Research Networking Programmes

Exchange visit – Science report

Cluster Census and Membership Determination in the Milky Way

1) Purpose of the visit

The study of the formation and evolution of open star clusters (OCs) and their stellar populations represents a backbone of research in modern astrophysics, with a strong impact on our understanding of key open issues, from the star formation process (SF), to the assembly and evolution of the Milky Way (MW) disc, and galaxies in general. Currently, about 2200 open clusters and cluster candidates have been identified in our Galaxy (Dias et al. 2002, A&A, 389, 871: DAML catalogue version 3.4, 2014 May 24).

Automated cluster searches in infrared survey data such as 2MASS (Skrutskie et al. 2006, AJ, 131, 1163; Froebrich et al. 2007 MNRAS, 374, 399), the UKIDSS Galactic Plane Survey (Lucas et al. 2008, MNRAS, 391, 136) have identified a large number of apparent stellar overdensities in the disk and in the high extinction regions whose nature still need to be ascertained. In spite of that, the cluster census is far from complete and some models predict that existing clusters in the MW should be as many as 100 times the actual number (Bonatto et al. 2006, A&A, 446, 121). Efficient automated searches are needed for several reasons:

- searches based on star counts alone produce false positives (identifying stellar overdensities which are not actual physical objects).
- searches based on star counts are not able to identify clusters projected against a dense background.
- the unprecedented amount of data expected from the Gaia mission or the LSST program will pose a technical challenge and forbid manual searches.

The cornerstone ESA Gaia mission will bring us in a new domain of cluster research. Gaia will be sensitive to stars down to G=20; it will measure distances for individual stars in OCs with a precision better than 1% for clusters closer than about 1 kpc and better than 10% for the entire OC family. Higher accuracies are expected for proper motions at end of mission, reaching individual tangential velocity accuracy of the order of $0.2 - 0.3$ km/s for low mass stars up to 1.5 kpc, and up to larger distances for bright O/B stars. Gaia has proven to have a spatial resolution comparable or even higher than HST data enabling the detection of less massive OCs. It is expected that Gaia will provide a fairly complete census of OCs within 3 kpc.

T. Cantat-Gaudin was a guest at the University of Lisbon during his PhD work, where he started working on extending the capabilities of the UPMASK cluster membership algorithm (Krone-Martins & Moitinho 2014, A&A, 561, 57) to higher-dimensionality spaces, allowing the use of parallaxes and kinematic information. The non-parametric approach of UPMASK can be adapted in order to build a cluster detection scheme that is able to take advantage of the high dimensionality of future all-sky catalogues such as Gaia.

2) Description of the work carried out during the visit

2.1) Estimating the significance of an overdensity in all quantities of the phase space, working with simulated data

Synthetic fields of view containing clusters and field stars were constructed, with a full 6D phase space (sky coordinates (α,δ), proper motions (μ_α,μ_δ), parallaxes π and radial velocities RV) and photometry in (U,B,V,I,J,H,K) bands. The philosophy of UPMASK is to identify stars grouping in the $(\mu_\alpha,\mu_\delta,\pi,RV,U,B,V,I,J,H,K)$ and check if their sky distribution (α,δ) appears statistically different from a random realisation of the same number of points. Our implementation does not only require that such groups are identified, but also requires that all such groups, taken as a whole, show a distribution that is significantly more "clumpy" than the field stars, in all available quantities $(\alpha,\delta,\mu_\alpha,\mu_\delta,\pi,RV)$.

The "D" parameter adopted in Krone-Martins & Moitinho (2014) is however not well-suited for the comparison of samples of different sizes (in a hunt for clusters, field stars are always expected to outnumber cluster stars), and produces ambiguous results when the reference distribution is not uniform (field stars can reasonably expected to be uniformly distributed in the

sky, but their proper motions, parallaxes and radial velocities are *not* well represented by a uniform distribution). As a result, even in cases where the method was efficiently able to sort cluster stars from field stars, the method was not able to automatically indicate overdensities in spaces other than sky coordinates.

A different approach based on a Minimum Spanning Tree (Allison et al. 2009, MNRAS, 395, 1449) was implemented and proved more efficient, as it is able to compare any two distributions regardless of the geometry of the reference distribution.

2.2) Application to observed data

Not having access to a full-sky 6D catalogue, we have worked in a 4D space by combining two all-sky catalogues: the **Tycho-2** catalogue which contains proper motions and *(B,V)* magnitudes, and the **2MASS** catalogue which contains *(J,H,Ks)* magnitudes. This combination constitutes a total sample of about 2.5 million stars. The sky was divided into 6643 overlapping tiles of 5°x5°.

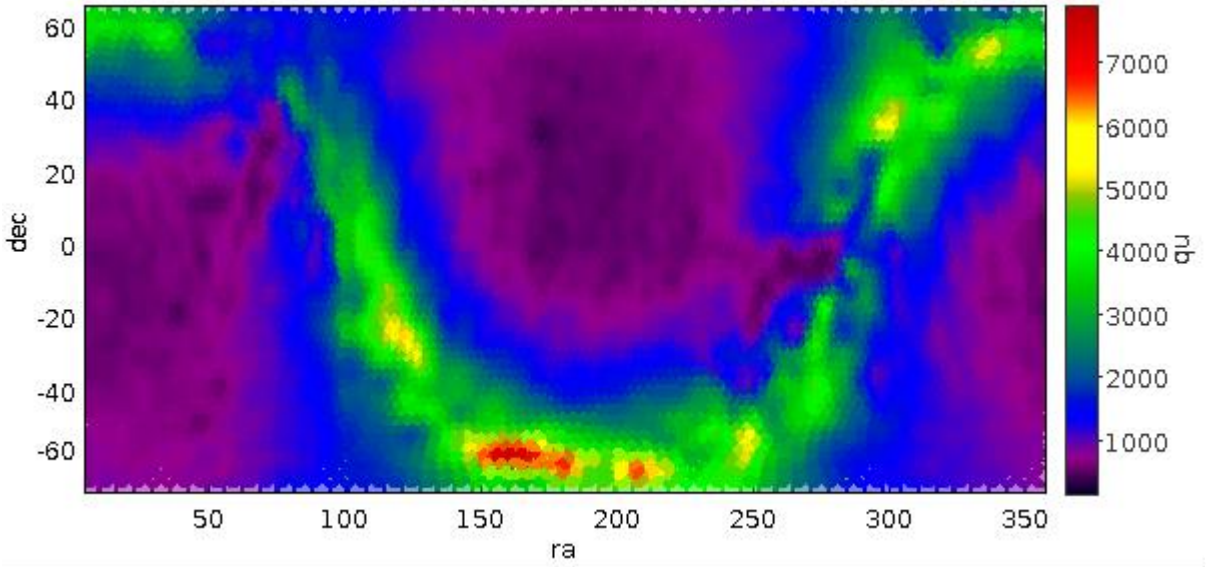

Fig. 1: number of stars in each tile as a function of the sky coordinates of the centre of the tile.

The number of stars in each tile varies from a few hundreds to more than 7000 in some dense fields towards the Galactic centre (see **Fig. 1**). Each tile was investigated following our method in order to tell potential cluster stars apart from field stars, and computing diagnostics such as the "sky clustering significance" and "proper motion clustering" of this identified "potential cluster" subset.

As described in Krone-Martins & Moitinho (2014), we use a k-means clustering method to build groups in the phase space. This method requires a random initialisation, which means it may identify slightly different groups if run several times. For this reason, we perform the analysis 10 times for each field of view under investigation. **Figure 2** shows the 1002 tiles for which a "potential cluster" was detected at least once. The colour code ranges from dark blue (1 potential detection) to dark red (10 potential detections).

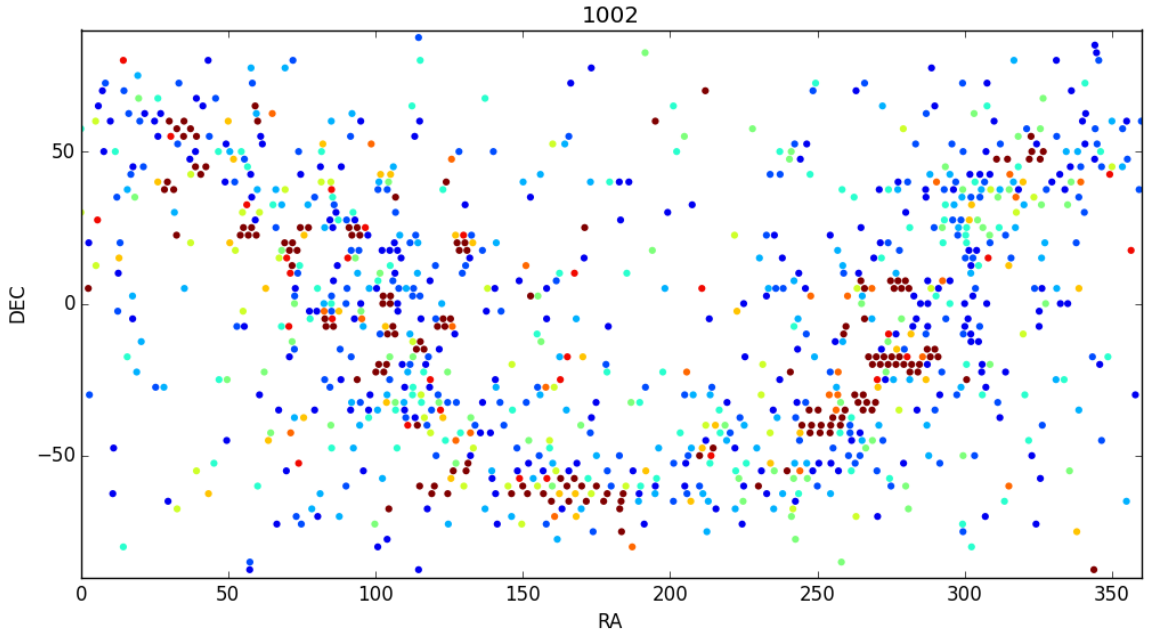

Fig. 2: position of all the tiles with at least one potential detection. The colour scale ranges from 1 (dark blue) to 10 detections (dark red).

For each one of these potential detections, the "sky clustering significance" and "proper motion clustering significance" is quantified (**Fig. 3**). We chose to discard as false positives all the potential detections with a sky significance under 1 (i.e. the identified group has to be more clumpy than the field by at least one sigma), and proper motion significance under 0 (the proper motions of the identified group can be as compact or more compact than the field, but not looser than the field).

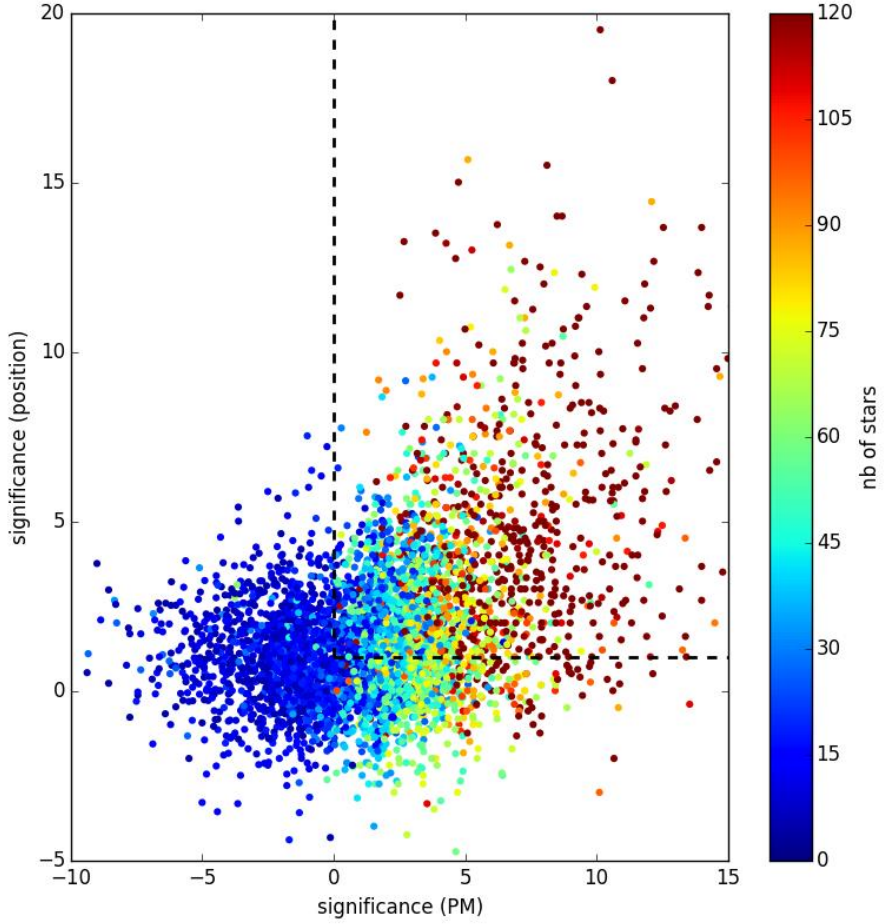

Fig. 3: significance in terms of proper motion clustering (x-axis) and sky clustering (y-axis) of all potential detections. Cases with sky position significance <1 and PM signifiifance <0 are considered false positives.

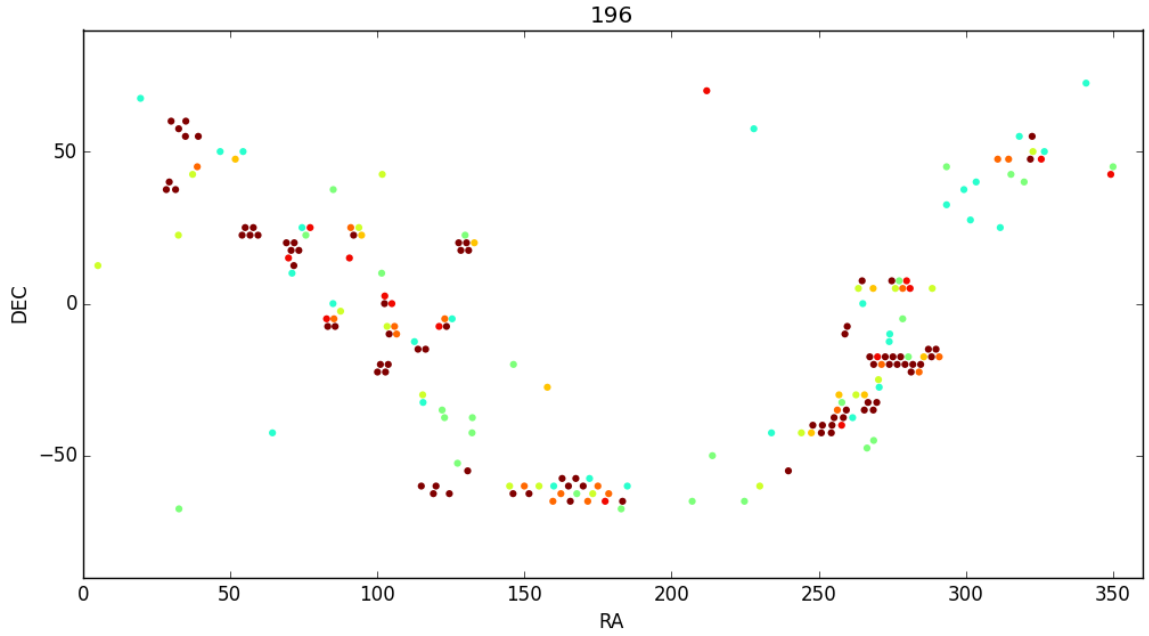

Fig. 4: sky coordinates of the 196 tiles with potential detection, after removing false-positives.

After operating this filtering, we end up with 196 remaining tiles that contain potential clusters (**Fig. 4**). At this stage, a visual inspection of those fields was performed in order to control the reality of those objects.

Figure 5 shows the example of a tile centred on the open cluster NGC2632. The cluster does not stand out as an overdensity. **Figure 6** shows that our method is capable of identifying a highly significant group of clustered stars. The "lambda" value of the sky clustering (by analogy with Allison et al., 2009) is 1.58±0.06 (i.e. a significance of 0.58/0.06=9.7) and the "lambda" in proper motions is 3.95±0.36 (significance 8.2).

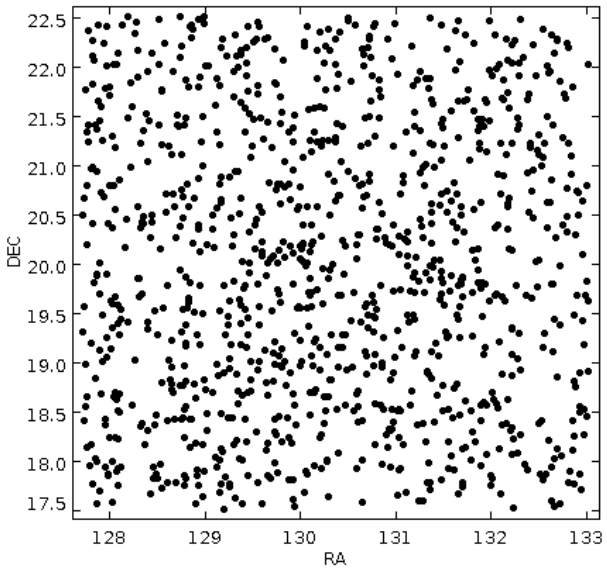

Fig. 5: coordinates of the stars in a tile centred on the open cluster NGC2632. The cluster does not appear as an overdensity. Figure 6 shows our method successfully disentangles the cluster from the field stars.

Diagnostic plots such as **Fig. 6** are created for all potential detection, facilitating eye control of the results.

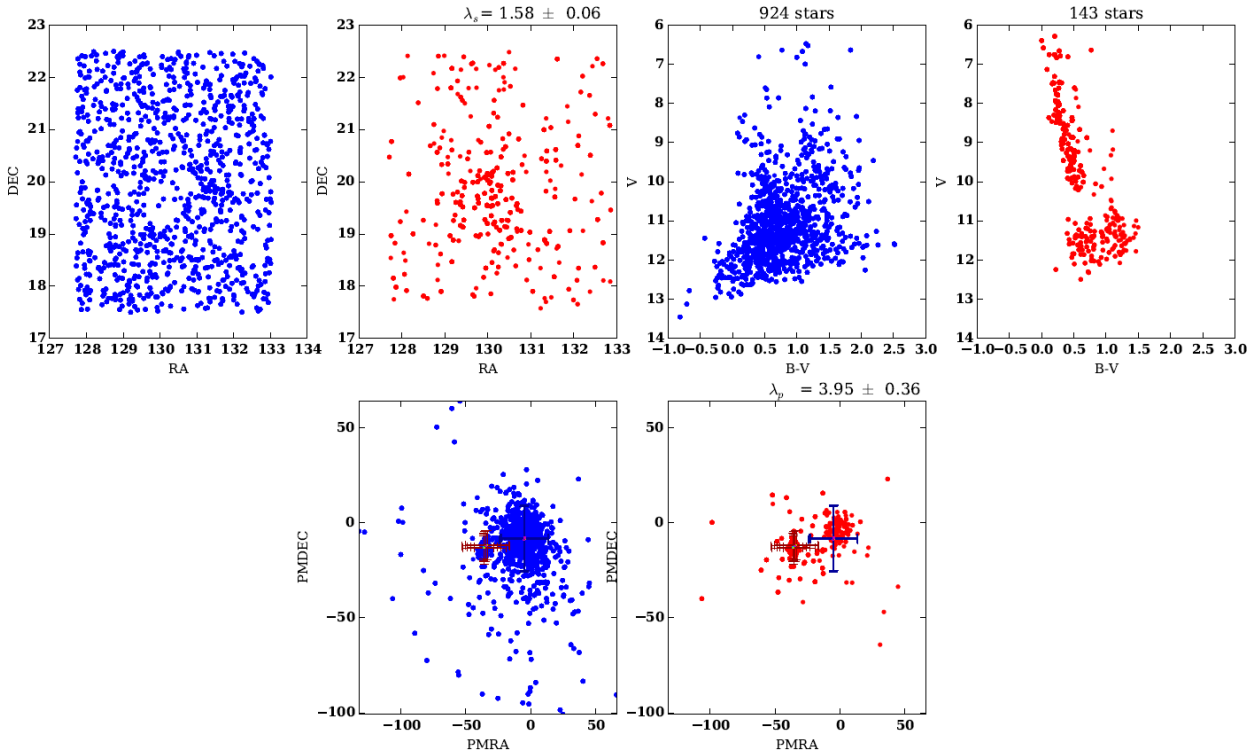

Fig. 6: sky coordinates (top-left), colour-magnitude diagram (top-right) and proper motions (bottom) for the stars identified as field (blue) and those identified as cluster (red).

3) Description of the main results obtained

The method is efficient at sorting stars into two categories (*field* stars and *potentially clustered* stars). The current thresholds imposed on the significance of the clustering (**Fig. 3**) still create a certain number of false positives, but drastically reduce the number of fields to investigate (here from 6643 to 196). Quantifying the significance also enables us to rank detections, and assign high priorities to fields with a stronger signal.

In its current implementation, the method detects groups of stars that show similar properties in most parameters, and separates them from the rest of the stellar population ("field"), and quantifies the statistical difference between this subset and the field stars. It produces figures that must be controlled by eye in order to confirm the reality of the object.

The diagnostic plots of all 196 tiles shown in **Fig. 4** were investigated by eye. The method successfully recovered 36 previously known clusters, in the distance range 200-1200pc. No previously unknown cluster was detected.

4) Ongoing work and future collaboration with host institution

A manual classification of the results was possible here, because of the relatively low number of tiles to control (196 fields). This will not be feasible when investigating the Gaia catalogue. In particular, the method should be able to:

- automatically link objects that have been detected in multiple tiles
- flag tiles containing multiple clusters

The first Gaia data release (which will contain all parameters except radial velocities) is planned for mid-2016. Our method will be able to investigate the whole sky. As a higher priority, a search could focus on the 3006 candidate clusters of Kharchenko et al. (2013, A&A, 558, 53) or the high-latitude associations detected by Schmeja et al. (2014, A&A, 568, 51). In addition to Gaia data, the

search can benefit from photometry from deep ground-based surveys (e.g. 2MASS, VVV, UKIDSS, VPHAS+).

5) Projected publications to result from the grant

A publication is planned, presenting the method and its application to the full-sky catalogue Tycho2 completed with 2MASS photometry. In addition, tests will be performed on mock Gaia data in order to assess the capabilities of the method in detection clusters of various ages and masses and quantify the contribution of the Gaia mission to the Galactic cluster census.

6) Other comments

The original version of UPMASK was implemented in R. The detection method developed here is based on T. Cantat-Gaudin's Python implementation of UPMASK. Taking advantage of numerical libraries such as NumPy and and sckikit-learn, the automated search is quite fast. Running the search on the full-sky **Tycho2 x 2MASS** catalogue, with a magnitude limit of V<10.5 (to avoid stars with poor proper motions) takes 3 hours on a single machine at the Lisbon computer cluster (the total cluster is made up of 12 such machines). The computing time in each investigated tile mostly scales linearly with the number of stars. A back of the envelope calculation suggests that using the full capacity of the Lisbon cluster, *half a billion* stars from the Gaia catalogue would take a very reasonable 10 days to explore.