Highlights on Spanish Astrophysics IX, Proceedings of the XII Scientific Meeting of the Spanish Astronomical Society held on July 18 – 22, 2016, in Bilbao, Spain. S. Arribas, A. Alonso-Herrero, F. Figueras, C. Hernández-Monteagudo, A. Sánchez-Lavega, S. Pérez-Hoyos (eds.)

Clusterix 2.0 for *Gaia*

L. Balaguer-Núñez[1], D. Galadí-Enríquez [2], M. López del Fresno[3,4], E. Solano[3,4], C. Jordi[1], T. Sézima[5], and E. Paunzen[5]

[1] Dept. FQA, Institut de Ciències del Cosmos. Universitat de Barcelona IEEC-UB, Barcelona, Spain
[2] Centro Astronómico Hispano Alemán CAHA, Almería, Spain
[3] Centro de Astrobiología (INTA-CSIC), Departamento de Astrofísica. P.O. Box 78, E-28691, Villanueva de la Cañada, Madrid, Spain
[4] Spanish Virtual Observatory
[5] Masaryk University, Brno, Czech Republic

Abstract

We present an advanced, VO-compliant version of Clusterix, a tool for the determination of membership probabilities in stellar clusters from proper motion data. Clusterix is a web-based, interactive application that allows the computation of membership probabilities from proper motions through a fully non-parametric method [7, 1]. Version 1.0 (`http://clusterix.cerit-sc.cz/`) was developed as a collaboration between the Masaryk University (Czech Republic) and the Universitat de Barcelona (Spain), as a complement to the WEBDA (`http://webda.physics.muni.cz`) database [15] of observational data on stars in open clusters.
Clusterix 2.0 (`http://clusterix.cab.inta-csic.es/clusterix/`) is oriented towards the exploitation of *Gaia* data products. With the participation of the Spanish Virtual Observatory, Clusterix now features an improved user interface for a faster, easier and more accurate interactive definition of the cluster and field proper motion distributions. The system provides fast feedback between membership probability determinations and the distribution of observables for the most probable members and field stars, with graphic tools to display, for instance, photometric diagrams on the fly. Furthermore, Clusterix 2.0 is fully VO-compatible, what opens interesting prospects for the astrophysical exploitation of the improved membership probabilities that will be capable to provide for many open clusters observed by *Gaia*.

1 Introduction

Open clusters are a valuable tool for undertaking studies of our Galaxy and stellar astrophysics. Clusters have been used to determine the spiral structure of the Galaxy and inves-

tigate star formation and evolution processes. They are good tracers of the dynamics [6] and the chemical evolution of our Galaxy's disk [5]. In the advent of high precision surveys such as *Gaia*, their contribution to astrophysical studies should become increasingly important.

Gaia Data Release 1 (DR1), published on 14 September 2016, includes positions, proper motions and parallaxes for about 2 milion stars in common between *Gaia* and the Hipparcos and Tycho-2 catalogues, computed as part of the Tycho-*Gaia* Astrometric Solution [4, 13]. Only the brightest clusters are included in this first release. For the *Gaia* DR2, expected for the end of 2017, we will have five parameter astrometric solutions of single stars covering at least 90% of the sky, plus integrated BP and RP photometry. Those catalogues open up a huge amount of open clusters with high quality homogeneous proper motions that will need new approaches on the tools developed up to now when only a few studies of a few clusters were available.

However, as noted by [6], the inaccuracy of current membership determinations poses difficulties in conducting these studies on a large scale. Accurate membership determination is thus essential for further astrophysical studies of clusters. The determination of the mean properties of open clusters (like radial velocity or metallicity) requires prior knowledge of their member stars to avoid the costly process of obtaining and reducing spectroscopic data on a large scale. Moreover, knowing the membership probability of the stars in a cluster area helps to select member stars in a colour-magnitude diagram and determine distance and age of the cluster by isochrone fitting. Hence, a precise identification of the stars that compose a cluster is critical to accurately determine the kinematic and fundamental parameters of the clusters(age, total mass, etc), which are essential for studies of the Galactic dynamics.

Clusterix is an easy tool for membership probabilities determination and also allows the possibility of gathering physical parameters (parallaxes, radial velocities, proper motions,...) from Virtual Observatory services and estimating effective temperatures, surface gravities and metallicities using VOSA (`http://svo2.cab.inta-csic.es/theory/vosa/`)

2 Membership probabilities from proper motions: non-parametric method

The classical approach to cluster field segregation is the parametric method. This method assumes the existence of two populations in the Vector-Point Diagram (VPD): cluster and field. The corresponding frequency functions are modelled as parametric Gaussian functions: a circular Gaussian model is adopted for the cluster distribution, while a bivariate (elliptical) Gaussian describes the field. See [10] for a full description. But there are two main drawbacks of the parametric method:

1. A circular bivariate function for the cluster probability density function (pdf) is good if the velocity dispersion is not resolved and there are no systematic differences of accuracy between the two proper motion axes, and

2. the choice of an elliptic bivariate function for the field pdf is non realistic. The proper motion distribution of field stars has an intricate structure dominated by the combi-

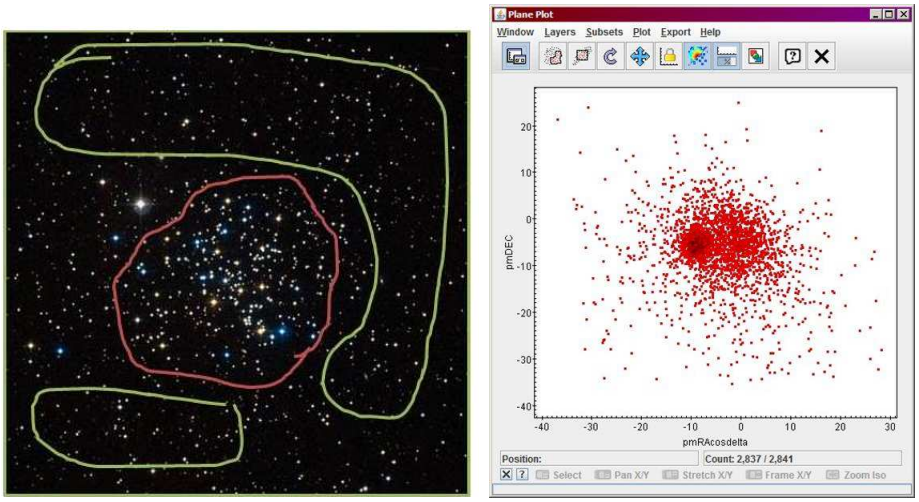

Figure 1: *Left*: M 67 open cluster areas covered by the two populations: cluster area marked in red and field areas in green. *Right*: Vector point diagram (VPD) of M 67, used as an example. Proper motions from ROA.

nation of solar motion and galactic differential rotation. Depending on the direction of observation and the proper motion accuracy, several asymmetries arise, the most evident of which is the tail in the direction of the solar antiapex. Furthermore, real field distribution wings are stronger than those predicted by a Gaussian model, and empirical data show that many other irregularities can be present in real world proper motion field distributions.

The cluster/field segregation from astrometry can also be analysed with a non-parametric approach, as explained in length in [1, 2, 3]. We perform an empirical determination of the frequency functions from the VPD, without relying on any previous assumption about their profiles. In the area occupied by the cluster, the frequency function $\Psi_{\rm c+f}$ is made up of two contributions: cluster, $\Psi_{\rm c}$, and field, $\Psi_{\rm f}$. To disentangle the two populations, we study the VPD corresponding to two areas: one surrounding the cluster to characterize the field (marked in green in Fig. 1), and one centred in the cluster (in red in Fig. 1).

The only two assumptions that we need to apply in the non-parametric approach are:

1. it is possible to select some area in the region under study relatively free of cluster stars (marked in green in the example of Fig. 1), and to determine the frequency function corresponding to the VPD of this area. This will provide a representation of the field frequency function with a small (or negligible) cluster contribution, and

2. this frequency function is representative of the field frequency function in the area occupied by the cluster (marked in red in Fig. 1).

It is important to perform tests with areas of very different size searching for a reasonable tradeoff between cleanness (absence of a significant amount of cluster members) and

signal-to-noise ratio (working area not too small, to avoid small number statistics). The kernel density estimator technique ([9]) is then applied in the VPD. Details of the procedure can be found in [8]. Circular Gaussian kernel functions are used, with Gaussian dispersion (also called smoothing parameter) h selected according to Silverman's rule ([11]), although this parameter can also be interactively tuned by the user. In general, the empirical frequency functions are computed for a grid with cell size of around 0.2 mas yr^{-1}, well below the proper motion errors of most current data. The procedure is tested for several subsamples applying different proper motion cutoffs with a default value of $|\mu| \leq 15$ mas yr^{-1}.

Once we find an area that yields a clean frequency function with low cluster contamination and low noise, we scale this field frequency function to represent the field frequency function in the cluster area, $\Psi_{\rm f}$, by simply applying a factor linked to the area. The cluster empirical frequency function can then be determined as $\Psi_{\rm c} = \Psi_{\rm c+f} - \Psi_{\rm f}$. These empirical frequency functions can be normalized to yield the empirical pdf's for the mixed population, for the field and for the clusters (non-field) population. The probability for a star in a node of the grid being a member of the cluster is $P_c = (\Psi_{c+f}$ - $\Psi_f)/\Psi_{c+f}$. The empirical tables can then be used to estimate the probability of a star being a cluster member according to the probability of its nearest node. These probability tables are then applied to all the stars in the surveyed area, both inside and outside the areas defined to determine the functions. Of course, the field estimated pdf in the outer area cannot be an absolutely perfect representation of the true field pdf in the whole area. This introduces undesired noise in the frequency function of the cluster. The negative density values usually found in several zones obviously lack physical meaning. These negative values allow us to estimate the typical noise level, γ, present in the result. To avoid meaningless probabilities in zones of low density we restrict by default the probability calculations to the stars with cluster pdf $\geq 3\gamma$.

Moreover, the non-parametric technique yields an expected number of cluster members, N_{mem}, from the integrated volume of the cluster frequency function in the VPD areas of high cluster density, where pdf $\geq 3\gamma$. Sorting this sample in order of decreasing membership probability, P_c, the first N_{mem} stars are the most probable cluster members.

3 C

lusterix

All this cumbersome procedure, originally in Fortran, has been implemented in an interactive web-based application (see [12]) and now extended and made fully VO compliant. It also allows the possibility of gathering physical parameters (parallaxes, radial velocities, proper motions,...) from VO services and estimating effective temperatures, surface gravities and metallicities using VOSA [14].

From the Clusterix main page you can choose to search by coordinates in a catalogue, to search in Webda database of clusters or you can upload your own file with positions and proper motions. From any of these inputs, you can calculate the membership determination from proper motions and you can also get some extra information from VO services (radial velocities, parallax and VOSA photometry).

Once you send your selected objects to calculate "Membership determination from proper motions", you get a first approach with basic information that will be used to determine membership probabilities for those stars. As a first approximation this version still uses a simplified area definition where the $\Psi_{\mathrm{c+f}}$ is calculated from the stars located in a circular area around the center of the cluster, and the Ψ_{f} from the stars located in the area between this inner radius and an outer one. It gives you the diameter of the area covered by your selection of stars, the center of them, the number of stars in the inner area selected ("cluster+field" selection), the number of stars in the outer area selected ("field" selection) and the total number of stars selected.

You can -and should- change the default values interactively: the coordinates of the center of the cluster, the inner radius, the outer radius and all the restrictions applied: the cutoffs in proper motion and in proper motion error, and the threshold in typical noise level, γ.

4 Results

We show here as an example the case for NGC 2682 (M 67). We have proper motions for 2841 stars in an area of 2° by 1.4° taken from ROA (see Fig. 1) and INT-WFC uvbyH$_\beta$ complete photometry for 1518 stars in an area of 45' by 45' (see Fig. 2).

We find the results marked in blue in Fig. 2. As you can appreciate in the figure, using proper motion information only we obtain a colour magnitude diagram almost clean from field stars.

5 Future enhancements

At the time of writing this proceedings, we have a beta version of Clusterix with the ability of choosing arbitrary areas (non linked, polygonal, any shape...) for field and cluster selection.

Future enhancements include in the short term, the posibility of using different parameters (e.g. tangential velocities instead of proper motions) and in the long term, the possibility of an n-dimensional approach (that may include for instance, radial velocities, or other observables).

Acknowledgments

Clusterix 2.0 is maintained by the Spanish Virtual Observatory at the Data Archive Unit of the CAB (INTA-CSIC). This work was supported by the MINECO (Spanish Ministry of Economy) - FEDER through grant ESP2014-55996-C2-1-R and MDM-2014-0369 of ICCUB (Unidad de Excelencia 'María de Maeztu') and the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement GENIUS FP7 - 606740. This research has made use of the Spanish Virtual Observatory (http://svo.cab.inta-csic.es) supported from the Spanish MINECO through grant AyA2014-55216. This research has made use of the WEBDA database, operated at the Department of Theoretical Physics and Astrophysics of the Masaryk University.

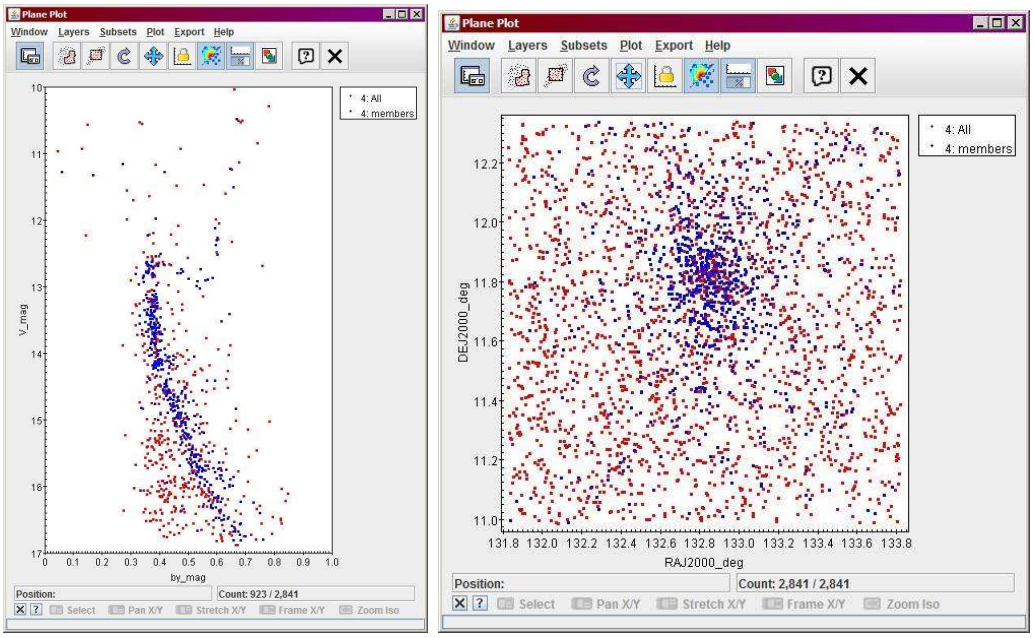

Figure 2: Colour-magnitude diagram and spatial distribution of our proper motions for M 67. Members chosen by Clusterix in blue.

References

[1] Balaguer-Núñez, L., Jordi, C., Galadí-Enríquez, D., Zhao, J.L. 2004, A&A, 426, 819

[2] Balaguer-Núñez, L., Jordi, C., Galadí-Enríquez, D. 2005, A&A, 437, 457

[3] Balaguer-Núñez, L., Galadí-Enríquez, D., Jordi, C. 2007, A&A, 470, 585

[4] *Gaia* Collaboration, Brown, A.G.A., Vallenari, A., et al., 2016, A&A,

[5] Casamiquela, L., Carrera, R., Jordi, C., et al., 2016, this volume

[6] Finchaboy, P.M. & Majewski, S.R., 2008, AJ 136, 118

[7] Galadí-Enríquez, D., Jordi, C., Trullols, E. 1998, A&A, 337, 125

[8] Galadí-Enríquez D., Jordi C., & Trullols E., 1998, A&A 337, 125

[9] Hand, D. 1982, Kernel Discriminant Analysis (Chichester Research Studies Press)

[10] Sanders, W.L., 1971, A&A 14, 226

[11] Silverman B.W., 1986, Density Estimation for Statistics and Data Analysis, J.W.Arrowshmith.

[12] Sezima, T., Galadí-Enríquez, D., et al. 2015, Highlights of Spanish Astrophysics VIII

[13] Lindegren, L., Lammers, U., Bastian, U., et al., 2016, A&A,

[14] Bayo, A., Rodrigo, C., Barrado y Navascués, D., et al., 2008, A&A 492,277B.

[15] Netopil, M., Paunzen, E., Stutz, C. 2012, ASSP 53