Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **458,** 3479–3488 (2016) doi:10.1093/mnras/stw398
Advance Access publication 2016 February 22

An all-sky support vector machine selection of *WISE* YSO candidates

G. Marton,[1★] L. V. Tóth,[2] R. Paladini,[3] M. Kun,[1] S. Zahorecz,[2,4] P. McGehee[3] and Cs. Kiss[1]

[1]*Konkoly Observatory, Research Centre for Astronomy and Earth Sciences, Hungarian Academy of Sciences, H-1121 Budapest, Hungary*
[2]*Department of Astronomy, Loránd Eötvös University, Pázmány P.s. 1/a, H-1117 Budapest, Hungary*
[3]*Infrared Processing Analysis Center, California Institute of Technology, 770 South Wilson Ave., Pasadena, CA 91125, USA*
[4]*European Southern Observatory, Karl-Schwarzschild-Str. 2, D-85748 Garching bei München, Germany*

Accepted 2016 February 18. Received 2016 February 9; in original form 2015 July 29

ABSTRACT

We explored the AllWISE catalogue of the *Wide-field Infrared Survey Explorer (WISE)* mission and identified Young Stellar Object (YSO) candidates. Reliable 2MASS and *WISE* photometric data combined with *Planck* dust opacity values were used to build our data set and to find the best classification scheme. A sophisticated statistical method, the support vector machine (SVM) is used to analyse the multidimensional data space and to remove source types identified as contaminants (extragalactic sources, main-sequence stars, evolved stars and sources related to the interstellar medium). Objects listed in the SIMBAD data base are used to identify the already known sources and to train our method. A new all-sky selection of 133 980 Class I/II YSO candidates is presented. The estimated contamination was found to be well below 1 per cent based on comparison with our SIMBAD training set. We also compare our results to that of existing methods and catalogues. The SVM selection process successfully identified >90 per cent of the Class I/II YSOs based on comparison with photometric and spectroscopic YSO catalogues. Our conclusion is that by using the SVM, our classification is able to identify more known YSOs of the training sample than other methods based on colour–colour and magnitude–colour selection. The distribution of the YSO candidates well correlates with that of the *Planck* Galactic Cold Clumps in the Taurus–Auriga–Perseus–California region.

Key words: methods: data analysis – methods: statistical – stars: pre-main-sequence – stars: protostars – infrared: general – infrared: stars.

1 INTRODUCTION

The amount of data collected by infrared (IR) satellites and observatories has been continuously increasing over the past three decades. The evolution of the detectors allowed us to explore the interstellar medium (ISM) and embedded objects in more and more detail. *IRAS* (Neugebauer et al. 1984) provided ∼350 000 objects with flux above 0.5 Jy at 12 μm. Recently, based on observations of the *Wide-field Infrared Survey Explorer* (*WISE*; Wright et al. 2010) IR satellite, more than 700 million sources with $>5\sigma$ accuracy above 1 mJy were catalogued in the AllWISE Data Release (Cutri et al. 2013). IR luminous objects cover a broad spectrum of object types. Extragalactic sources, especially galaxies with ongoing star formation or active galactic nuclei (AGNs), show similar spectral energy distribution (SED) to that of the Young Stellar Objects (YSOs). Evolved stars eject dust into their outer envelopes, which has IR colours analogous to the dust surrounding YSOs in their early evolutionary stages. In this work, we identify the AllWISE sources by searching for a close counterpart in the SIMBAD data base, using a 5 arcsec radius. The closest SIMBAD source within this radius is associated with the AllWISE object. Fig. 1 illustrates the surface density of the known extragalactic sources, main-sequence (MS) stars, evolved stars and YSOs in the *WISE* $W1$–$W2$, $W2$–$W3$ colour–colour plane and it shows that the different object types have highly overlapping *WISE* colours. Linear methods would fail to separate the different object types and result in samples with high percentage of contamination, so the separation requires special attention.

The complexity of the observable properties makes the object classification a fundamental and challenging problem. However, the commonly used schemes do not take advantage of all the available information. Sources are often identified on colour–colour and colour–magnitude diagrams: Gutermuth et al. (2008) characterized the *Spitzer* (Werner et al. 2004) IRAC (Fazio et al. 2004) colour and magnitude properties of proto- and pre-MS stars in NGC 1333, and completed the data set with MIPS (Rieke et al. 2004) and J, H and K_s 2MASS (Cutri et al. 2003; Skrutskie et al. 2006) data. This method was then extended by Gutermuth et al. (2009) to

[★]E-mail: marton.gabor@csfk.mta.hu

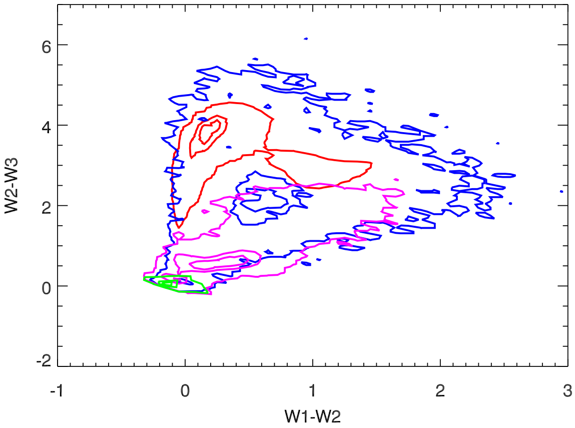

Figure 1. Object types identified with SIMBAD on the *W*1–*W*2, *W*2–*W*3 plane. The figure demonstrates that different object types are overlapping, boundaries between them are non-linear. Contour lines show the 5 per cent, 50 per cent and 75 per cent of the maximal surface density of extragalactic sources (red), field stars (green), evolved stars (magenta) and YSOs (blue). The surface density of different types was calculated in bins of 0.1 mag.

several star-forming clouds, located within 1 kpc of the Sun. Rebull et al. (2010) also used IRAC and MIPS colours together with 2MASS data to identify YSOs in the Taurus Molecular Cloud (TMC). They note that the method does not seem to successfully weed out all the galaxies. Harvey et al. (2007) set criteria for YSO identification on large-scale maps of star-forming regions, with a primary goal of mitigation of extragalactic contamination. This kind of large-scale classification is problematic because of the wide variation of extinction. Rebull et al. (2011) also identified YSO candidates in the TMC by using *WISE* photometry, as described in Koenig et al. (2012). By using far-IR (four bands between 65 and 160 μm) *AKARI* (Murakami et al. 2007) *FIS* (Kawada et al. 2007) colours and flux densities, Pollo, Rybka & Takeuchi (2010) successfully separated the sources of the *AKARI* Bright Source Catalogue (Yamamura et al. 2010) in low-extinction regions by classifying them as either extragalactic sources or Milky Way stars. Based on a combination of far-IR *AKARI* and mid-IR *WISE* data, Tóth et al. (2014) used quadratic discriminant analysis (QDA) to identify YSO candidates. Their comparison to the known YSOs of the SIMBAD data base showed that 90 per cent of the training sample YSOs were successfully reclassified as YSO candidates, while the fraction of known contaminants remained <10 per cent.

In this paper, we built a multidimensional data set containing near-IR 2MASS and mid-IR *WISE* data and we apply the support vector machine (SVM) method to identify potential YSO candidates. Our goal is to create a catalogue of carefully selected YSO candidates that can be used for statistical studies, and that can also provide a list of potential targets for future follow-up observations. We show that SVM (Vapnik 1995), which is a commonly used tool in pattern recognition and in multidimensional classification, is able to identify higher fraction of the training sample YSOs than the regularly used polygonal selections on colour–colour and colour–magnitude planes. We have to note that, due to the lack of spectroscopic data, our training samples are based on SIMBAD identifications. Therefore, most of our comparisons are also estimates based on SIMBAD. We identify extragalactic contaminants, Galactic field stars, Galactic evolved stars and other Galactic contamination. As a result, YSO candidates are identified and an attempt is made to separate them based on their evolutionary stages. Verification of our method and comparison to existing methods of YSO selection has been done. We also investigate, in a well-known star-forming region (i.e. the Taurus–Auriga–Perseus–California molecular cloud), the potential spatial correlation between the candidate YSOs and the Cold Clumps listed in the Planck Catalogue (Planck Collaboration 2015).

2 DATA AND METHOD

2.1 Data

To search for the YSO candidates, we used the AllWISE Data Release (Cutri et al. 2013), which is an improved version of the *WISE* All-Sky Data Release (Cutri et al. 2012). The AllWISE catalogue contains information on 747 634 026 sources. We used not only brightness values in all four *WISE* passbands (3.6, 4.6, 12 and 22 μm), but also the extended source flag (*ext*), which indicates whether or not the morphology of a source is consistent with the *WISE* point spread function, and also if the source is associated with or superimposed on a previously known extended object from the 2MASS Extended Source Catalog. We also used 2MASS J, H, K_s magnitudes, which are provided in the AllWISE catalogue, based on 2MASS Point Source Catalogue (Cutri et al. 2003) associations. Instead of the whole AllWISE catalogue, we used only those sources that matched the following criteria: (1) signal-to-noise ratio (SNR) > 3 in all four *WISE* bands and (2) 2MASS J, H, K_s magnitudes are available with photometric errors lower than 0.1 mag. Applying these criteria resulted in 8956 636 sources. These form our initial sample, which hereafter we call the *W0* sample.

At the position of each source, we estimated the dust optical depth (τ) by using the 353 GHz R1.2 Planck dust opacity maps (Planck Collaboration XI 2014). This operation allowed us to include the effect of interstellar reddening in our analysis.

2.2 Support vector machines

For classification and pattern recognition in multidimensional data, one can use several statistical methods. We used the SVM, a class of supervised learning algorithm, developed by Vapnik (1995) as an extension to non-linear models of the generalized portrait algorithm. SVM calculates decision planes between different known classes of objects and applies the decision planes to objects of unknown classes. These unknown objects are classified based on their position in the multidimensional parameter space with respect to the separation boundaries. A more detailed description of the method can be found in Małek et al. (2013). Various statistical methods are usually among the major ingredients of professional statistical software packages. We used the *R* implementation of SVM in our work.

SVM is a supervised learning algorithm, therefore it needs a training set, which is used to determine the boundaries in the parameter space between the different object types. To find out the object types of our sources, we searched the SIMBAD data base for counterparts within 5 arcsec radius. If more than one object was located within this radius, then the type of the closest entry was used. More information about the SIMBAD object identification can be found in Ochsenbein & Dubois (1992). Following this strategy, we were able to identify 890 552 sources of the *W0* sample. We are aware that individual source identifications in SIMBAD might not be fully reliable. However, for our purposes, we only need training samples that are statistically reliable. The number of sources found per object type are listed in the *W0* column of Table A1.

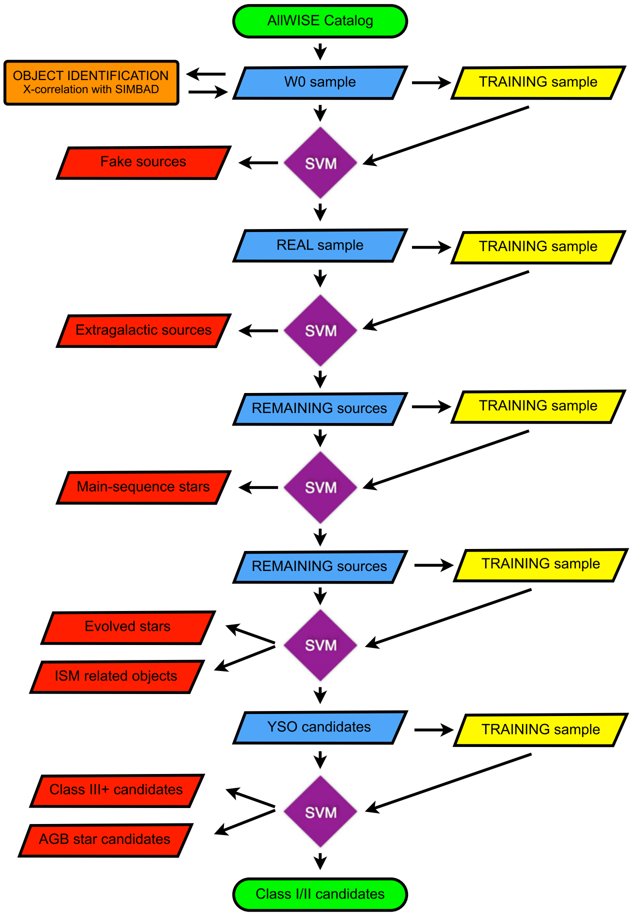

Figure 2. Steps from the AllWISE catalogue to the final Class I/II candidate selections. Steps are described in details in the corresponding subsections.

Unless noted otherwise, in the remaining of this paper the expression 'known object' always refers to a SIMBAD identification. The average colours and magnitudes for each SIMBAD object type in our *W0* sample are listed in Tables B1–B9.

Because the number of known sources from the different object types was very inhomogeneous, e.g. much more known extragalactic objects and field stars were found than YSOs, we checked if the classification is affected by the number of sources used in the training samples. To this end, we considered two cases: (a) we used all objects of given object types, (b) we used a maximum number of 1000 sources of each object types. We found that the number of false positives and false negatives differs by less than 1 per cent between case (a) and (b). Therefore, in our classification scheme, the number of objects used in the training sample was always limited to 1000, allowing us to speed up our classification process.

3 CLASSIFICATION STEPS AND RESULTS

A multistep classification scheme was developed to remove the contaminating sources from our sample, and to identify the YSO candidates with high accuracy. The steps of our classification are shown in Fig. 2. Each step and the corresponding training samples are described in detail below. Although SVM is a very powerful tool, the necessity of a multistep classification scheme can be explained by considering the complex structure of the ISM. The interstellar reddening has an impact on the apparent colours of the sources, therefore it is important to take into account where different object types are typically found as a function of ISM column density. For this purpose, and by using the Planck dust opacity maps, we binned our sources according to the dust opacity value registered at their position in the sky.

3.1 Spurious source identification

Koenig & Leisawitz (2014) performed a careful examination of the AllWISE catalogue: they inspected, at the positions of the AllWISE sources, the higher resolution *Spitzer* images. In addition, in selected regions of the sky, they compared *Spitzer* source catalogues to AllWISE lists of sources. Their analysis led them to conclude that several AllWISE catalogued sources are spurious, and that many are likely ISM knots, or *cirrus*. Following their finding, our very first step was to identify and remove the spurious sources from our *W0* sample.

To train the SVM, we checked the *WISE W*3 and *W*4 images of *W0* positions in five different regions: the Galactic Centre, the California Molecular Cloud, the Galactic anticentre, the ρ Oph star-forming region and the Cepheus molecular complex. In these regions, we selected 680 positions where we were able to clearly identify a point source, and 664 positions where visual source identification was not possible. Examples for these real and spurious AllWISE detections are shown in Fig. C1 and Fig. C2, respectively. The training samples included, for these sources, the AllWISE catalogue information on the SNR, the reduced χ^2 value, the number of times the source was detected with SNR > 3 and the number of profile fits in the *W*3 and *W*4 bands.

With the help of the training sample described above, we classified the *W0* sample into two classes: real and spurious sources. The misclassification rate, i.e. the rate of false positive and false negatives, was investigated as a function of the number of elements used in the training sample. Fig. 3 shows that the rate of misclassified sources does not change significantly with the number of elements used. The ratio of false negatives was 7.2 ± 1.4 per cent in the test. The lowest misclassification rate was achieved with the maximum number of elements in the training sample (5.7 per cent). The fraction of false positives was found to be 1.8 ± 0.6 per cent. With the maximum number of elements in the training sample, it was 1.7 per cent. The spurious and real sources were classified in their own class with 98.3 per cent and 94.3 per cent success rate, respectively. Applying the determined classification boundaries to

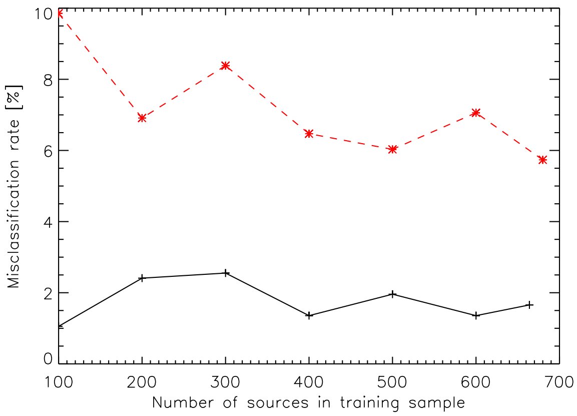

Figure 3. The rate of false positive (black solid line) and false negative (red dashed line) as a function of elements used in the training sample. On average, only 1.8 per cent of the spurious sources were classified as a real source.

our initial sample, we classified 5366 238 sources as spurious and 3 590 398 sources as real. The latter constitutes what we refer to as *real* sample.

3.2 YSO identification process

The SIMBAD data base lists 235 object types. This large variety makes the training of the algorithm rather inefficient. To make the algorithm more powerful and to have statistically more robust training samples, we binned the object types based on similarities in their $J-H$, $H-K_s$, $K_s - W1$, $W1-W2$, $W2-W3$, $W3-W4$ colours and on the *ext* (extended source) parameter. Figs D1–D4 show the distribution of the average colours and the average *ext* values.

Three large groups of SIMBAD extragalactic objects were created. SIMBAD types belonging to the G1 group have low $H-K_s$ and $W1-W2$ colours and have mostly high *ext* values. Source types of G2 group are less extended based on the *ext* value and have high $W1-W2$ values. The remaining sources were classified as G3, and are mostly extended like the G1 objects but have higher colour indices at shorter wavelengths.

Evolved stars were also grouped in three large bins. E1 type objects have rather small colour indices and are more compact, while E2 types have high $J-K$ colours compared to the other object types. E3 objects have $W2-W3$ colour higher than all the other evolved types.

Two bins of young SIMBAD objects were created: All colour indices of Y1 sources are higher than those of Y2 objects. Y1 sources also appear to be less compact because they have higher *ext* values.

The SIMBAD type '*' (single star) was not further binned.

Finally, two groups of SIMBAD source types of ISM-related objects were defined. ISM1 sources appear to be more compact than ISM2.

The 11 subtypes are listed below, including all SIMBAD types associated with each of them:

(i) G1: Galaxy, Part of a Galaxy, Galaxy in Cluster of Galaxies, Brightest Galaxy in a Cluster, Galaxy in Group of Galaxies, Galaxy in Pair of Galaxies, radio Galaxy, H II Galaxy, Low Surface Brightness Galaxy, Emission-line galaxy, Starburst Galaxy, Blue compact Galaxy, LINER-type Active Galaxy Nucleus.

(ii) G2: Active Galaxy Nucleus, Seyfert 2 Galaxy, Blazar, Seyfert Galaxy.

(iii) G3: Broad Absorption Line system, Gravitationally Lensed Image, Gravitationally Lensed Image of a Quasar, Seyfert 1 Galaxy, Quasar, Absorption Line system, Damped Ly-alpha Absorption Line system, Possible Quasar, BL Lac - type object.

(iv) Y1: Young Stellar Object, Variable Star of FU Ori type, Young Stellar Object Candidate.

(v) Y2: Variable Star of Orion Type, T Tau-type Star, T Tau Star Candidate.

(vi) E1: Horizontal Branch Star, S Star, Red Giant Branch star, Possible Carbon Star, Possible S Star, Carbon Star, Yellow supergiant star, Asymptotic Giant Branch Star, Evolved supergiant star, Variable Star of Mira Cet type, Possible Yellow supergiant star, Possible Horizontal Branch Star, Possible Red supergiant star, Red supergiant star.

(vii) E2: Possible Supergiant star, Possible Asymptotic Giant Branch Star, OH/IR star.

(viii) E3: Post-AGB Star, Post-AGB Star Candidate.

(ix) S: Single star.

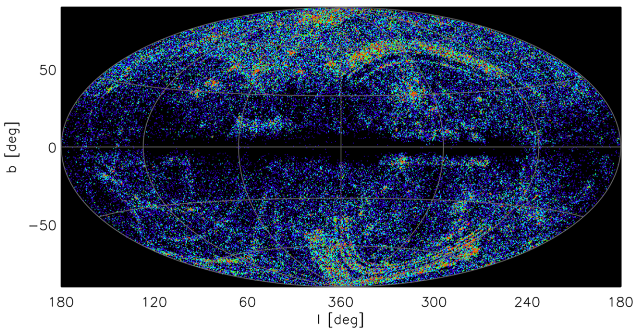

Figure 4. Surface density of extragalactic sources identified with SIMBAD, shown in galactic equal-area Aitoff projection. The direction of the Galactic mid-plane is almost completely galaxy-free. Values were calculated in $0.^{\circ}5 \times 0.^{\circ}5$ bins, and are represented on linear scale from 0 to 4.

(x) ISM1: Interstellar matter, HI shell, High-velocity Cloud, Emission Object, Planetary Nebula, Possible Planetary Nebula, SuperNova Remnant, SuperNova Remnant Candidate, Dark Cloud (nebula), Cloud, Part of Cloud, Outflow.

(xi) ISM2: Molecular Cloud, HII (ionized) region, Galactic Nebula, Globule (low-mass dark cloud).

3.2.1 *Removal of extragalactic sources*

First, the *real* sample was analysed with the goal of removing the extragalactic sources. We were able to identify 105 564 known extragalactic sources, including all SIMBAD object types that belong to the Galaxy type.[1] As seen in Fig. 4, their surface density distribution in the sky is very inhomogeneous. Regions close to the Galactic mid-plane and around known giant molecular clouds, like Orion (at $l \simeq 200$, $b \simeq -10$) or Cepheus (at $l \simeq 110$, $b \simeq 10$) are almost free of extragalactic sources. These regions contain high amount of ISM compared to the surroundings, therefore the extragalactic source distribution is biased in these regions. On one hand, the ISM is opaque at visual and near-IR wavelengths. Most of the SIMBAD extragalactic counts were made in the visual and near-IR regime, resulting in incomplete catalogues. On the other hand, IR-bright ISM features are able to cover fainter extragalaxies, making them undetectable. Third, the interstellar reddening modifies the apparent colours of the background objects. We found that 90 per cent of the extragalactic sources are located in regions where $\tau < 1.25 \times 10^{-5}$ and only 1 per cent are located in regions where $\tau > 5.05 \times 10^{-5}$ (see Fig. 5). 9 percent are located in regions between the two values. After investigating which colours and brightness values provide the highest separation between the extragalactic sources and all the other object types, we found that our trainer would provide the best separation if the $J-H$, $J-W4$, $K_s - W4$, $W2-W3$ colours, the J and $W2$ band magnitudes and the *ext* parameter are used. In each of these three regions (with characteristic τ values), we performed an SVM classifications using all subtypes listed in the previous section. Those sources classified as either G1, G2 or G3, were identified as extragalactic objects.

As a result, 377 126 sources have been classified as extragalactic sources (G1, G2 or G3), and were removed to build the *galaxy-free* sample. Out of the 105 564 already known extragalactic objects 105 376 were re-classified as extragalactic, and only 188 were misclassified (false negative). The *real* sample contained 5 685

[1] http://cds.u-strasbg.fr/cgi-bin/Otype?X

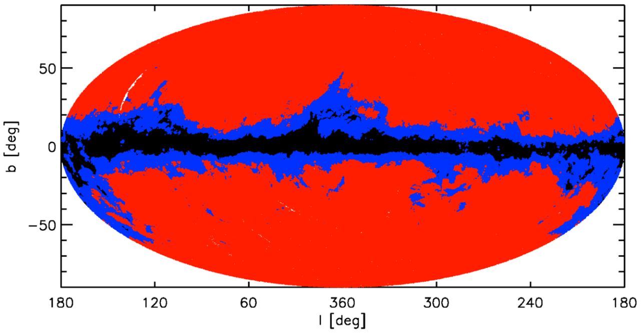

Figure 5. Regions of the sky containing 90 (red), 9 (blue) and 1 per cent (black) of the SIMBAD identified extragalactic sources. Regions are represented in galactic equal-area Aitoff projection.

known YSOs, of which only 117 were classified as extragalactic source (false positive).

3.2.2 Removal of MS stars

After removing the sources classified as extragalactic objects, our goal was to remove the MS stars from the remaining 3 213 272 sources. This *galaxy-free* sample contained 932 531 known MS stars. As it was done in case of the galaxy removal, we checked again the τ distribution of the MS stars and divided the sky into three regions. Those regions contained one third of the known MS stars, having $\tau < 1.1 \times 10^{-4}$, $\tau > 2.8 \times 10^{-4}$ and in-between. Also (as in the previous step), the average colours and brightness values of the different subtypes were calculated to achieve a maximum possible separation between MS stars and all the other object types. Performing this task led us to use the colours $H-W4$, $K_s - W4$, $W1-W2$, $W2-W3$, $W3-W4$, the H and $W4$ band magnitude values and the *ext* parameter.

Our results showed that we successfully removed 90.1 per cent of the known MS stars of the *galaxy-free* sample, by classifying 2 052 410 sources as MS star. At the same time, only 327 of the known YSOs were classified as MS star, meaning that 92.5 per cent of the known YSOs of the *real* sample were still kept.

3.2.3 Classification into evolved stars, ISM-related objects and YSO candidates

In the next step, we wanted to divide the remaining (*galaxy- and MS star-free*) 1 160 862 sources into three main categories: evolved stars (E1, E2 and E3 subtypes), ISM-related objects (ISM1 and ISM2 subtypes) and YSO candidates (Y1 and Y2 subtypes), with the goal to keep sources classified as Y1 or Y2. We were not able to identify regions, where one of the object types was dominant, thus cuts in the τ value were not applied. The training sample was prepared by using $J-H$, $W2-W3$, $W3-W4$ colours, the $W2$ magnitude and the *ext* parameter.

This classification step resulted in losing only 210 of the known YSOs, while keeping 88.8 per cent of the known YSOs in the *real* sample. We also removed 11 589 of the known evolved stars. Compared to the number of known evolved stars in the *real* sample, 27.4 per cent of them were still present at this stage. We were also able to remove 62.3 per cent of the known ISM-related objects. The resulting YSO candidate sample (sources classified as Y1 or Y2) contained 751 628 sources.

3.2.4 Classification into YSO evolutionary classes

The last step was to separate the Class I, II and III sources (Lada 1987) with the ultimate goal of finding reliable Class I and Class II YSO candidates. Class I and II sources have a significant IR excess that originates from the dust in their circumstellar envelopes or protoplanetary discs. The Class III sources have IR colours more similar to MS stars, but still showing IR excess. The SIMBAD data base does not list information on the actual evolutionary stage of the known YSOs, therefore our training sample was prepared based on YSO catalogues from the literature, preferably listing the evolutionary classes. We used catalogues from the following papers: Allen et al. (2012), Billot et al. (2010), Chavarria et al. (2008), Connelley, Reipurth & Tokunaga (2008), Evans et al. (2009), Gutermuth et al. (2008), Gutermuth et al. (2009), Kirk et al. (2009), Koenig, Allen & Gutermuth (2008), Megeath et al. (2012), Rebull et al. (2011), Rivera-I. et al. (2011). We note that these papers and their classification methods are based on IR data, mainly obtained with the *Spitzer Space Telescope* (Werner et al. 2004). We also used Connelley & Greene (2010) and Fang et al. (2009), which are catalogues of spectroscopically confirmed YSOs, and Winston et al. (2007) and Winston et al. (2010) listing YSOs with X-ray data.

Based on these catalogues, we created a training sample that contained 247 Class I, 1925 Class II and 313 Class III objects (the latter category includes sources of Koenig et al. (2008), being stars in star-forming regions, but showing mostly photospheric colours). In order to simplify the classification, where listed, we considered Transition Disc and Flat SED objects as Class II sources.

The first attempt to classify our YSO candidates into evolutionary stages failed, and the contamination caused by asymptotic giant branch star candidates ('AB?' as listed in SIMBAD) was still high. Therefore, the following step was additionally performed: the SVM was trained by using three subtypes, the Class I/II, the ClassIII+ and the 'AB?' stars. Majority of the 'AB?' objects are those identified by Robitaille et al. (2008), and we disagree with their statement 'YSOs and AGB stars can be mostly separated by simple colour–magnitude selection criteria'. The $J-H$, $H-W3$, $W1-W2$, $W1-W4$ and $W2-W3$ colours were used along with the $W2$ magnitude and the *ext* parameter, to create our training sample.

As a final result, we classified 133 980 sources as Class I/II candidates. Their surface density distribution is shown on Fig. 6. Compared to the training sample of 247 Class I, 1925 Class II and 313 Class III objects, the resulting Class I/II candidate sample contains 240 of the known Class I, 1824 of the known Class II and 79 of the known Class III sources. This means that 95 per cent of the known Class I+II sources were successfully kept, while 74.8 per cent of the Class III sources were removed. Likewise, in this last step 63.7 per cent of the known AGB star candidates were also removed. Fig. 7 illustrates the robustness of our method, as it shows the significant overlap between the known extragalactic sources, the known field stars, and our samples classified as Class I/II candidates and Class III+ candidates. Our candidate catalogues are available via the VizieR service.

4 DISCUSSION

4.1 Reliability, false positives

We carefully investigated the possible contamination present in our Class I/II candidate sample. In our *real* sample, the number of sources identified with SIMBAD was 1151 956, including 5685 YSOs (sources with object types 'Y*O', 'TT*', 'Or*', 'FU*',

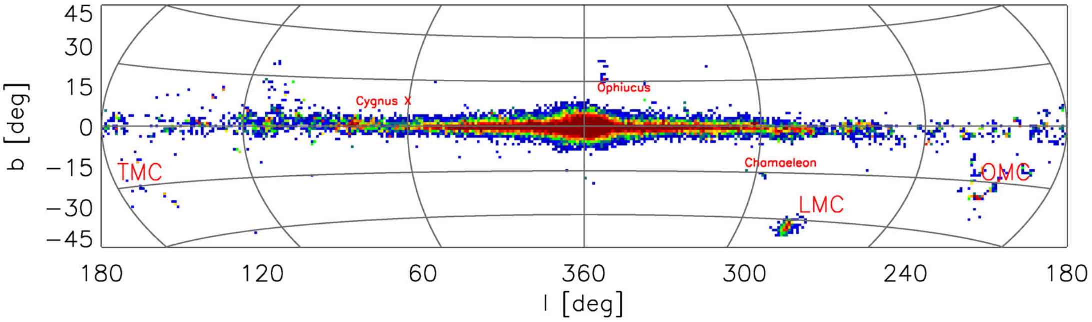

Figure 6. Surface density of the Class I/II YSO candidate sources classified with SVM, shown in galactic equal-area Aitoff projection. Values were calculated in 1° × 1° bins, and are represented on linear scale from 1 to 100. The major and best known star forming regions can be easily identified.

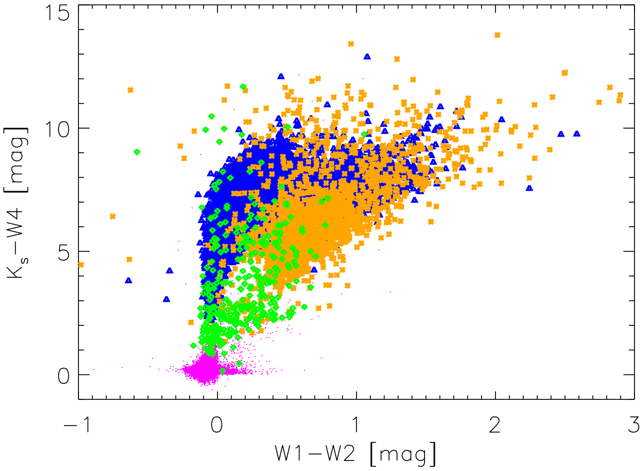

Figure 7. $W1-W2$ versus K_s-W4 colour–colour diagram with our Class I/II candidate sources (orange crosses) overlaid on sources classified as Class III+ (green diamonds), known MS stars from SIMBAD (magenta dots) and known SIMBAD extragalactic sources (blue triangles).

‘Y*?’ or ‘TT?’). We found only 21 568 (1.9 per cent) false-positive classifications that remained in our SVM classified Class I/II candidate sample. This means that we were able to remove 98.1 per cent of the contamination, as compared with the SIMBAD training set. The 21 568 false-positive sources were further analysed in order to learn what fraction of the contamination is coming from the different SIMBAD object types. The complete list is shown in column ‘SVM’ in Table A1. Here, we created four different categories to summarize our findings, representing groups of object types that are (i) most probably contamination, like known extragalactic objects or known evolved stars, (ii) candidate SIMBAD object types, (iii) sources that are assigned a generic object type in SIMBAD, such as IR source or star, but for which there is a non-zero probability to be instead YSOs and (iv) sources of flux that might be YSOs, or are closely related to them (e.g. maser).

(i) Obvious contamination.

(a) Extragalactic source – 101 (/105 564 – 0.1 per cent).
(b) Evolved star – 830 (/13 121 – 6.3 per cent).
(c) Other – 1955.

(ii) Possible contamination – candidate object type – 1128 (/7033 – 16 per cent). Here, we have to note that 931 of the 1128 candidate type objects are from the ‘AB?’ type. 925 of them are classified as ‘AB?’ by Robitaille et al. (2008).

(iii) Possible YSOs.

Table 1. Number of known sources in our samples of the selection process. First column indicates the name of the subtype (as defined in Section 3.2). Column 2, 3 and 4 are the *W0*, the *real* and the final Class I/II candidate samples.

Subtype	*W0*	Real	Class I/II
G1	148 267	90 840	65
G2	10 311	6 152	7
G3	12 729	8572	29
E1	13 208	12 940	631
E2	3429	2925	1118
E3	109	106	43
ISM1	912	745	183
ISM2	515	319	173
S	973 629	932 733	11 990
Y1	9268	4930	3705
Y2	1128	755	637

(a) Single star – 11 990 (/932 531 – 1.3 per cent).
(b) IR source – 1618 (/6701 – 24.1 per cent).
(c) Sources related to ISM – 413 (/1226 – 33.7 per cent).
(d) Variable star – 1424 (/16 537 – 8.6 per cent).

(iv) Possibly related to YSOs.

(a) Radio, mm, sub-mm or X-Ray source – 185 (/3089 – 6 per cent).
(b) Maser 49 (/71 – 69 per cent).

As it is shown in the list, the total number of obvious and possible contamination is very low in the candidate sample, 4013 (0.35 per cent) sources in total. The number of obvious contaminants only, is even lower, 2886 sources (0.25 per cent).

We also made a comparison based on the binned SIMBAD subtypes defined in Section 3.2. Table 1 details the number of elements for each subtype, found in the samples indicated by the table columns. We notice that only 101 sources (0.1 per cent) of the galaxy subtypes (G1, G2 and G3) are still present in the final Class I/II sample. The remaining number of evolved stars was found to be 1792 that is 11.2 per cent of the total E1+E2+E3 subtypes in the *real* sample. We note again that 925 of the 1792 (52 per cent) are asymptotic giant branch star candidates (‘AB?’) of Robitaille et al. (2008). Without these objects, the total number of E1+E2+E3 sources would be only 867. The fraction of remaining single stars (as listed in SIMBAD) is only 1.3 per cent. This small fraction corresponds to 11 990 sources, which is higher than the number of SIMBAD YSOs, but it is also a very generic object type that does not prevent these sources from being YSOs in reality. The number of objects from ISM-related types (ISM1+ISM2) is 356, corresponding to 33.5 per cent of the *real* sample. We emphasize once again that the SIMBAD associations are rather generic, therefore some of the sources might actually turn out to be YSOs.

The Sloan Digital Sky Survey DR-9 (SDSS DR-9; Ahn et al. 2012) flags the object types for all their sources indicating whether the source is thought to be a galaxy or a star based on morphology. A cross-correlation with SDSS DR-9 allowed us a different estimation on the fraction of false-positive classifications. Of the SIMBAD YSOs in the *real* sample, 1029 of the 5685 known YSOs have a counterpart in SDSS DR-9 (using a searching radius of 5 arcsec). 14 per cent of them (144) were flagged as galaxy in their catalogue. We also cross-checked the known YSOs in our final Class I/II candidate sample. 814 of them were found in the SDSS of which 106 (13 per cent) are flagged as galaxy. Out of the total 133 980 sources classified as Class I/II, 5840 were found in the SDSS catalogue, 580 of them are flagged as galaxy (9.9 per cent). We conclude that our final SVM sample of candidate Class I/II does not contain a higher faction of extragalactic contaminants than the sample of known YSOs in SIMBAD.

4.2 Completeness, false negatives

In the process of identifying the Class I/II sources candidates, a number of known YSOs were lost by either misclassification (false negatives) or because they were not detected by *WISE*. The completeness of our sample was investigated in three different ways. (i) First, we searched our *W0*, *real* and final Class I/II sample for the known YSOs in SIMBAD; (ii) then we looked, in the same samples as in (i), for YSOs listed in public photometric catalogues (see Section 3.2.4) (iii) finally, again using the samples as in (i), we looked for spectroscopically confirmed YSOs.

4.2.1 Comparison to SIMBAD YSOs

First, we considered all the known SIMBAD YSOs and searched for them in our *W0*, *real* and final Class I/II samples. The total number of known YSOs in the SIMBAD data base was 46 453 at the time of our investigation (21 186 ‘Y*O’, 20 716 ‘Y*?’, 1831 ‘TT*’, 237 ‘TT?’, 33 ‘FU*’ and 2450 ‘Or*’). After cross-correlating with SIMBAD, we found a total number of 10 396 known YSOs (4496, 4039, 733, 61, 20 and 1047, respectively) in our *W0* sample – this represents the number of known YSOs observed by *WISE* and detected above S/N > 3 in all *WISE* bands with 2MASS photometric errors below 0.1 mag. After the spurious source identification, our *real* sample contained 5685 of the known YSOs. The number of known YSOs in the final Class I/II sample is 4342. Based on this comparison, the completeness of our Class I/II sample is 9.3 per cent compared to the total number of YSOs known to SIMBAD. The completeness is 40.4 per cent compared to the *W0* sample, and 61.7 per cent compared to the *real* sample. The distributions for the normalized magnitude of the known YSOs in the *W0*, *real* and Class I/II samples are shown in Table E1. These plots reveal that the known YSOs lost in the selection process are mainly located at the faint end of the magnitude distributions, suggesting that our selection method is more sensitive to the brighter YSOs.

4.2.2 Comparison to photometric YSO catalogues

The next step consisted in searching in our *W0*, *real* and Class I/II samples for YSOs listed in published photometric catalogues. A separate search in the literature was performed for Class I and II and Class III sources. The results are listed in Table 2. Due to the initial selection conditions, only 36.4 per cent of the literature Class I and II sources and 30.4 per cent of the literature Class III sources were found in the *W0* sample. The *real* sample contains 20.8 per cent of the Class I and II and 12.0 per cent of the Class III sources. Our final Class I/II contains 12.9 per cent of the literature Class I and II sources and only 3.3 per cent of the literature Class III sources. In addition, 52.6 per cent of the Class I and II sources in the *W0* sample are classified again as such, while 10.8 per cent of the Class III sources are erroneously classified as Class I/II. For the real sample, 92.8 per cent of the Class I and II sources were successfully classified, while 27.5 per cent of the Class III sources were re-classified as Class I/II.

4.2.3 Comparison to spectroscopic YSO catalogues

Finally, the steps described in Section 4.2.2 were repeated in the case of spectroscopically confirmed YSOs, which are mostly located in nearby star-forming regions. The results of the comparison are provided in Table 3. Only about 42 per cent of the spectroscopically confirmed YSOs fulfilled our initial W0 condition. This also shows the limitations of a *WISE*–2MASS selection,

Table 2. Comparison of our selection to existing photometry-based YSO catalogues. Column 1 lists the corresponding paper. Column 2 lists the region of the sky that was the subject of the study. Column 3 and 4 give the number of Class I/II and Class III YSOs listed in the paper. Column 5 and 6 list the number of Class I/II and Class III objects found in the *W0* sample. Columns 7 and 8, and columns 9 and 10 give the number of corresponding objects in the *real* and Class I/II sample.

	Region	Paper		W0		Real		SVM Class I/II	
		Class I/II	Class III	Class I/II	Class III	Class I/II	Class III	Class I/II	Class III
Allen et al. (2012)	Cepheus OB3	1135	1440	435	279	209	29	198	9
Billot et al. (2010)	Vul OB1	703	153	259	82	160	54	134	7
Chavarria et al. (2008)	S254-S258	252	210	63	42	18	10	16	0
Connelley et al. (2008)	–	220	–	51	–	47	–	44	
Evans et al. (2009)	–	942	79	294	37	214	28	204	4
Gutermuth et al. (2008)	NGC 1333	133	–	57	–	46	–	43	–
Gutermuth et al. (2009)	–	2548	–	926	–	549	–	518	–
Kirk et al. (2009)	Cepheus flare	128	13	97	10	76	5	69	1
Megeath et al. (2012)	Orion A and B	2284	329	1023	168	631	69	603	53
Rebull et al. (2011)	NAN complex	1149	112	498	98	264	84	217	1
Rivera-I. et al. (2011)	W3 molecular cloud	1566	–	312	–	57	–	55	–
Winston et al. (2007)	Serpens cloud core	115	22	46	2	31	2	22	1
Winston et al. (2010)	NGC 1333	54	41	36	11	34	6	31	3
Total		11 229	2399	4097	729	2336	287	2154	79

Table 3. Comparison of our selection to existing spectroscopic YSO catalogues. The first column list the papers used in the analysis. Column 2 lists the corresponding sky region. Column 3 gives the number of YSOs used from the paper. The fourth column lists the number of YSOs found in our *W0* sample. Column 5 and 6 list the number of YSOs classified as *real*, and further classified as Class I/II YSO candidate. (a) and (b) stands for the Classical T Tauri and weak-line T Tauri objects, respectively.

	Region	Paper	W0	Real	SVM Class I/II
Alcala et al. (2014)	Lupus	36	35	28	25
An et al. (2011)	Galactic Centre	35	1	1	1
Connelley & Greene (2010)	–	88	38	37	36
Cooper et al. (2013)	–	180	49	48	48
Erickson et al. (2015)	Serpens	63	33	22	18
Fang et al. (2009)	L1630N and L1641	330	183	106	100
Kumar et al. (2014)	Carina nebula	241	23	2	1
Kun et al. (2009)	Cepheus flare	77	68	68	63
Mooley et al. (2013)	Taurus–Auriga	13	10	9	4
Oliveira et al. (2009)	Serpens	58	49	46	40
Rebollido et al. (2015)	ρ Oph	48	48	44	39
Szegedi–E. et al. (2013) (a)	ONC	372	131	55	52
Szegedi–E. et al. (2013) (b)	ONC	187	56	9	7
Total		1728	724	475	434

Table 4. Number of sources of our binned SIMBAD subtypes (first column) in the initial *W0* sample (second column), in our SVM classified Class I/II candidate sample (third column) and in the YSO sample of the KL14 method (last column). KL14 successfully removes the Galactic contamination, but is less successful in identifying the known YSOs and in removing the extragalactic contamination.

Subtype	W0	SVM Class I/II	KL14 YSOs
G1	148 267	65	1398
G2	10 311	7	1188
G3	12 729	29	3303
E1	13 208	631	58
E2	3429	1118	613
E3	109	43	16
ISM1	912	183	90
ISM2	515	173	50
S	973 629	11 990	1043
Y1	9268	3705	2695
Y2	1128	637	650

i.e. a large proportion of YSOs do not have reliable fluxes in all *WISE* and/or 2MASS bands. An additional 249 sources (14.4 per cent) turned out to be spurious *WISE* detections, while 434 of the remaining 475 (91.3 per cent) were successfully re-classified.

4.3 Comparison with the Koenig & Leisawitz (2014) method

Recently, Koenig & Leisawitz (2014, hereafter KL14) published a method to reduce the number of spurious sources and to identify potential YSO candidates in the *WISE* and AllWISE catalogue. In this comparison, we applied their method to our *W0* sample to be able to compare the results, including the spurious source removal. As a result, the KL14 method identified 124 608 YSO candidates from the *W0* sample.

In Table 4, we compare the number of sources of our 11 binned subtypes in the following three samples: (1) the initial *W0* sample (S/N > 3 in all *WISE* bands and 2MASS photometric errors <0.1), (2) our SVM-selected Class I/II sample and (3) the YSO sample resulted from applying the KL14 method to the initial *W0* sample. Because the KL14 method has its own criteria for spurious source mitigation, the comparison is fair if we apply it to our *W0* sample. We can see that the number of false-positive extragalactic objects in our Class I/II selection is 101 (0.05 per cent compared to the *real* sample), while it is 5889 (3.4 per cent) in the KL14 sample. The fraction of known SIMBAD YSOs that were recovered with our method is higher (41.7 per cent) than with the KL14 one (32.1 per cent). On the other hand, the KL14 method successfully eliminated 95.9 per cent of the evolved stars (89.3 per cent with the SVM). We conclude that KL14 method is very efficient in identifying and removing the Galactic contamination, but allows us to retrieve a lower number of known YSOs. Our method is more successful in removing extragalactic contamination, and it misclassifies known YSOs in a smaller fraction. A summary of the comparison between the two methods is provided in Table A1.

We also investigated whether the KL14 method is more sensitive to Class I/II YSOs rather than to more evolved class types. Using the same sources that we used in Section 3.2.4, we found that KL14 method finds 95 of the 247 Class I sources (SVM finds 240), 922 of the 1925 Class II sources (1824 with SVM) and 47 of the 313 Class III objects (SVM result is 79). The combined results show that SVM is able to retrieve 2064 of the 2172 Class I/II objects (95 per cent), while the KL14 method finds only 589 (27 per cent). These results suggest that the overlap between the two methods is rather small; however, the results based on the SIMBAD comparison show a somewhat better agreement. To find out the reason for such a discrepancy, we calculated the average magnitudes in each 2MASS and *WISE* band for the SIMBAD YSOs found in each selection. The results are listed in Table 5. We also analysed the magnitude distribution of the SIMBAD YSOs in the KL14 selection (see Fig. E1). Our conclusion is that the KL14 method is overall more sensitive to the fainter sources than our method. This can explain why, on one side, the KL14 method classifies more extragalactic sources as YSOs and successfully removes bright Galactic contaminants while, on the other, it recovers a smaller fraction of YSOs.

4.4 Comparison with the QDA

In our previous work (Tóth et al. 2014), the QDA (McLachlan 1992) technique was used to identify YSO candidates using far-IR *AKARI* and mid-IR *WISE* data. In this section, we compare the

Table 5. Average brightness of the SIMBAD YSOs classified as YSO with the KL14 method (second column) and with our SVM method (third column). The KL14 method appears to be more sensitive to the fainter sources.

Band	KL14 YSOs (mag)	SVM Class I/II (mag)
J	13.9 ± 1.7	13.1 ± 2.0
H	12.5 ± 1.4	11.6 ± 1.8
K_s	11.6 ± 1.4	10.7 ± 1.7
$W1$	10.6 ± 1.4	9.6 ± 1.7
$W2$	9.8 ± 1.5	8.9 ± 1.7
$W3$	7.5 ± 1.7	6.6 ± 1.8
$W4$	5.0 ± 1.7	4.4 ± 2.0

currently used SVM method and the QDA by repeating the steps of the analysis described above, and using the same training samples, to find out which one suits the problem better. The main difference between the two methods is their approach to the decision boundaries. Discriminant analysis techniques perform dimensionality reduction, and project the data into a subspace where they maximize the separation. On the contrary, the SVM maps the data into a higher dimensional space and a hyperplane is calculated that provides the best separation of the classes.

As a first step, the classification of real and spurious sources was repeated. QDA successfully classified as such only 77.9 per cent of the spurious sources (compared to 98.3 per cent for SVM), and only 37.9 per cent of the real sources (compared to 94.2 per cent for SVM). In this case, SVM clearly outperforms the QDA method.

As a second step, we repeated the removal of the extragalactic sources. 99.3 per cent of the sources in the training sample were successfully re-classified as extragalactic source and only 0.7 per cent (725) were misclassified by QDA. This is three times more than with SVM, for which only 237 sources were misclassified. By using SVM, we were not able to recover 2.1 per cent of the known SIMBAD YSOs, while we lost 2.3 per cent of them with QDA.

In the third step, we repeated the removal of field stars. With QDA, 17.3 per cent of the SIMBAD single stars remained in our sample, while with SVM only 8.9 per cent. Also, with SVM 5.5 per cent of the YSOs were lost while applying QDA to the same training sample resulted in the loss of 7.8 per cent of them.

As a last step of the QDA and SVM comparison, we repeated the classification of the remaining sources into three main object types, YSO candidates, evolved stars and ISM-related objects. By using SVM, we successfully re-classified 96 per cent of the known YSOs as YSO candidate. Using QDA, the success rate was only 89 per cent. The contamination caused by the remaining evolved stars is also higher with QDA. The number of evolved objects was found to be 6194 while it was 4382 with SVM.

This comparison clearly shows that while QDA and SVM are comparable in some of the steps, the overall performance of SVM is better than QDA, and it is more efficient for classifying Class I/II YSOs.

4.5 Correlation with the PGCCS

The *Planck* Catalogue of Galactic Cold Clumps (PGCC; Planck Collaboration 2015) is an all-sky catalogue of Galactic cold clump candidates detected by Planck. The PGCC catalogue contains 13 188 Galactic sources spread across the whole sky with a median temperature between 13 and 14.5 K and their size is described with the major and minor full width at half-maximum (FWHM) of a fitted elliptical Gaussian. Cold clumps represent the early stages of star formation (McKee & Ostriker 2007) and a spatial correlation between the cold clump and the YSO distribution is therefore expected.

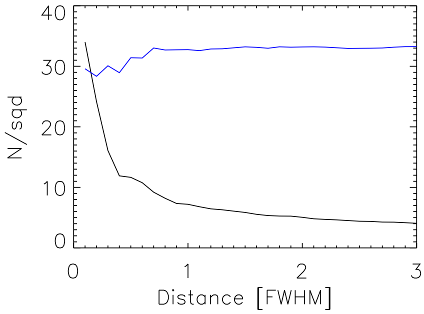

Figure 8. Surface density of the Class I/II candidates (black solid line) and the MS star candidates (blue line) as a function of the distance relative to the PGCC objects.

To further test the robustness of our YSO classification, we analysed their position relative to the PGCCs in the Taurus–Auriga–Perseus–California molecular complex ($150 < l < 180$, $-25 < b < -1$), which is a well-known star-forming region. As a function of the major FWHM, we calculated the surface density of our Class I/II candidates and of the objects that we classified as MS stars. As seen on Fig. 8, the surface density of the Class I/II candidates is highest close to the PGCCs and then rapidly decreases, while that of the MS star candidates is independent of the distance measured from the cold clumps. This result strongly suggests that our candidate Class I/II sources are indeed related to the Planck cold clump population.

5 SUMMARY

The AllWISE catalogue was investigated to identify potential YSO candidates. A subset of the catalogue was used with S/N > 3 and available 2MASS J, H, K_s data with photometric errors <0.1. We applied the SVM method to the initial data set of 8956 636 sources in a multistep process. Different combinations of colours and magnitudes were used, in combination with the extended source flag, in order to generate a multidimensional training samples and to remove contaminating sources. Sources of known Galactic and extragalactic types were identified with the help of the SIMBAD data base, using a 5 arcsec radius to match the AllWISE sources. As many as 133 980 objects were classified as YSO Class I/II candidates. The contamination of sources with well-known object types is <1 per cent, in comparison with our SIMBAD training set. We also compared our method to that described in KL14 and to the results obtained with a different approach, the QDA. We found that SVM outperforms the KL14 method in preserving the known YSOs and in identifying the extragalactic contamination and it is more effective than QDA.

A positional correlation analysis with the PGCC sources was performed in the case of the Taurus–Auriga–Perseus–California regions. Our Class I/II candidates appear to be characterized by a higher surface density in the proximity of the cold clumps while the MS star candidates had a uniform surface density in the field.

ACKNOWLEDGEMENTS

We thank our anonymous referee for all the useful comments that helped us to improve the manuscript. This publication makes use of data products from the *WISE*, which is a joint project of the University of California, Los Angeles and the Jet Propulsion Laboratory/California Institute of Technology, funded by the National

Aeronautics and Space Administration. This research has made use of the SIMBAD data base, operated at CDS, Strasbourg, France. This research has made use of the VizieR catalogue access tool, CDS, Strasbourg, France. This research was supported by the OTKA grants NN 111016 and K101393.

REFERENCES

Ahn C. P. et al., 2012, ApJS, 203, 21
Alcala J. M. et al., 2014, A&A, 561, 2
Allen T. S. et al., 2012, ApJ, 750, 125
An D. et al., 2011, ApJ, 736, 133
Billot N., Noriega-Crespo A., Carey S., Guieu S., Shenoy S., Paladini R., Latter W., 2010, ApJ, 712, 797
Chavarria L. A., Allen L. E., Hora J. L., Brunt C. M., Fazio G. G., 2008, ApJ, 682, 445
Connelley M. S., Greene T. P., 2010, AJ, 140, 1214
Connelley M. S., Reipurth B., Tokunaga A. T., 2008, AJ, 135, 2496
Cooper H. D. B. et al., 2013, MNRAS, 430, 1125
Cutri R. M. et al., 2003, 2MASS All Sky Catalog of point sources, NASA/IPAC Infrared Science Archive. Available at: http://irsa.ipac.caltech.edu/applications/Gator/
Cutri R. M. et al., 2012, VizieR On-line Data Catalog: II/31
Cutri R. M. et al., 2013, VizieR On-line Data Catalog: II/311
Erickson K. L., Wilking B. A., Meyer M. R., Kim J. S., Sherry W., Freeman M., 2015, AJ, 149, 103
Evans N. J. et al., 2009, ApJS, 181, 321
Fang M., Van Boekel R., Wang W. Carmona A., Sicilia-Aguilar A., Henning Th., 2009, A&A, 504, 461
Fazio G. et al., 2004, ApJS, 154, 10
Gutermuth R. A. et al., 2008, ApJ, 674, 336
Gutermuth R. A., Megeath S. T., Myers P. C., Allen L. E., Pipher J. L., Fazio G. G., 2009, ApJS, 184, 18
Harvey P. et al., 2007, ApJ, 663, 1149
Kawada M. et al., 2007, PASJ, 59, 389
Kirk J. M. et al., 2009, ApJS, 185, 198
Koenig X. P., Leisawitz D. T., 2014, ApJ, 791, 131 (KL14)
Koenig X. P., Allen L. E., Gutermuth R. A., 2008, ApJ, 688, 1142
Koenig X. P., Leisawitz D. T., Benford D. J., Rebull L. M., Padgett D. L., Assef R. J., 2012, ApJ, 744, 130
Kumar B. et al., 2014, A&A, 567, 109
Kun M., Balog Z., Kenyon S. J., Mamajek E. E., Gutermuth R. A., 2009, ApJS, 185, 451
Lada C. J., 1987, in Peimbert M., Jugaku J., eds, Proc. IAU Symp. 115, Star Forming Regions. Reidel, Dordrecht, p. 1
McKee C. F., Ostriker E. C., 2007, ARA&A, 45, 565
McLachlan G. J., 1992, Discriminant Analysis and Statistical Pattern Recognition. Wiley, New York
Małek K. et al., 2013, A&A, 557, 16
Megeath S. T. et al., 2012, AJ, 144, 192
Mooley K., Hillenbrand L., Rebull L., Padgett D., Knapp G., 2013, ApJ, 771, 110
Murakami H. et al., 2007, PASJ, 59, 369
Neugebauer G. et al., 1984, ApJ, 278, 1
Ochsenbein F., Dubois P., 1992, ESO Conf. Workshop Proc. No. 43, Astronomy from Large Databases, II. European Southern Observatory, Garching, p. 405
Oliveira I. et al., 2009, ApJ, 691, 672
Planck Collaboration XI, 2014, A&A, 571, 11
Planck Collaboration 2015, submitted to A&A-PLX5
Pollo A., Rybka P., Takeuchi T. T., 2010, A&A, 514, A3
Rebollido I. et al., 2015, A&A, 581, 30
Rebull L. M. et al., 2010, ApJS, 186, 259
Rebull L. M. et al., 2011, ApJS, 193, 25
Rieke G. H. et al., 2004, ApJS, 154, 25
Rivera-Ingraham A., Martin P. G., Polychroni D., Moore T. J. T., 2011, ApJ, 743, 39
Robitaille T. P. et al., 2008, AJ, 136, 2413
Skrutskie M. F. et al., 2006, AJ, 131, 1163
Szegedi-Elek E., Kun M., Reipurth B., Pál A., Balázs L. G., Willman M., 2013, ApJS, 208, 28
Tóth L. V. et al., 2014, PASJ, 66, 17
Vapnik V. N., 1995, The Nature of Statistical Learning Theory. Springer, Berlin
Werner M. et al., 2004, ApJS, 154, 1
Winston E. et al., 2007, ApJ, 669, 493
Winston E. et al., 2010, AJ, 140, 266
Wright E. L. et al., 2010, AJ, 140, 1868
Yamamura I. et al., 2010, VizieR On-line Data Catalog: II/298, available at: http://irsa.ipac.caltech.edu/data/AKARI/documentation/AKARI-FIS_BSC_V1_RN.pdf

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Table A1. Comparison of SIMBAD sources found in our initial *W0* sample, in the SVM selected Class I/II YSO sample and in the YSO candidate sample identified by the KL14 method.
Table B1. Average colours and magnitudes of SIMBAD object types.
Table B2. Average colours and magnitudes of SIMBAD object types.
Table B3. Average colours and magnitudes of SIMBAD object types.
Table B4. Average colours and magnitudes of SIMBAD object types.
Table B5. Average colours and magnitudes of SIMBAD object types.
Table B6. Average colours and magnitudes of SIMBAD object types.
Table B7. Average colours and magnitudes of SIMBAD object types.
Table B8. Average colours and magnitudes of SIMBAD object types.
Table B9. Average colours and magnitudes of SIMBAD object types.
Figure C1. Example of sources used as real for the spurious source identification training sample.
Figure C2. Example of sources selected as spurious for the spurious source identification training sample.
Figure D1. Three main groups of SIMBAD extragalactic objects.
Figure D2. Average colour indices and *ext* values of three main groups of SIMBAD source types of evolved objects.
Figure D3. Average colour indices and *ext* values of two groups of SIMBAD source types of young objects.
Figure D4. Average colour indices and *ext* values of two groups of SIMBAD source types of ISM-related objects.
Figure E1. Fraction of known YSOs in the *W0* (black solid line), *real* (red dotted line), Class I/II (blue dashed line) and KL14 (green-dash–dotted line) samples as a function of brightness in the different 2MASS and *WISE* bands.
(http://mnras.oxfordjournals.org/lookup/suppl/doi:10.1093/mnras/stw398/-/DC1).

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

This paper has been typeset from a TEX/LATEX file prepared by the author.