. R. D., Carraro G., Paunzen E., Alessi B., Villela L., 2021, MNRAS, 504, 356
Dreyer J. L. E., 1888, Mem. RAS, 49, 1
Efron B., 1979, The Annals of Statistics, 7, 1
Ester M., Kriegel H.-P., Sander J., Xu X., 1996, in Knowledge Discovery and Data Mining.
Ferreira F. A., Santos J. F. C., Corradi W. J. B., Maia F. F. S., Angelo M. S., 2019, MNRAS, 483, 5508
Ferreira F. A., Corradi W. J. B., Maia F. F. S., Angelo M. S., Santos J. F. C. J., 2020, MNRAS, 496, 2021
Ferreira F. A., Corradi W. J. B., Maia F. F. S., Angelo M. S., Santos J. F. C. J., 2021, MNRAS, 502, L90
Friel E. D., 1995, ARA&A, 33, 381
Froebrich D., Scholz A., Raftery C. L., 2007, MNRAS, 374, 399
Gaia Collaboration et al., 2016, A&A, 595, A1
Gaia Collaboration et al., 2022, arXiv e-prints, p. arXiv:2208.00211
Gran F., et al., 2022, MNRAS, 509, 4962
Hao C., Xu Y., Wu Z., He Z., Bian S., 2020, PASP, 132, 034502
Hao C. J., et al., 2021, A&A, 652, A102
Hao C. J., Xu Y., Wu Z. Y., Lin Z. H., Liu D. J., Li Y. J., 2022, A&A, 660, A4
He Z.-H., Xu Y., Hao C.-J., Wu Z.-Y., Li J.-J., 2021, Research in Astronomy and Astrophysics, 21, 093
He Z., et al., 2022a, ApJS, 260, 8
He Z., Wang K., Luo Y., Li J., Liu X., Jiang Q., 2022b, ApJS, 262, 7
He Z., Luo Y., Wang K., Ren A., Peng L., Cui Q., Liu X., Jiang Q., 2023a, arXiv e-prints, p. arXiv:2305.10269
He Z., Liu X., Luo Y., Wang K., Jiang Q., 2023b, ApJS, 264, 8
Herschel W., 1786, Philosophical Transactions of the Royal Society of London Series I, 76, 457
Høg E., et al., 1997, A&A, 323, L57
Huchra J. P., Geller M. J., 1982, ApJ, 257, 423
Hunt E. L., Reffert S., 2021, A&A, 646, A104
Hunt E. L., Reffert S., 2023, A&A, 673, A114
Hunter J. D., et al., 2007, Computing in science and engineering, 9, 90
Jaehnig K., Bird J., Holley-Bockelmann K., 2021, ApJ, 923, 129
Jones E., Oliphant T., Peterson P., et al., 2001, SciPy: Open source scientific tools for Python, `http://www.scipy.org/`
Kharchenko N. V., Piskunov A. E., Schilbach E., Röser S., Scholz R. D., 2012, A&A, 543, A156
Kos J., et al., 2018, MNRAS, 480, 5242

[10] `http://www.numpy.org/`
[11] `http://www.scipy.org/`
[12] `http://matplotlib.org/`
[13] `https://scikit-learn.org/`
[14] `https://github.com/asteca`
[15] `http://www.astropy.org`

Kounkel M., Covey K., Stassun K. G., 2020, AJ, 160, 279
Krone-Martins A., Moitinho A., 2014, A&A, 561, A57
Li Z., Mao C., 2023, ApJS, 265, 3
Li Z., et al., 2022, ApJS, 259, 19
Liu L., Pang X., 2019, ApJS, 245, 32
Loktin A. V., Popova M. E., 2017, Astrophysical Bulletin, 72, 257
Lynga G., 1987, VizieR Online Data Catalog, p. VII/92A
Mermilliod J.-C., 1995, in Egret D., Albrecht M. A., eds, Vol. 203, Information & On-Line Data in Astronomy. p. 127, doi:10.1007/978-94-011-0397-8_12
Messier C., 1774, Mémoires de l'Académie Royale des Sciences, pp 435–461
Ochsenbein F., Bauer P., Marcout J., 2000, A&AS, 143, 23
Pedregosa F., et al., 2011, Journal of Machine Learning Research, 12, 2825
Pera M. S., Perren G. I., Moitinho A., Navone H. D., Vazquez R. A., 2021, A&A, 650, A109
Perren G. I., Vázquez R. A., Piatti A. E., 2015, A&A, 576, A6
Perryman M. A. C., et al., 1997, A&A, 323, L49
Qin S.-M., Li J., Chen L., Zhong J., 2021, Research in Astronomy and Astrophysics, 21, 045
Qin S., Zhong J., Tang T., Chen L., 2023, ApJS, 265, 12
Ripley B. D., 1976, Journal of Applied Probability, 13, 255–266
Ripley B. D., 1979, Journal of the Royal Statistical Society. Series B (Methodological), 41, 368
Ryu J., Lee M. G., 2018, ApJ, 856, 152
Santos-Silva T., et al., 2021, MNRAS, 508, 1033
Sim G., Lee S. H., Ann H. B., Kim S., 2019, Journal of Korean Astronomical Society, 52, 145
Skrutskie M. F., et al., 2006, AJ, 131, 1163
Tarricq Y., Soubiran C., Casamiquela L., Castro-Ginard A., Olivares J., Miret-Roig N., Galli P. A. B., 2022, A&A, 659, A59
Tian H.-J., 2020, ApJ, 904, 196
Tremmel M., et al., 2013, ApJ, 766, 19
Van Der Walt S., Colbert S. C., Varoquaux G., 2011, Computing in Science & Engineering, 13, 22
Vasiliev E., Baumgardt H., 2021, MNRAS, 505, 5978
Zari E., Hashemi H., Brown A. G. A., Jardine K., de Zeeuw P. T., 2018, A&A, 620, A172
van Rossum G., 1995, Report CS-R9526, Python tutorial. pub-CWI, pub-CWI:adr

APPENDIX A: DATABASES CLEANING

We describe here a summary of the cleaning and standardizing processes applied on almost all of the catalogues mentioned in Table 1. Most of these are small tasks, like manually editing names so that they are consistent across databases, but they were necessary (and rather time consuming). For convenience we abbreviate Kharchenko et al. (2012) as KHAR12 and Bica et al. (2019) as BICA19. The catalogue used for identifying globular clusters (GCs) is that of Vasiliev & Baumgardt (2021) with the addition of the Gran 2, 3, 4, and 5 objects from Gran et al. (2022).

Kharchenko et al. (2012): Selected all entries with *class* equal to "OPEN STAR CLUSTER", resulting in 2858 entries. The astrometry values in this database are of low quality, we do not use these values in the membership estimation process. Added the preferred denomination VDBH to vdBergh-Hagen and VDB to vdBergh clusters, per CDS recommendation (this is done across all the catalogues). Removed entries that match galactic clusters: ESO 456-29 (Gran 1), FSR 1716, FSR 1758, VDBH 140.

Loktin & Popova (2017): Many proper motion values in this catalogue are clearly wrong (e.g: the values for NGC 2516). We thus do not use these values in the membership estimation process. Removed entries that match galactic clusters: Berkeley 42 (NGC 6749), Lynga 7 (BH 184). Fixed the name of four listed OCs: Sauer5 → Saurer 5; Teusch61 → Teutsch 61; AlessiJ2327+55→ Alessi J2327.0+55; Sigma Ori → Sigma Orionis.

Bica et al. (2019): The Vizier table contain 10978 entries, we keep only those with class OC (open cluster) or OCC (open cluster candidate). This reduces the list to 3564 entries.

This is the only DB that lists the Ryu & Lee (2018) clusters. The original article claims to have found 721 new OCs (923 minus 202 embedded). BICA19 (page 11) says that the Ryu & Lee article lists 719 OCs (921 minus 202 embedded). BICA19 lists in its Vizier table only 711 Ryu OCs, 4 of which are listed with alternative names (Teutsch J1814.6-2814 → Ryu 563, Quartet → Ryu 858, "GLIMPSE 70, Mercer 70" → Ryu 273, "LS 468|La Serena 468" → Ryu 094). Hence there are 707 Ryu clusters in the final BICA19 Vizier table. Removed entries: MWSC 2776, FSR 523, FSR 847, FSR 436 (listed twice, removed one of the entries); ESO 393-3 (listed twice and no available data in CDS to decide, removed both); MWSC 1025, 1482, 948, 3123, 1997, 1840, 442, 1808, 2204 (listed twice, removed name from both entries as they were not found in KHAR12); ESO 97-2 (removed from Loden 848 as it matches the position of Loden 894); FSR 972, OCL 344, Collinder 384, FSR 179 (listed twice, removed from both entries as we found no available data in CDS to check); MWSC 206 (listed twice, removed the entry that also showed FSR 60 since the coordinates for FSR 60 are a better match in KHAR12 for the entry with the single FSR 60 name); Alessi J0715.6-0722 (removed as its position matches that of Alessi J0715.6-0727).

Fixed the name of the following entries: FSR 429.MWSC 3667 → FSR 429,MWSC 3667; Carraro 1.MWSC 1829 → Carraro 1,MWSC 1829; Cernik 39 → Czernik 39; de Wit 1 → Wit 1 (to match KHAR12); JS 1 → Juchert-Saloran 1 (to match KHAR12); ESO 589-26,MW → ESO 589-26; Alessi J2327.6+5535 → Alessi J2327.0+55; TRSG 1 → RSG 1; Dol-Dzim 9 added DoDz 9 (to match KHAR12); Dol-Dzim 11 added DoDz 11 (to match KHAR12). Removed entries that match GCs: ESO 456-29,MWSC 2761 (Gran 1); ESO 93-8,MWSC 1932; FSR 1758,MWSC 2617; VDBH 140,vdBergh-Hagen 140,FSR 1632,MWSC 2071.

Sim et al. (2019): Added a 'plx' column estimated as the inverse of the distance (the distance in parsecs is included in the catalogue).

Liu & Pang (2019, LIU19): Added the identifier 'FoF' to all the entries to match HUNT23. Changed 'LP' for 'FoF' in all the catalogues where it appeared, for consistency.

Cantat-Gaudin et al. (2020): Changed Sigma Ori → Sigma Orionis.

Castro-Ginard et al. (2020): Fixed wrong right ascension value for UBC 595 and UBC 181.

Hao et al. (2020): Added the acronym 'HXWHB' to match HUNT23.

He et al. (2021): Added 'CWNU' acronym for consistency across catalogues.

Dias et al. (2021): Lists 177 LIU19 clusters because it includes clusters not listed as new by the authors. We remove all except those listed as new in LIU19. Cluster LP 866 was duplicated (listed also as LP 0866), entry was removed. Changed Sigma Ori → Sigma Orionis.

Hao et al. (2021): This database contains dozens of duplicated

entries and even some that are listed thrice, e.g: ESO 130-06, ESO 368-11, ESO 368-14 and Basel 11a. Furthermore, duplicated clusters are assigned very different fundamental parameters (e.g., Alessi 44 is listed twice and assigned logarithmic ages of 7.82 and 8.42). Of the almost 4000 listed OCs, ∼15% show a difference in the mean parallax value with those from CANTAT20 larger than 50%. Finally some clusters have wildly incorrect astrometric parameters. For example the cluster Melotte 25 is assigned a parallax of 0.264 mas in this catalogue when its true value is larger than 21 mas. We thus unfortunately decided to exclude this catalogue from our list.

Hao et al. (2022): Removed the listed entry OC 0586 as a duplicate of the GC BH 140.

Hunt & Reffert (2023): Removed GCs listed as OCs: Palomar 2, 7 (listed as IC 1276) 8, 10, 11, 12; ESO 452-11 (1636-283); Pismis 26 (Ton 2); Lynga 7 (BH 184). New candidates HSC 2890 and HSC 134 were removed as their position and astrometry match those of the GCs Gran 3 and 4. Candidate HSC 2605 has very similar coordinates and proper motions to GC NGC 5139 but its parallax is different, so it was not removed. Removed moving groups and Theia objects from Kounkel et al. (2020). Fixed: ESO 429-429 → ESO 429-02 (position corresponds to this OC); AH03 J0748+26.9 → AH03 J0748-26.9; Juchert J0644.8+0925 → Juchert J0644.8-0925; Teutsch J0718.0+1642 → Teutsch J0718.0-1642; Teutsch J0924.3+5313 → Teutsch J0924.3-5313; Teutsch J1037.3+6034 → Teutsch J1037.3-6034; Teutsch J1209.3+6120 → Teutsch J1209.3-6120; Collinder 302 changed position to (RA=246.525, DEC=-26.233), it was centred on GC NGC 6121.

For ∼160 HSC candidates we updated their centre values in coordinates. These are extended and irregular objects for which the median positions of their members was more than 1 deg away from the stored values in the HUNT23 database.

Li & Mao (2023): Database lists 56 'LISC' clusters but only 35 are kept as real objects. The parallax distances are in very bad agreement with the the estimated distance moduli. HUNT23 recovers 0% of these clusters. We keep the catalogue but advise caution.

Chi et al. (2023a): The article mentions 83 clusters but only 82 are visible in the article table that lists them. No Vizier data was available at the moment of writing this article and no answer was received after enquiring the author. answer. Added 'LISC-III' to the names to match HUNT23.

This paper has been typeset from a TEX/LATEX file prepared by the author.