Identifying 46 New Open Cluster Candidates in Gaia EDR3 Using a Hybrid pyUPMASK and Random Forest Method

Huanbin Chi (迟焕斌)[1], Shoulin Wei (卫守林)[1], Feng Wang (王锋)[2,3], and Zhongmu Li (李忠木)[4]

[1] School of Management and Economics, Faculty of Information Engineering And Automation, Kunming University of Science and Technology, Kunming 650500, People's Republic of China; weishoulin@astrolab.cn

[2] Center for Astrophysics and Great Bay Center of National Astronomical Data Center, Guangzhou University, Guangzhou 510006, People's Republic of China fengwang@gzhu.edu.cn

[3] Peng Cheng Laboratory, Shenzhen, 518000, People's Republic of China

[4] Institute of Astronomy, Dali University, Dali, 671003, People's Republic of China

Received 2022 September 16; revised 2022 December 21; accepted 2022 December 21; published 2023 March 1

Reporter: Baisong Zhang (张百松)

2024-09-20

1. Introduction

- Stars in open clusters (OCs) simultaneously formed from the same molecular cloud, and gravitationally bound stellar systems were born in the same starburst.
- Therefore, OCs are a kind of natural laboratory and are valuable tracers for studying galactic structure, chemical composition, and dynamical evolution, as well as providing validation and constraints on galaxy evolution models *(Spina et al. 2022).*
- Young OCs are often assigned to analyze the structure of galaxies.
- Old OCs also provide information about the chemical history of the Galaxy, e.g., the relationship between age and metallicity, mixing processes, and cluster destruction processes caused by interactions with other clusters.
- However, due to the limitations of Galactic dust extinction and contamination of field stars (foreground and background stars), identifying OCs is still a challenging issue *(Deb et al. 2022).*

1. Introduction

- Various methods based on unsupervised machine-learning clustering algorithms have been used to search for OCs.
  - One of the most successful searching methods is the Density-based Spatial Clustering of Applications with Noise (DBSCAN) algorithm *(Ester 1996).*
  - A series of DBSCAN variants based on DBSCAN is capable of effectively identifying OCs *(CastroGinard et al. 2018, 2019, 2020; He et al. 2021; Castro-Ginard et al. 2022).*
  - In addition, *Kounkel & Covey (2019), Kounkel et al. (2020), and Hunt & Reffert (2021)* used the improved method **Hierarchical** Density-based Spatial Clustering of Applications with Noise (HDBSCAN) based on DBSCAN to detect many new clusters.
  - *Cantat-Gaudin et al. (2018)* applied an unsupervised membership assignment code (Unsupervised Photometric Membership Assignment in Stellar Clusters; UPMASK) to Gaia DR2 data contained within the fields of those clusters.
  - *Gao (2018)* and *Moriarty et al. (2020)* identified cluster members using a Gaussian mixture model (GMM) clustering method.

1. Introduction

- Besides DBSCAN and its variants, the friends-of-friends (FoF) algorithm *(Yang et al. 2008)* is also applied to identify OCs.
  - *Liu & Pang (2019)* found 76 candidate star clusters from Gaia DR2 using the FoF algorithm.
  - *Li et al. (2022)* also used the FoF algorithm to perform a blind search for OCs in Gaia EDR3 within 25° of the Galactic plane. As a result, 61 new OCs were reported among the 868 candidates.
  - The advantage of the FoF algorithm to group stars is that clustering considers a five-dimensional weighted parameter space of parallax, position, and velocity.
  - However, its disadvantage is that it is not sensitive to the size of the cluster radius and the uneven distribution of star density which changes at different distances.

1. Introduction

- *Liu & Pang (2019)* proposed a high-performance approach (i.e., SHiP) to calculate bFoF in each data region, which has been successful in finding many OCs *(Li et al. 2022).*
  - However, the approach has some minor deficiencies.
  - It is relatively ineffective in searching for member stars in some sparse spaces of star clusters.
  - The minimum number of clusters is set to a predetermined fixed value, e.g., 50. This may lead to some member stars being incorrectly included in a cluster during the merging process of the method.
- In this study, we refer to the study of *Li et al. (2022)* to obtain a data set of OC candidates using the FoF algorithm. We then present an improved hybrid algorithm to identify OCs from OC candidates found by FoF more robustly.

2. An Improved Identification Method for OCs

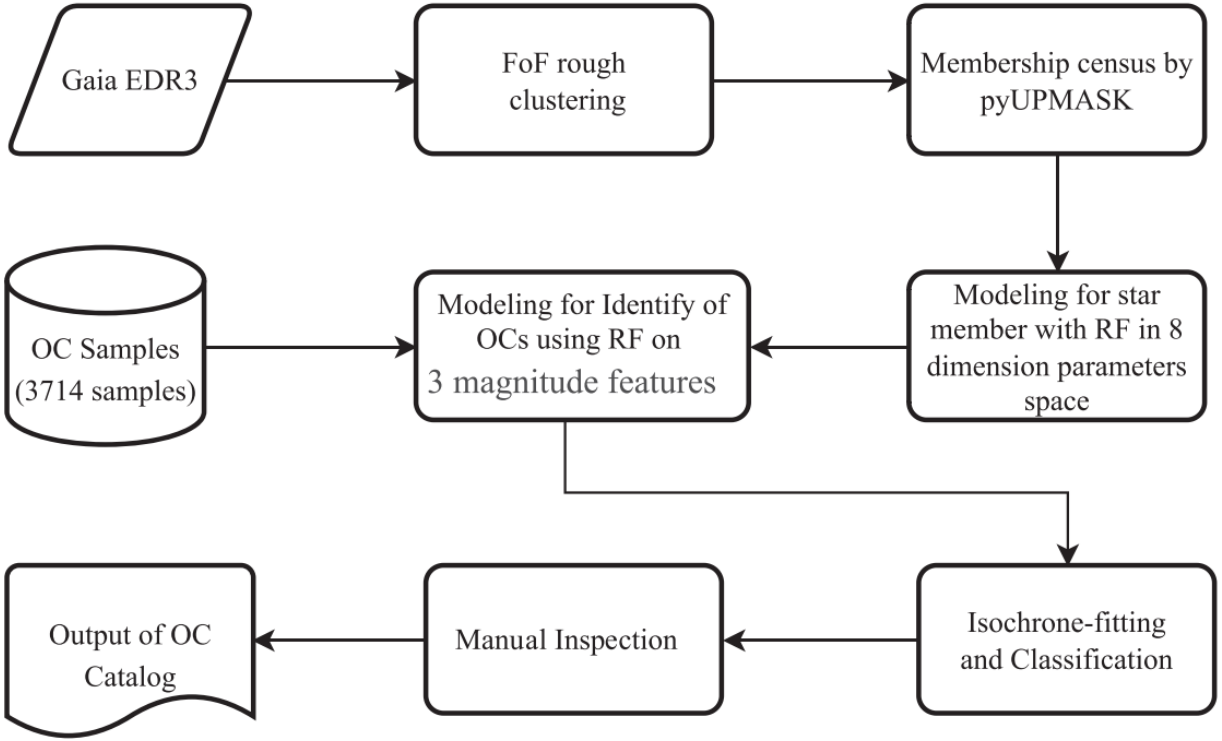


Figure 1. Flowchart of the hybrid identification approach. The process of this approach consists of three stages. Stage 1 deals with rough FoF clustering within Gaia EDR3. Then, each OC's membership is identified in stage 2. After refining the star member, we implement recognition of OCs, including a submodule of modeling for identifying OCs, isochrone fitting, and visual inspection in stage 3.

2.1. Member Star Determination Method

- We presented a hybrid algorithm based on pyUPMASK *(Pera et al. 2021)* and the RF algorithm to eliminate false member stars (field stars) among those star clusters.
  - pyUPMASK is a Python package for UPMASK *(KroneMartins & Moitinho 2014)* used to estimate the membership probability of each input star.
  - pyUPMASK has been widely used in the determination of member stars of OCs based on astrometric parameters *(Cantat-Gaudin et al. 2019; He et al. 2022b; Bai et al. 2022; Dias et al. 2022)*.
  - The essence of the UPMASK algorithm is to calculate the kernel density estimation (KDE) likelihood of the member stars of the OC candidates.

2.1. Member Star Determination Method

$$P_{\mathrm{star}} = \frac{\mathrm{KDE}_m}{(\mathrm{KDE}_m + \mathrm{KDE}_{nm})},$$

— where P_{star}, KDE_m, and KDE_{nm} are the membership probability of a star and the KDE of the members and field stars, respectively.

— After the pyUPMASK calculation, we refer to *Gao (2018)* and set the value of membership probability as **0.8.**

2.2. Identification Model for OC

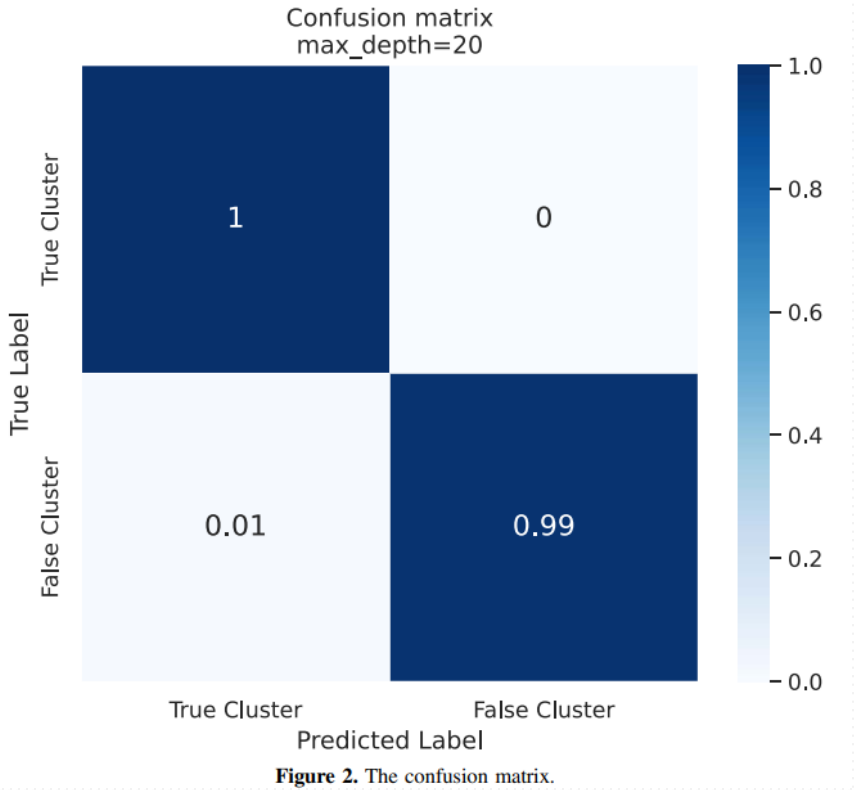


Figure 2. The confusion matrix.

- We applied an RF classifier to detect OCs among the potential candidates.
- The positive OC samples: 1229 collected from Gaia DR2 and 628 from Gaia EDR3.
- The negative OC samples: which have the same number of positive OC samples, are synthesized with stars.
- We finally obtained a training set of 3714 OC samples for modeling.
- The precision of the model is 99.35%.

3. Identification of OCs and the Results

- We strictly followed the data preparation method of *Li et al. (2022)* to generate the OC candidate data set.
- First, we filter suitable samples to exclude observational artifacts due to faintness in Gaia EDR3 based on the stellar position, proper motion, and parallax parameters

$$\text{and parallax parameters } (\varpi \text{ between } 0.2\,\text{mas and } 0.7\,\text{mas}, G < 18 \text{ mag}, \mu_\alpha \cos\delta < 30\,\text{mas yr}^{-1}, \text{and } \mu_\delta < 30\,\text{mas yr}^{-1}).$$

- Meanwhile, considering most OCs are centered near the Galactic disk, we set $|b| < 25^\circ$.
- Second, based on 180 million sources extracted, to facilitate this procedure, we roughly divided the data into many data regions according to Galactic longitude (l), Galactic latitude (b), and parallax (ϖ). The number of divisions for ϖ, b, and l are 8, 8, and 64, respectively.

3. Identification of OCs and the Results

- After carrying out the above scheme, the whole search volume is divided into 4091 data regions.
- We used FoF clustering for each data region to find local clusters and aggregate them to **obtain 3597 candidate clusters**.
- After crossmatching, we got 807 new cluster candidates. Using the membership determination RF model, we removed the field stars from each of these 807 candidates. We then classified these 807 candidates using an identification RF model. **801 candidates were classified as OCs, and the model rejected the other 6 candidates.**

3.2. Validation of the OC Candidates

- We gathered most of the known star cluster catalogs and labeled them as MWSC *(Dias et al. 2012),* CG2017 *(CantatGaudin & Anders 2020*), Hao3794 *(Hao et al. 2021),* UBC series *(Castro-Ginard et al. 2018, 2019, 2020, 2022),* CWNU *(He et al. 2022b), Dias1743 (Dias et al. 2021),* and Hao704 *(Hao et al. 2022)*, respectively.

- Similarly, we have applied the same crossmatch method to some previous catalogs, i.e., *Liu & Pang (2019), Ferreira et al. (2020), Hao et al. (2020), Hunt & Reffert (2021), and Li et al. (2021).* In particular, using the same method, we performed an integral crossover of 46 recently reported star clusters by *He et al. (2022b)*.

- After crossmatching, 501 of the 801 candidates were already identified. We **obtained 300 candidate clusters** that have not been identified and reported, which is the data set used for subsequent OC identification.

3.3. Color–Magnitude Diagram Fitting

- In theory, member stars in an OC are born from the same gas cloud in a single episode of star formation. Most of them are expected to follow a single isochrone in CMDs and have the same metallicity and age.
- Therefore, we used the isochrone-fitting method with the PARSEC theoretical isochrone models *(Bressan et al. 2012)* updated to the Gaia EDR3 passbands using the photometric calibrations from ESA/Gaia to derive their physical parameters (age and metallicity).

$$\bar{d}^2 = \frac{\sum_{k=1}^{n}(\boldsymbol{x}_k - \boldsymbol{x}_{k,nn})^2}{n},$$

- An objective fitting function was applied to the 300 OC candidates, where n is the number of selected members in a cluster candidate and x_k and $x_{k,nn}$ are the positions of the member stars and the points on the isochrone that are closest to the member stars, respectively.

3. Identification of OCs and the Results

We classified the 300 OC candidates into three classes:

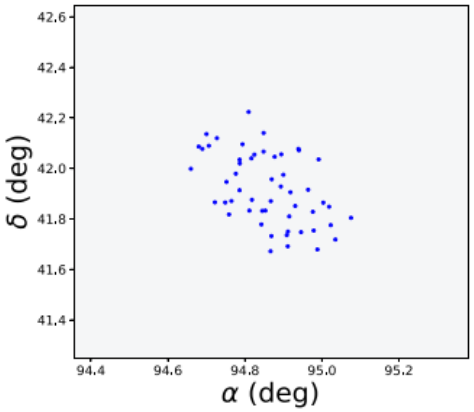
1. Class A: $n_{\rm star} \geqslant 20$, $r_{\rm n} < 0.1$, $\bar{d}^2 < 0.05$;
2. Class B: $n_{\rm star} \geqslant 20$,$r_{\rm n} < 0.1$;
3. Class C: other cases.

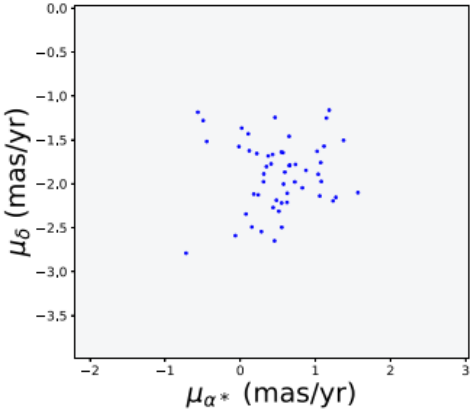
class A (20%, 59/300) has a sufficient number of member stars and clear CMDs;

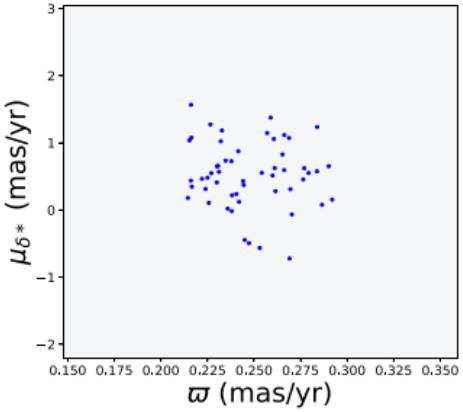
1. $n_{\rm star}$: number of member stars brighter than $G < 17$ mag.
2. $\bar{d}^2$: the average square of the distance between cluster stars and an isochrone used to measure the isochrone fitting errors.
3. $r_{\rm n}$: the narrowness of the main sequence in the CMD calculated as $\left| \frac{v_1}{v_2} \right|$. v_1 and v_2 are the two eigenvalues of the covariance matrix of the distribution of stars in the CMD.

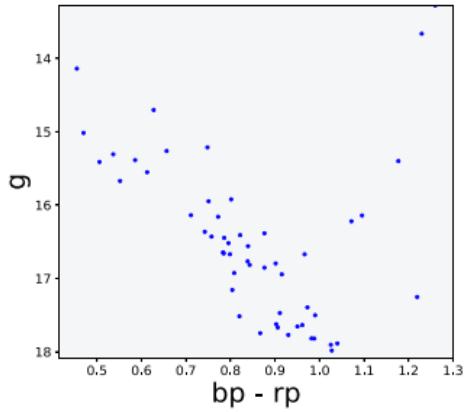
Class B (7%, 21/300) includes candidates with unclear isochrone fitting and loose CMD distributions.

ID:3093

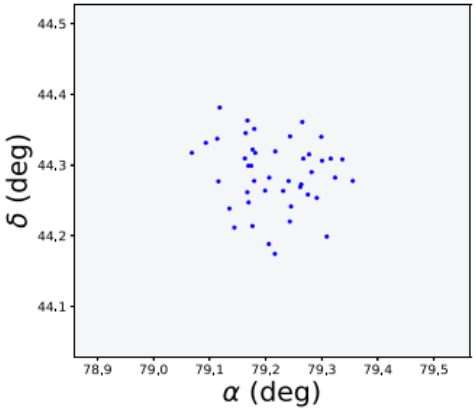


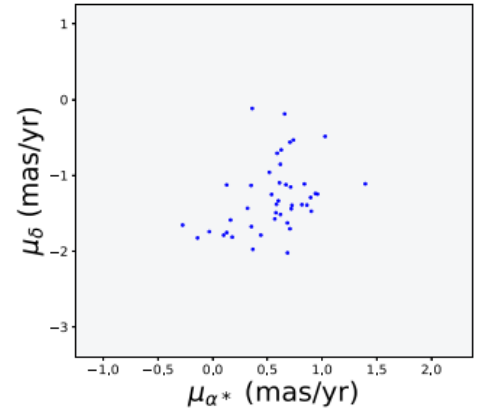


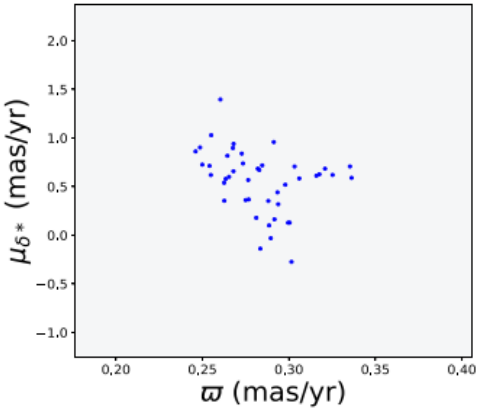


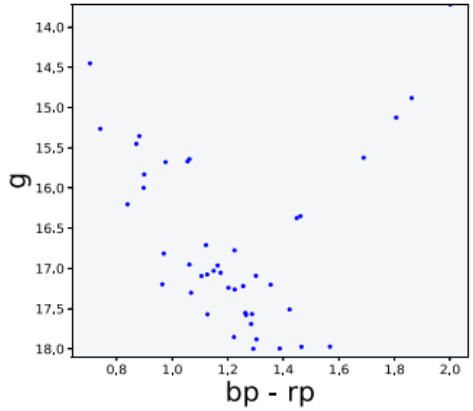


ID:3066









3. Identification of OCs and the Results

- The 80 candidates (class A and class B)
- manual visual inspection
- 46 candidates were finally considered as possible real OCs

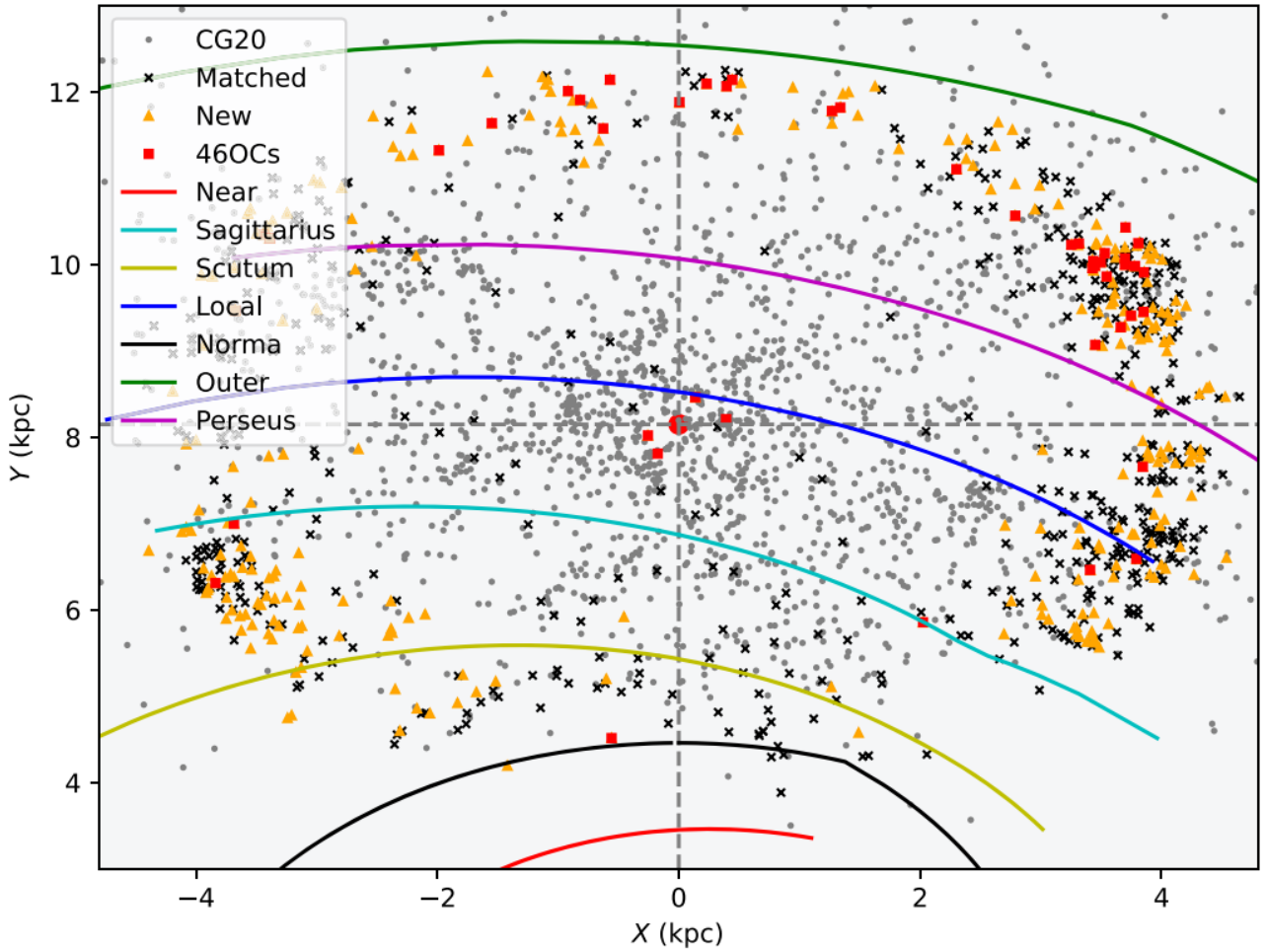


Figure 4. Distributions of the matched and new OCs in the Galactic X−Y plane in relation to the spiral arms.

The image is viewed from the north pole of the Galaxy, around which it rotates clockwise.

3. Identification of OCs and the Results

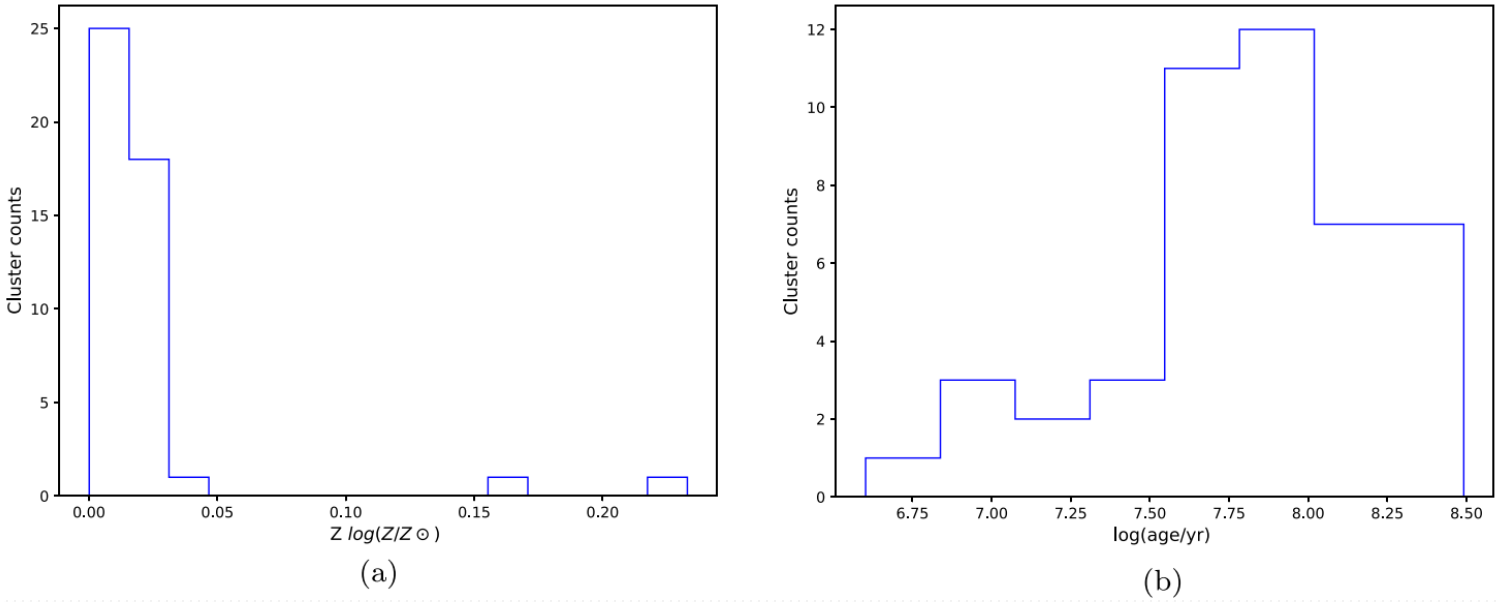


Figure 6. Histograms of Z and age for the 46 OCs.

- The distributions of age and metallicity of the newly identified OCs.
- These OCs are younger than 3.0 Gyr.
- Additionally, most of them are metal-poor.

4. Discussions and Future Works

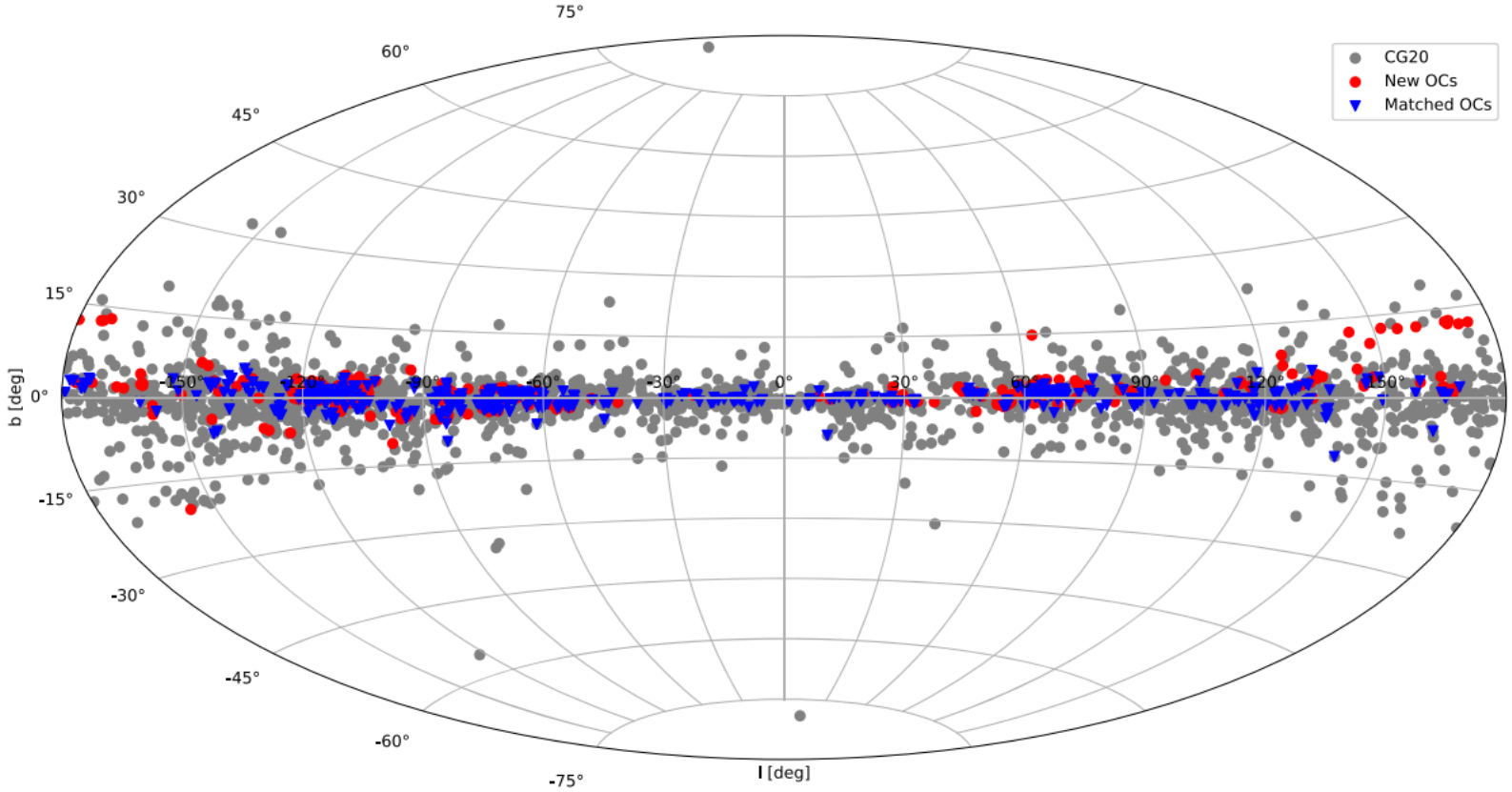


Figure 11. Comparison of distribution of the OC population in (l, b) Galactic coordinates. The black circles present OCs known prior to this study, reported by CG20. Blue triangles represent the OCs found in this work matched with known OCs. Red points represent the new OCs found in this work using Gaia EDR3.

4. Discussions and Future Works

- We identified 46 reliable clusters among 300 OC candidates.
- However, we cannot regard the rest of the 254 candidates as not being OCs. It can only be said that the method we proposed in the study cannot accurately identify these 254 candidates.
- We still suspect that there are OC samples among these 254 candidates.
- We need to find other methods in the future. In addition, multiview learning should be further introduced in the future.

5. Conclusions

- In this study, we proposed a robust approach to identifying OCs.
- For the given OC sample data, a hybrid pyUPMASK and RF method is first used to remove field stars.
- Then an identification model based on the RF algorithm and Gaia EDR3 data is used to identify OC candidates.
- Finally, OC candidates are obtained after isochrone fitting and manual visual inspection.
- Based on the proposed approach, we obtained 46 new reliable OC candidates that have not been reported before, which proved that the method proposed in the study is reasonable.