**RAA 2017** Vol. **17** No. **6**, 58 (6pp) doi: 10.1088/1674–4527/17/6/58

*http://www.raa-journal.org* *http://iopscience.iop.org/raa*

*Research in Astronomy and Astrophysics*

# An application of the KNND method for detecting nearby open clusters based on Gaia-DR1

Xin-Hua Gao

School of Information Science and Engineering, Changzhou University, Changzhou 213164, China; *xhgcczu@163.com*

Received 2017 January 2; accepted 2017 March 10

**Abstract** This paper presents a preliminary test of the $k$-th nearest neighbor distance (KNND) method for detecting nearby open clusters based on Gaia-DR1. We select 38 386 nearby stars ($<$ 100 pc) from the Gaia-DR1 catalog, and then use the KNND method to detect overdense regions in three-dimensional space. We find two overdense regions (the Hyades and Coma Berenices (Coma Ber) open clusters), and obtain 57 reliable cluster members. Based on these cluster members, the distances to the Hyades and Coma Ber clusters are determined to be 46.0$\pm$0.2 and 83.5$\pm$0.3 pc, respectively. Our results demonstrate that the KNND method can be used to detect open clusters based on a large volume of astrometry data.

**Key words:** open clusters and associations: individual (Hyades, Coma Ber) — astrometry — parallaxes — stars: distances

## 1 INTRODUCTION

Nearby open clusters (OCs) provide an important opportunity to study many problems in astrophysics (e.g., stellar evolution and Galactic distance scale), since their distances can be directly determined based on accurate trigonometric parallaxes (Perryman et al. 1998; de Bruijne et al. 2001; van Leeuwen 1999, 2009; de Grijs 2013). Hipparcos data (ESA 1997) have offered an important opportunity for the first time to investigate several nearby OCs (Perryman et al. 1998; van Leeuwen 1999, 2009). The good news is that the Gaia astrometric satellite is collecting high precision positions, parallaxes and proper motions (PMs) for about 1 billion stars throughout our Galaxy. Gaia's main scientific goal is to clarify the composition, formation and evolution of our Galaxy (Gaia Collaboration et al. 2016b). The first Gaia Data Release 1 (Gaia-DR1) contains positions, parallaxes and PMs of about 2 million bright stars ($\sim$11.5 mag). The typical error is about 0.3 mas for parallaxes and positions, and about 1 mas yr$^{-1}$ for PMs (Gaia Collaboration et al. 2016a; Lindegren et al. 2016). High precision astrometric data in the Gaia catalog will offer a new opportunity to study nearby OCs. However, effective methods for fast detection of OCs in large volumes of data, such as Hipparcos and Gaia catalogs, are lacking.

In this paper, we attempt to use the $k$-th nearest neighbor distance (KNND) method to detect nearby OCs within 100 pc of the Sun, using precise positions and parallaxes from Gaia-DR1. At the distance of the Hyades cluster ($\sim$46 pc) (Perryman et al. 1998), a parallax error of 0.3 mas leads to a distance error of $\sim$0.6 pc, which is much less than the cluster tidal radius ($\sim$10 pc) (Perryman et al. 1998). Even at a distance of 100 pc, the typical error of Gaia-DR1 parallaxes only leads to a distance error of $\sim$ 3 pc. In other words, Gaia-DR1 can provide sufficient spatial resolution for stars within 100 pc of the Sun, which allows us to analyze their spatial distribution in three-dimensional (3D) space. Although the KNND method has been confirmed to be effective in OC membership determination based on 3D kinematic data (Gao 2016), its effectiveness in detecting OCs in large volumes of data has never been tested. The remainder of this paper is organized as follows.

In Section 2, we use the KNND method to detect OCs and obtain reliable cluster members. In Section 3, the distances of detected OCs are investigated based on parallaxes of reliable cluster members. The last section summarizes the main conclusions of this work.

## 2 METHOD AND MEMBERSHIP

In data mining, cluster analysis aims to divide multivariate data into different subgroups (or clusters) using various algorithms (Everitt et al. 2011). We have used the DBSCAN and KNND clustering algorithms to determine members of several OCs (Gao 2014, 2016; Gao et al. 2015). The major advantage of our methods is that prior knowledge about star distribution (or cluster model) in OCs is not necessary. In other words, both the DBSCAN and KNND clustering algorithms have the ability to detect arbitrarily shaped cluster structures in given data sets. Very recently, the DBSCAN clustering algorithm was used to analyze the morphology of two OCs, Czernik 20 and NGC 1857, and an unknown high-density region was identified (Bhattacharya et al. 2017). The KNND clustering algorithm is simple and easy to implement. Its basic principle can be summarized as follows: (1) calculate Euclidean distance $d_{ij}$ between the $i$-th and $j$-th sample star ($i \neq j$) in a given data set; (2) sort the distances $d_{ij}$ in ascending order; (3) determine the distance of the $k$-th nearest neighbor from the $i$-th star. The KNND clustering method is based on the assumption that cluster members can be regarded as outliers with significantly smaller KNNDs (with higher point density) compared to field stars (Gao 2016).

We select 38 386 nearby stars within 100 pc of the Sun from the Tycho-Gaia astrometric solution (TGAS) of the Gaia-DR1 catalog (Gaia Collaboration et al. 2016a; Lindegren et al. 2016). Figure 1 shows that most of our sample stars ($\sim$92%) have a small relative error ($\sigma\pi/\pi < 5\%$). Let ($\alpha$, $\delta$, $\pi$) be respectively the right ascension, declination and parallax of a sample star. The corresponding rectangular coordinates ($X$, $Y$, $Z$) in pc can be computed using the following formulas

$$D = 1000 \times \pi^{-1}, \quad (1)$$

$$X = D \times \cos(\alpha) \times \cos(\delta), \quad (2)$$

$$Y = D \times \sin(\alpha) \times \cos(\delta), \quad (3)$$

$$Z = D \times \sin(\delta), \quad (4)$$

where $D$ is the distance to the star in pc. The Euclidean distance $d_{ij}$ between the $i$-th and $j$-th star in the $X$-$Y$-$Z$ space can be calculated using the following formula

$$d_{ij} = \sqrt{(X_i - X_j)^2 + (Y_i - Y_j)^2 + (Z_i - Z_j)^2}. \quad (5)$$

The KNND of the $i$-th star in the $X$-$Y$-$Z$ space can be easily determined by sorting $d_{ij}$ in ascending order.

As shown in Figure 2, a conservative value of $k$ =10 is adopted to enhance the "contrast" between cluster and field stars. Meanwhile, we use a threshold value of 3 pc to segregate cluster members from field stars, and 60 outliers (candidate members) are obtained for further identification (see Fig. 2). Figures 3 and 4 clearly show that the 60 candidate members can be divided into two star groups (the Hyades and Coma Berenices (Coma Ber) clusters) in the 3D $X$-$Y$-$Z$ space and two-dimensional (2D) projected space. Figure 5 shows that the PM vector-point diagram (VPD) of the 60 candidate members can also be divided into two star groups except for three isolated outliers (field stars). Among these candidate members, 49 stars belong to the Hyades cluster, eight stars belong to the Coma Ber cluster and three stars are likely field stars.

## 3 CLUSTER DISTANCES

In consideration of the Hipparcos distance controversy related to the Pleiades cluster (Melis et al. 2014; Mädler et al. 2016), we decided to redetermine distances to the two OCs using Gaia-DR1 parallaxes. Having obtained reliable cluster members, we can compute their weighted mean parallaxes using the following formulas

$$p_i = (\sigma\pi_i)^{-2}, \quad (6)$$

$$\overline{\pi} = \frac{\sum(\pi_i \times p_i)}{\sum p_i}, \quad (7)$$

$$\sigma\overline{\pi} = \left[\frac{\sum((\overline{\pi} - \pi_i)^2 \times p_i)}{(N-1) \times \sum p_i}\right]^{1/2}, \quad (8)$$

where $p_i$ is the weight of the $i$-th member, $\overline{\pi}$ and $\sigma\overline{\pi}$ are the weighted mean parallax and the corresponding uncertainty respectively, $\pi_i$ and $\sigma\pi_i$ are parallax of the $i$-th member and the corresponding observational error respectively, and $N$ is the total number of cluster members. According to the mean parallax $\overline{\pi}$ and its corresponding uncertainty $\sigma\overline{\pi}$, the mean distance $\overline{D}$ and the corresponding uncertainty $\sigma\overline{D}$ can be calculated using the following formulas

$$\overline{D} = 1000 \times \overline{\pi}^{-1}, \quad (9)$$

$$\sigma\overline{D} = \frac{\overline{D}^2 \times \sigma\overline{\pi}}{1000}. \quad (10)$$

The mean parallax of the Hyades cluster is determined to be 21.49±0.10 mas, which corresponds to a distance of 46.5±0.2 pc. The mean parallax and distance of

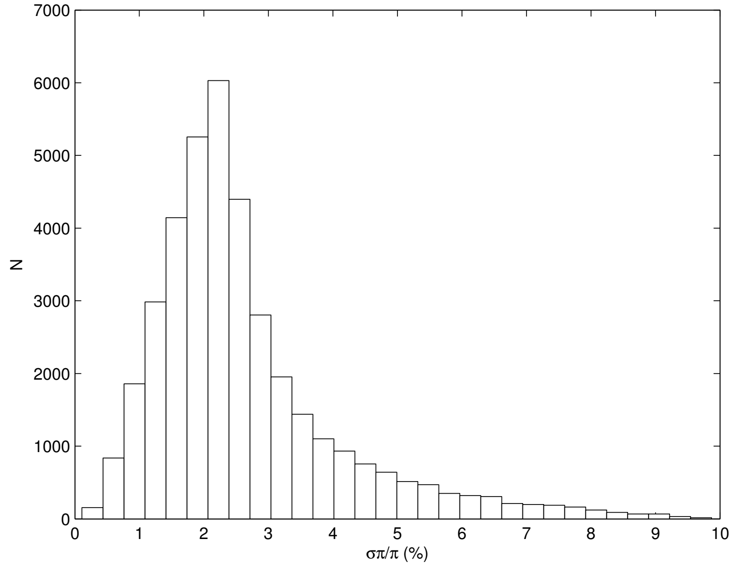


**Fig. 1** The relative parallax errors of 38 386 sample stars within 100 pc of the Sun.

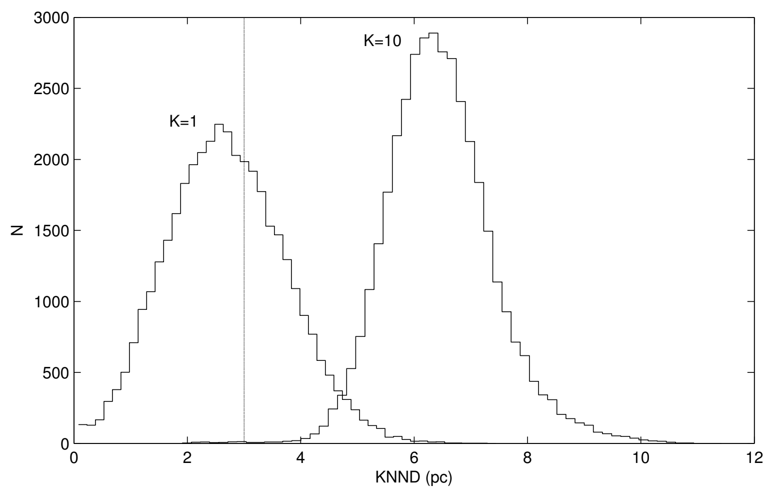


**Fig. 2** The KNND distribution of the 38 386 sample stars ($k = 1, 10$). The vertical line indicates the threshold value (3 pc) used for segregating cluster members from field stars.

the Coma cluster are determined to be 11.72±0.04 mas and 85.3±0.3 pc, respectively. However, a systematic offset of $\sim -0.25$ mas has been found in Gaia-DR1 parallaxes (Stassun & Torres 2016; Jao et al. 2016). Therefore, the distances to the Hyades and Coma Ber clusters are adjusted to be 46.0±0.2 and 83.5±0.3 pc, respectively. The Hyades distance derived in this work is quite consistent with the values derived by other authors (Peterson & Solensky 1988; Torres et al. 1997b,a; van Altena et al. 1997b,a; Perryman et al. 1998; Percival et al. 2003; van Leeuwen 2009; Francis & Anderson 2012) and the Coma Ber distance is also consistent with the values derived by other authors (van Leeuwen 1999, 2009; Francis & Anderson 2012).

It should be noted that the cluster members used to determine the above distances are located in high-density central regions of the two clusters, because they are obtained based on conservative selection criteria (see Fig. 2). Although many sparse members have been neglected in this work, we obtain reliable distances to the cluster centers.

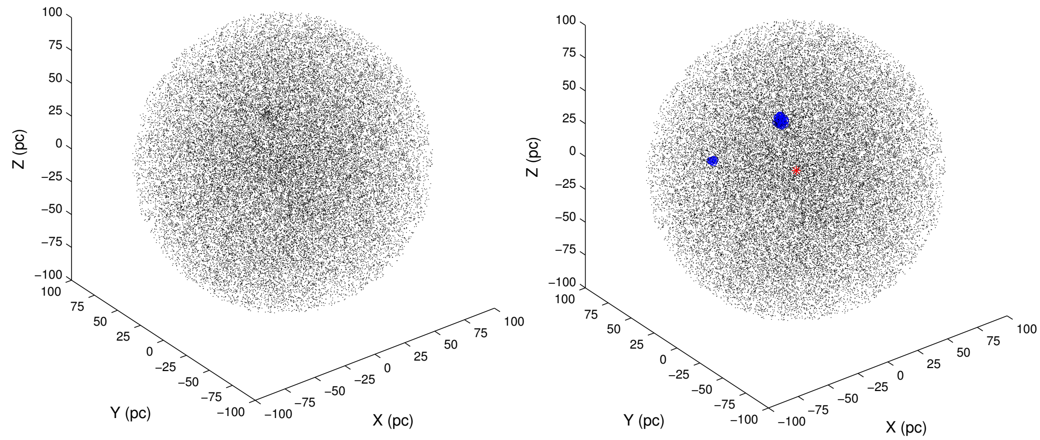


**Fig. 3** (*Left*) Spatial distribution of the 38 386 sample stars in the 3D $X$-$Y$-$Z$ space. (*Right*) The blue circles indicate the 60 candidate members derived based on the KNND method ($k = 10$) and the Sun (*red asterisk*) is located at the origin (0, 0, 0) of the coordinate system.

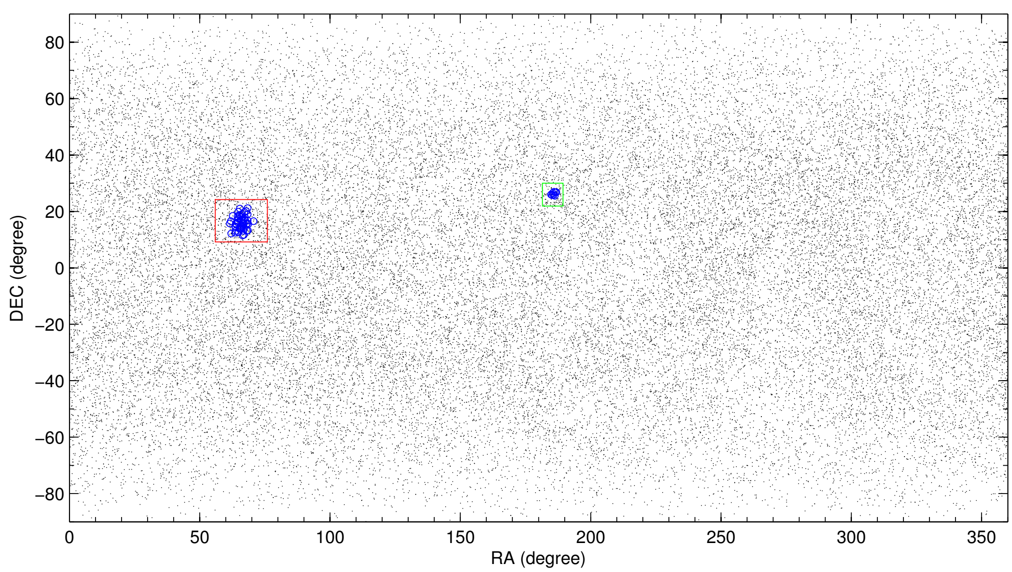


**Fig. 4** The 2D projected spatial distribution of the 60 candidate members (*blue circles*) and field stars (*black dots*). The red and green rectangles represent candidate members of the Hyades and Coma Ber clusters, respectively.

## 4 CONCLUSIONS AND DISCUSSION

In this paper, we use the KNND method to detect nearby OCs based on Gaia-DR1 data, and obtain 57 reliable cluster members that belong to the Hyades and Coma Ber OCs. Based on Gaia-DR1 parallaxes of 57 cluster members, the mean distances to these two clusters are determined. Our distances are in good agreement with values derived by other authors, which implies that the KNND method is effective in detecting dense cluster structures in 3D space. The major advantage of the KNND method is that it is a non-parametric (or non-model-based) clustering method, so prior knowledge about the nature of OCs is not necessary. The KNND method focuses on the

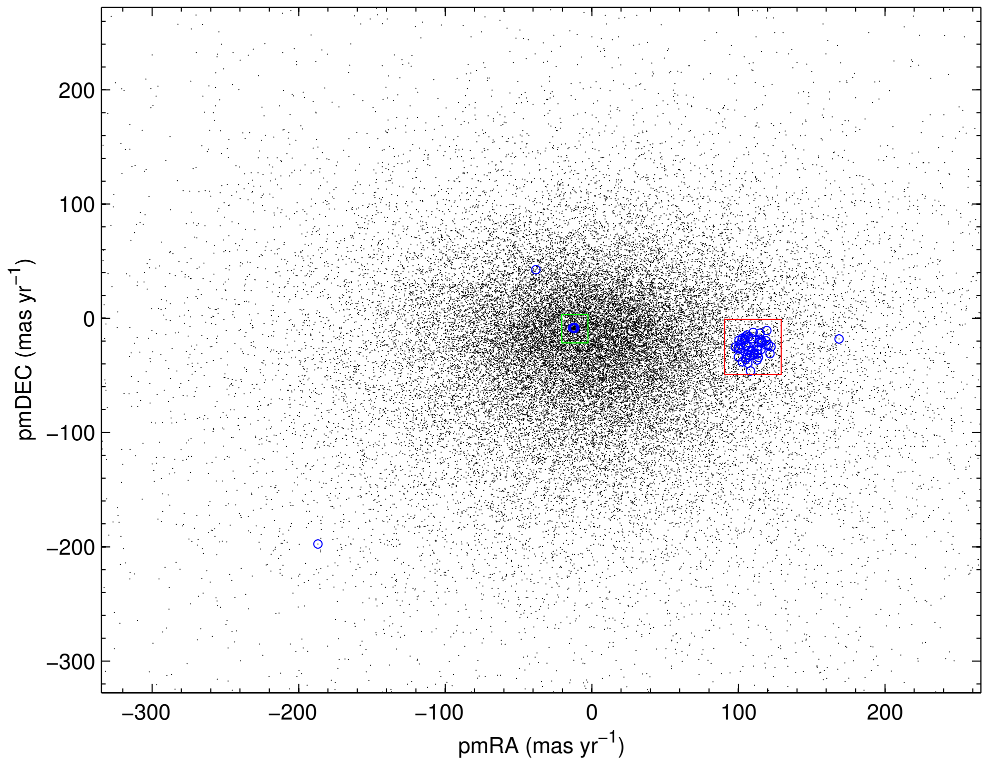


**Fig. 5** The PM VPD of the 60 candidate members (*blue circles*) and field stars (*black dots*). The red and green rectangles represent candidate members of the Hyades and Coma Ber clusters, respectively.

distance between a star and its $k$-th nearest neighbor, not on distribution of the stars, which allows us to detect a low number of cluster members from a large sample of stars. It should be noted that this is a result based on preliminary, not complete, parallaxes. Further tests will be conducted to verify whether the KNND method can be used to detect more known and unknown OCs in a larger volume (out to many kpc) using forthcoming Gaia data releases.

**Acknowledgements** This research was supported by the National Natural Science Foundation of China (NSFC, Grant No. 11403004). This research has made use of the VizieR catalogue access tool, CDS, Strasbourg, France.

## References

Bhattacharya, S., Mahulkar, V., Pandaokar, S., & Singh, P. K. 2017, Astronomy and Computing, 18, 1

de Bruijne, J. H. J., Hoogerwerf, R., & de Zeeuw, P. T. 2001, A&A, 367, 111

de Grijs, R. 2013, in IAU Symposium, 289, Advancing the Physics of Cosmic Distances, ed. R. de Grijs, 351

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. 2011, Cluster Analysis, 5th Edition, Wiley Series in Probability and Statistics, Established by Walter A. Shewhart & Samuel S. Wills (Wiley Online Library)

Francis, C., & Anderson, E. 2012, Astronomy Letters, 38, 681

Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2016a, A&A, 595, A2

Gaia Collaboration, Prusti, T., de Bruijne, J. H. J., et al. 2016b, A&A, 595, A1

Gao, X.-H. 2014, RAA (Research in Astronomy and Astrophysics), 14, 159

Gao, X.-H., Xu, S.-K., & Chen, L. 2015, RAA (Research in Astronomy and Astrophysics), 15, 2193

Gao, X.-H. 2016, RAA (Research in Astronomy and Astrophysics), 16, 184

Jao, W.-C., Henry, T. J., Riedel, A. R., et al. 2016, ApJ, 832, L18

Lindegren, L., Lammers, U., Bastian, U., et al. 2016, A&A, 595, A4

Mädler, T., Jofré, P., Gilmore, G., et al. 2016, A&A, 595, A59

Melis, C., Reid, M. J., Mioduszewski, A. J., Stauffer, J. R., & Bower, G. C. 2014, Science, 345, 1029
Percival, S. M., Salaris, M., & Kilkenny, D. 2003, A&A, 400, 541
Perryman, M. A. C., Brown, A. G. A., Lebreton, Y., et al. 1998, A&A, 331, 81
Peterson, D. M., & Solensky, R. 1988, ApJ, 333, 256
Stassun, K. G., & Torres, G. 2016, ApJ, 831, L6
Torres, G., Stefanik, R. P., & Latham, D. W. 1997a, ApJ, 485, 167
Torres, G., Stefanik, R. P., & Latham, D. W. 1997b, ApJ, 474, 256
van Altena, W. F., Lee, J. T., & Hoffleit, E. D. 1997a, Baltic Astronomy, 6, 27
van Altena, W. F., Lu, C.-L., Lee, J. T., et al. 1997b, ApJ, 486, L123
van Leeuwen, F. 1999, A&A, 341, L71
van Leeuwen, F. 2009, A&A, 497, 209