Astronomical Data Analysis Software and Systems XXI
ASP Conference Series, Vol. 461
Pascal Ballester, Daniel Egret, and Nuria P. F. Lorente, eds.

Mining the UKIDSS GPS: Star Formation and Embedded Clusters

Otto Solin,[1,2] Esko Ukkonen,[1] Lauri Haikala,[3] and Sami Maisala[2]

[1]*University of Helsinki, Department of Computer Science, Finland*

[2]*University of Helsinki, Department of Physics, Division of Geophysics and Astronomy, Finland*

[3]*Finnish Centre for Astronomy with ESO, Finland*

Abstract. The aim of this research is to develop methods to locate previously unknown stellar clusters from the UKIDSS Galactic Plane Survey catalogue data release 7. The cluster candidates were computationally searched from pre-filtered catalogue data using a recently proposed method that fits a mixture model of Gaussian densities and background noise using the Expectation Maximization algorithm. The pre-filtering of the data involves both removing data artefacts and searching for sources classified as non-stellar due to associated surface brightness thus directing the search in particular to embedded stellar clusters. The findings were further screened by visual inspection of images, and SIMBAD was used to study sources in the direction of the candidates. The search covered an area of 1302 $\square^\circ$ and 111 previously unknown cluster candidates and 19 previously unknown sites of star formation were located.

1. Introduction

Major part of star formation, be it low- or high-mass stars, takes place in clusters. The clusters are not bound and will eventually disrupt e.g. because of the Galactic differential rotation. The stellar clusters trace therefore the recent Galactic star formation. The younger the clusters are the more compact they are and the more closely they are associated with the interstellar gas and dust clouds they formed in. Detailed study of young clusters still associated with their parent cloud will provide information on the star formation process and the stellar initial mass function (IMF).

At the moment some 2000 Galactic stellar clusters are known. This is only a small fraction of the estimated total population of which a major part is obscured by interstellar dust to us and can not be observed in optical wavelengths. However, the extinction decreases at longer wavelengths and already at 2.2 microns in the NIR the extinction in magnitudes is only 11 percent of that in the *V* band.

2. Search Method

The search method takes pre-filtered catalogue data, divided into overlapping bins of size 4′ by 4′, and performs a maximum likelihood fitting of a mixture of a Gaussian density and a uniform background. On each bin the fitting is done using the standard Expectation Maximization (EM) algorithm that is widely applied in a variety of sci-

ences, and generally for data clustering in machine learning. Our method is similar to the one used by Mercer et al. (2005) and Lucas (2011). It has so far been applied to the UKIDSS GPS DR7 covering an area of 1302 $\square^{\circ}$. In addition to the UKIDSS GPS catalogue, stars brighter than 10^{m} in K from the 2MASS survey are used, because the brighter stars saturate in UKIDSS and moreover tend to produce false positives around them.

Scrutiny of the data base and the survey images reveals that the pipeline source detection algorithm tends to classify most of the objects within regions of variable surface brightness as non-stellar (parameter `mergedClass` = +1), whereas objects with intensity profiles similar to the UKIDSS WFCAM point spread function are classified as star-like (`mergedClass` = −1). Clustering non-stellar sources directs the search to stellar clusters either embedded in or near molecular/dust clouds. Besides stellar clusters, the search targets also the locations of non-clustered star formation and single embedded stars with associated nebulosities. The surface brightness, either due to outflow activity or reflection, will produce 'cluster" detections.

A fraction of the catalogue sources are due to data artefacts. The artefacts cause highly varying extended surface brightness which causes the pipeline to classify most of the sources within the artefact as non-stellar sources. In addition sharp features in the artefacts produce nonexistent sources. The data artefacts must be filtered out from the data before the EM-algorithm is applied as otherwise too many false clusters due to artefacts will be located. The following artefacts have been addressed: Diffraction patterns of bright stars and diffraction spikes due to secondary mirror supports, bright stars at or near the border of the detector array, beams, 'bow-ties', cross-talk images and persistence images.

The classification of sources fainter than 17^{m} in K as star/non-stellar objects is highly unreliable. These sources were filtered out from the data. UKIDSS DR7 contains 631 117 002 sources measured in the K filter. Out of these 343 737 754 i.e. 54% satisfy the criteria K magnitude brighter than 17^{m} and `k_1ppErrBits` < 524288. These sources are divided according to the `mergedClass` parameter so that a negligible fraction are probable galaxies or noise, 5% probable stars, 74% stars, and 19% galaxies. We end up using for the detection algorithm sources with K magnitude brighter than 17^{m}, `k_1ppErrBits` < 524288 and `mergedClass` = +1. These amount to 66 149 194 sources ($\sim$ 10% out of all sources in UKIDSS DR7).

The automated search proceeds in the following steps. Only the K band data is used in the search.

1. The pre-filtered catalogue data is divided into smaller overlapping spatial bins of size $4'$ by $4'$. Apart from bins at the edges each bin overlaps one half of its neighbouring bins.

2. Remove false `mergedClass` = +1 classifications around bright stars and in the direction of the 8 diffraction spikes.

3. In order to track clusters with bright members the detection algorithm is run five times: once with all (filtered) input data and then using 80, 60, 40 and 20% of these sources arranged in descending order of the K magnitude.

4. The spatial coordinates are rescaled to the interval [0,1] to make all bins equally important but still allowing them to have differing means and variances. This step

is relevant only for bins at the edges of the survey and which are smaller than 4′ by 4′.

5. In order to initialize the model parameters (μ, Σ, τ) the data bin is divided into 16 subgrids to find the area with the highest spatial density. The initial value of the cluster mean μ is the center point of the subgrid with the highest density. The covariance matrix of the data points assigned to the subgrid with the highest density give the initial values for the cluster covariance Σ. The weights τ have as initial values the same value: $\tau_0 = \tau_1 = 0.5$. Coefficient τ_0 gives the proportion of stars belonging to the background and τ_1 gives the proportion of stars belonging to the cluster.

6. Each data bin is represented by a mixture model of a background component and one Gaussian cluster component.

7. The EM-algorithm returns for each data bin a candidate cluster, i.e. an ellipse with the center point at the mean μ and half-axes determined by the covariance Σ.

8. Remove false positives created by bright stars at or just outside a frame border.

9. Rearrange candidates in descending order of the BIC, and at the same time merge cluster candidates closer than one arcmin to each other.

10. Remove from the list the cluster candidates catalogued in Bica et al. (2003a,b) (165 in UKIDSS DR7), Mercer et al. (2005) (25 in DR7), Froebrich et al. (2007) (168 in DR7) and Lucas (2009) (331 cluster candidates from DR4).

3. Results

The search located 111 cluster and 19 non-clustered star formation location candidates which, to our knowledge, are previously unknown. An example is presented in Fig. 1. As expected most of the detected clusters or star formation locations are tightly concentrated on the Galactic plane. Relatively few clusters were detected in the direction of the northern Galactic plane. This possibly indicates that most of the northern clusters have already been discovered as this part of the plane has been more thoroughly investigated than the southern plane. Also, the interstellar extinction is less severe in the northern plane than in the south. However, some of the new northern clusters in the direction of the Galactic anticentre are massive and deserve to be investigated in more detail.

Most images of the new cluster candidate areas show clear signs of reflected light in particular in the *K* band thus indicating embedded clusters or sites of star formation.

SIMBAD was used to search for astronomical objects within 2′ from the candidates. An IRAS point source is seen in the direction of almost all new clusters and locations of star formation. Besides other indications of a still ongoing star formation (e.g. (sub)mm, MSX and maser sources) are detected in the direction or near a large part of the detected clusters. Only six cluster candidates and one embedded star formation candidate were not associated with any object in the SIMBAD data base. The number of indicators seen in the direction of the candidates gives confidence the new clusters or embedded star formation locations are real entities and not produced by chance nor are

due to catalogue artefacts. In general radio surveys find circumstellar dust envelopes and disks, and cold cores of molecular clouds. In areas where a radio telescope sees only a point source or signs of e.g. an ultracompact HII region, the UKIDSS images show structures of surface brightness and single stars thus verifying the results of the millimetre/submillimetre radio surveys of suspected star forming regions.

A workflow implementation of the search algorithm is presented in a poster at this conference (Maisala et al. 2012).

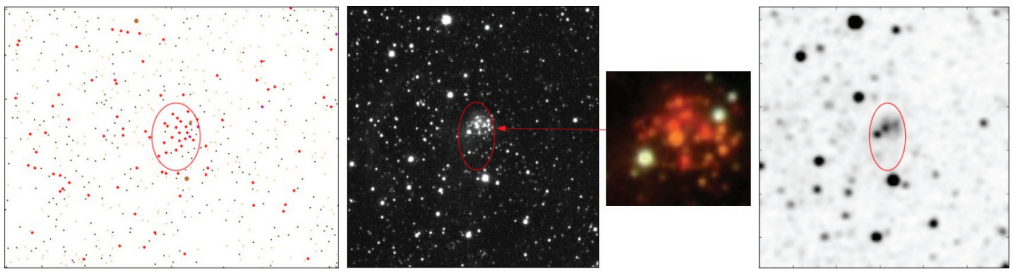

Figure 1. A cluster candidate at the location ($l = 52.367°, b = -1.044°$) identified previously as an infrared point source. In the leftmost panel are the UKIDSS catalogue entries in the cluster area. The red points are UKIDSS non-stellar sources brighter than 17^{m} in K, black points other sources brighter than 17^{m} in K, yellow points sources fainter than 17^{m} in K, and brown points sources listed in 2MASS but not in UKIDSS GPS. The red confidence ellipse is the cluster area given by the EM-algorithm. In the two middle panels are the K band and JHK false colour images of the cluster area. In the 2MASS image (the rightmost panel) of the same area no cluster can be seen.

Acknowledgments. This work was funded by the Finnish Ministry of Education under the project "Utilizing Finland's membership in the European Southern Observatory". This work was supported by the Academy of Finland under grant 118653 (ALGODAN), and by the Finnish Funding Agency for Technology and Innovation (TEKES) under the project MIFSAS.

References

Bica, E., Dutra, C. M., & Barbuy, B. 2003a, A&A, 397, 177
Bica, E., Dutra, C. M., Soares, J., & Barbuy, B. 2003b, A&A, 404, 223
Froebrich, D., Scholz, A., & Raftery, C. L. 2007, MNRAS, 374, 399
Lucas, P. W. 2009, 331 cluster candidate images with central coordinates. URL `http://star-www.herts.ac.uk/~pwl/Lucas/clusters`
— 2011, in Science from UKIDSS III. URL `http://wiki.astrogrid.ac.uk/pub/UKIDSS/Jan11Workshop/Lucas-GPS.pdf`
Maisala, S., Oittinen, T., Takala, T., Huovelin, J., & Solin, O. 2012, in ADASS XXI, edited by P. Ballester, D. Egret, & N. P. F. Lorente, vol. 461 of ASP Conf. Ser., 99
Mercer, E. P., et al. 2005, ApJ, 635, 560