Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **476,** 3981–3990 (2018)
Advance Access publication 2018 March 3

doi:10.1093/mnras/sty518

Improved selection criteria for H II regions, based on *IRAS* sources

Qing-Zeng Yan,[1,2,3,4★] Ye Xu,[2★] A. J. Walsh,[3] J. P. Macquart,[3] G. C. MacLeod,[5] Bo Zhang,[1] P. J. Hancock,[3] Xi Chen[1,6★] and Zheng-Hong Tang[1,7]

[1]*Shanghai Astronomical Observatory, Chinese Academy of Sciences, Shanghai 200030, China*
[2]*Purple Mountain Observatory, Chinese Academy of Sciences, Nanjing 210008, China*
[3]*International Centre for Radio Astronomy Research, Curtin University, GPO Box U1987, Perth, WA 6845, Australia*
[4]*University of Chinese Academy of Sciences, 19A Yuquanlu, Beijing 100049, China*
[5]*Hartebeesthoek Radio Astronomy Observatory, PO Box 443, Krugersdorp 1740, South Africa*
[6]*Center for Astrophysics, GuangZhou University, Guangzhou 510006, China*
[7]*School of Astronomy and Space Science, University of Chinese Academy of Sciences, 19A Yuquanlu, Beijing 100049, China*

Accepted 2018 February 20. Received 2018 February 20; in original form 2017 October 19

ABSTRACT

We present new criteria for selecting H II regions from the *Infrared Astronomical Satellite* (*IRAS*) Point Source Catalogue (PSC), based on an H II region catalogue derived manually from the all-sky *Wide-field Infrared Survey Explorer* (*WISE*). The criteria are used to augment the number of H II region candidates in the Milky Way. The criteria are defined by the linear decision boundary of two samples: *IRAS* point sources associated with known H II regions, which serve as the H II region sample, and *IRAS* point sources at high Galactic latitudes, which serve as the non-H II region sample. A machine learning classifier, specifically a support vector machine, is used to determine the decision boundary. We investigate all combinations of four *IRAS* bands and suggest that the optimal criterion is $\log(F_{60}/F_{12}) \geqq (-0.19 \times \log(F_{100}/F_{25}) + 1.52)$, with detections at 60 and 100 μm. This selects 3041 H II region candidates from the *IRAS* PSC. We find that *IRAS* H II region candidates show evidence of evolution on the two-colour diagram. Merging the *WISE* H II catalogue with *IRAS* H II region candidates, we estimate a lower limit of approximately 10 200 for the number of H II regions in the Milky Way.

Key words: stars: evolution – stars: massive – stars: statistics – H II regions – infrared: ISM – infrared: stars.

1 INTRODUCTION

High-mass stars, whose masses exceed 8 $M_\odot$ (Zinnecker & Yorke 2007), are OB stars that emit strong ultraviolet (UV) radiation, thereby ionizing surrounding atomic and molecular gases. Composed mainly of ionized hydrogen, these ionized gases are usually called H II regions. Most H II regions essentially trace high-mass stars, hence the evolution and distribution of H II regions is useful for investigating high-mass stars in the Milky Way.

Although the amount of H II regions is an important evolutionary indicator of the Milky Way, their total number is still unclear due to the difficulty in detecting and identifying them due to their large distances and rapid evolution (Zinnecker & Yorke 2007). In their early stages, when H II regions are compact, high-mass stars are deeply embedded in cold thick molecular clouds, whose typical temperature is about 30 K (Wolfire & Churchwell 1994; Garay & Lizano 1999). In this phase, high-mass stars are invisible at optical wavelengths due to high extinction caused by dust grains, but they are observable at longer wavelengths, notably in the infrared or radio bands. The infrared output of H II regions is generally due to the thermal emission of their internal or surrounding dust grains (Garay & Lizano 1999; Churchwell 2002), whereas the radio output is generated by their internal free–free emission (Kurtz, Churchwell & Wood 1994; Walsh et al. 1998). However, the properties of infrared emission vary with temperature, while the optical depth of free–free emission hinges on the frequency (Kurtz et al. 1994).

Wood & Churchwell (1989b, hereafter WC89) investigated the population and distribution of embedded high-mass stars in the Milky Way using the all-sky *Infrared Astronomical Satellite* (*IRAS*) Point Source Catalogue (PSC; Neugebauer et al. 1984; Helou & Walker 1988) over four infrared bands at 12, 25, 60, and 100 μm. If there is no detection in one band, the flux density quality of this band is marked with an upper limit. With the help of some previously identified ultracompact (UC) H II regions (Wood & Churchwell 1989a), they derived a criterion for embedded high-mass stars: $\log(F_{60}/F_{12}) \geqq 1.30$ and $\log(F_{25}/F_{12}) \geqq 0.57$, where

★ E-mail: qzyan@shao.ac.cn(Q-ZY); xuye@pmo.ac.cn (YX); chenxi@shao.ac.cn (XC)

F_{12}, F_{25}, and F_{60} represent fluxes at 12, 25, and 60 μm, respectively. They further rejected sources whose flux density quality at either 25 or 60 μm is marked by an upper limit. They identified 1717 UC H II region candidates and potentially missed many evolved H II regions.

Using known H II regions, including UC H II regions, Hughes & MacLeod (1989, hereafter HM89) investigated the H II regions based on a two-colour diagram of *IRAS* sources. They provided a decision boundary of $\log(F_{25}/F_{12}) \geqq 0$ and $\log(F_{60}/F_{25}) \geqq 0$. They also imposed extra constraints on F_{100} and Galactic latitudes, and the total number of H II region candidates identified was 2298. However, the sample of known H II regions they used is far from complete, and therefore the criterion was not well constrained.

Recently, Anderson et al. (2014) created a catalogue of H II region candidates, providing an opportunity to improve the selection criteria for H II region candidates. This H II region candidate catalogue is based on the all-sky *Wide-field Infrared Survey Explorer* (*WISE*; Wright et al. 2010). They created this catalogue by identifying infrared bubbles (Churchwell et al. 2006) manually. Infrared bubbles essentially are H II regions produced by high-mass stars. The 12 μm band emission traces polycyclic aromatic hydrocarbon (PAH) molecules, delineating the edge of H II regions, while the 24 μm band emission mostly traces internal thermal emission from dust grains heated by ionized gases. This is the most complete catalogue of H II regions in the Milky Way, because the covering area is much larger than that of the *Spitzer*/Galactic Legacy Infrared Mid-Plane Survey Extraordinaire (GLIMPSE) survey (Benjamin et al. 2003), and the number (8399) of identified H II regions exceeds that provided by another similar undertaking: the Milky Way Project (MWP; Simpson et al. 2012).

However, the catalogue of Anderson et al. (2014) can potentially miss those H II regions that have small angular sizes or that are not easily identified visually. In this paper, we investigate possible criteria for selecting H II regions from *IRAS* sources to obtain a more complete census of H II regions in the Milky Way, using publicly available radio and infrared surveys and sophisticated algorithms. Essentially, the criteria are defined by the decision boundary of two samples: *IRAS* sources that are associated with known H II regions and those sources that are not H II regions. The *WISE* H II region candidates (Anderson et al. 2014) are used as a basis for the H II region sample (subject to further selection criteria based on radio detection), while the *IRAS* point sources at high Galactic latitudes serve as the non-H II regions. Support vector machines (SVMs; Vapnik 1995), which are machine learning algorithms used to do supervised classification, are applied to derive the decision boundary of the two samples. A 3D simulation of the expansion of H II regions (Tremblin et al. 2014) enables us to estimate the age of H II regions and to investigate the evolution of H II regions on two-colour diagrams of *IRAS* sources.

2 ANALYSIS

In this section, we present our method of producing selection criteria for H II regions, based on the *IRAS* PSC and the *WISE* H II region catalogue. Essentially, the criteria are determined by two types of *IRAS* point sources: H II regions, and non-H II regions. We match the *IRAS* PSC to the positions of the *WISE* H II region candidates that have radio counterparts, yielding the H II region sample, whereas the non-H II region sample is built from high Galactic latitude sources ($|b| > 8^\circ$). The decision boundary of the two samples is produced by the SVM algorithm and is subsequently applied to the *IRAS* PSC to identify H II region candidates.

2.1 Catalogues

Our analysis is based on the *IRAS* PSC (Helou & Walker 1988), the *WISE* H II region catalogue (Anderson et al. 2014), and three radio continuum source catalogues (Condon et al. 1998; Mauch et al. 2003; Murphy et al. 2007).

The *IRAS* PSC (version 2.1) includes four infrared fluxes at 12, 25, 60, and 100 μm, with the resolution ranging from 45 arcsec to 3 arcmin. We use F_{12}, F_{25}, F_{60}, and F_{100} to denote fluxes at these four bands, respectively, and use Q_{12}, Q_{25}, Q_{60}, and Q_{100} to denote their corresponding qualities. The flux quality values of 1, 2, and 3 represent an upper limit (i.e. non-detection), moderate quality, and high quality, respectively. After eliminating galaxies and quasars identified by Fullmer & Lonsdale (1989), we further rejected those *IRAS* sources that are matched with extragalactic objects within 1 arcmin, including nearby galaxies (Kraan-Korteweg 1986; Karachentsev, Makarov & Kaisina 2013; Bai et al. 2015) and unresolved very long baseline interferometry (VLBI) calibrators identified by Xu et al. (2006) and Immer et al. (2011). Our study is based on the remaining 234 261 *IRAS* point sources.

The H II region catalogue that we use to extract H II regions from the *IRAS* PSC is created by Anderson et al. (2014) from *WISE* data. This catalogue contains 8399 H II region candidates, covering $|b| \leqq 8^\circ$ and five high-mass star-forming regions at high Galactic latitudes. In this catalogue, 1413 H II regions have their distances determined, and we estimate their ages based on a numerical simulation of the 3D expansion of H II region (Tremblin et al. 2014).

In their catalogue, Anderson et al. (2011) found that the co-existence of radio continuum and mid-infrared emission can identify H II regions at a 95 per cent confidence level. Therefore, in order to examine the quality of H II region candidates, we use three radio continuum surveys: the NRAO VLA Sky Survey (NVSS; Condon et al. 1998), the Sydney University Molonglo Sky Survey (SUMSS; Bock, Large & Sadler 1999; Mauch et al. 2003), and the second epoch Molonglo Galactic Plane Survey (MGPS-2; Murphy et al. 2007). NVSS, SUMSS, and MGPS-2 have similar sensitivities and spatial resolutions and collectively cover the whole sky. At frequencies $\nu \leqq 8$–15 GHz, radio free–free emission is optically thick (Kurtz et al. 1994; Protheroe et al. 2008) for UC H II regions, rendering UC H II regions undetectable at these frequencies. Nonetheless, the proportion of sources associated with radio continuum emission is still an excellent indicator of the quality of the selection criteria, because only those H II regions at very early stages are missed.

We summarize the five catalogues in Table 1, where from left to right, we list the name, the telescope, the observed band, the spatial resolution, the covering area, and the reference for each catalogue.

2.2 Support vector machines

The SVM algorithm is used to decide the decision boundary of H II regions. SVMs, developed by Vladimir Vapnik (Cortes & Vapnik 1995; Vapnik 1995), are algorithms used to do classification and regression analysis in supervised machine learning. For two groups of points, which are linearly separable, SVMs determine their decision boundary by maximizing the gap between them. However, if they are not linearly separable, SVMs can still perform classification by mapping them into a higher dimensional space using specific kernels.

We adopted linear SVM classifiers, because the overlapping area of H II regions and non-H II regions is not large, meaning they are well separated. We use the PYTHON package SKLEARN to perform linear SVMs. An important parameter of this algorithm is the penalty

Table 1. Catalogues of *IRAS* point sources, H II regions, and three radio continuum surveys.

Catalogue	Telescope	Band	Resolution	Coverage	Reference
IRAS v2.1	*IRAS*	12, 25, 60, and 100 μm	45 arcsec–3 arcmin	All sky	Neugebauer et al. (1984) and Helou & Walker (1988)
H II regions	*WISE*	12 and 22 μm	6.5 and 12 arcsec	$\|b\| \leqq 8^\circ$	Anderson et al. (2014)
NVSS	VLA	1.4 GHz	45 arcsec	$\delta > -40^\circ$	Condon et al. (1998)
SUMSS	Molonglo	843 MHz	∼45 arcsec	$\delta < -30^\circ (\|b\| > 10^\circ)$	Bock et al. (1999) and Mauch et al. (2003)
MGPS-2	Molonglo	843 MHz	∼45 arcsec	$245^\circ < l < 365^\circ$, $\|b\| < 10^\circ$	Murphy et al. (2007)

parameter for misclassification, denoted by C. Larger values of C impose higher penalties for misclassification, while smaller values of C permit more misclassification.

Because of the presence of overlapping areas, we allow a small fraction of misclassification, but the misclassification needs to be firmly constrained to avoid involving a larger number of high Galactic latitude sources. Consequently, we adopt an intermediate value $C = 1$. In Section 2.3, we find the effect caused by a small shift of C is not significant.

2.3 Producing the criteria

HM89 investigated the selection criteria for H II regions based on an incomplete sample of H II regions, while we use a more complete H II region catalogue and more sophisticated algorithms to improve their result in this subsection. The entire process of creating selection criteria is divided into five main steps, details of which are described in the rest of this subsection. We choose the optimal criterion according to a metric, which is defined below (in equation 2).

The process of producing criteria includes five main steps.

(i) Identify *IRAS* sources associated with H II regions, which serve as the sample of known H II regions.
(ii) Select *IRAS* sources at high Galactic latitudes, which serve as the sample of non-H II regions.
(iii) Use SVMs to determine the criteria based on the samples of H II regions and non-H II regions for all possible colour combinations.
(iv) Apply these criteria to the *IRAS* PSC.
(v) Determine the optimal criterion according to their scores.

In the first step, we matched the *IRAS* PSC to the *WISE* H II region catalogue. Following Anderson et al. (2014), we only adopted *WISE* sources that have small angular sizes (radii <4 arcmin) and ignore those lack detected radio continuum emission (the classification '*Q*'). In total, 1773 *IRAS* sources are matched with at least one H II region, and these *IRAS* sources serve as the sample of H II regions. For a particular colour combination, these sources are further filtered with bands required to have good qualities (better than an upper limit).

The second step is to build the sample of non-H II regions, by selecting *IRAS* sources at high Galactic latitudes. This is because high-mass stars are generally far away from the Sun and are tightly constrained to the Galactic plane (Zinnecker & Yorke 2007). Furthermore, they are not overlapping with any of the giant molecular clouds (GMCs) in the Milky Way identified by Rice et al. (2016). At high Galactic latitudes (except the Orion nebula, which is a high-mass star-forming region at a distance of about 400 pc from the Sun), most *IRAS* point sources are low-mass stars or extragalactic objects. Although some extragalactic objects possess similar colours to H II regions, there are relatively few, on account of the reddening caused by the intergalactic medium (IGM; Wright 1981; Assef et al. 2013).

Because the *WISE* H II region catalogue extends up to a Galactic latitude of 8°, we selected those *IRAS* sources whose absolute values of Galactic latitude are greater than 8° ($|b| > 8^\circ$), serving as the sample of non-H II regions. We further rejected those sources in three prominent regions: Orion, the Large Magellanic Cloud (LMC), and the Small Magellanic Cloud (SMC). According to a CO survey performed by Wilson et al. (2005), the ranges of Galactic longitude and latitude for Orion are [200°, 220°] and [−22°, −8°], respectively. Based on the position and size of galaxies provided by Cook et al. (2014), the Galactic longitude ranges of the LMC and the SMC are [275°, 286°] and [299°, 305.°5] and their Galactic latitude ranges are [−38.°5, −27.°5] and [−47°, −41.°5], respectively.

We used the criteria of WC89 and HM89 (no constraints on Galactic latitudes) to examine the quality of those sources at high Galactic latitudes. In total, 6200 sources are tested by the inequalities of WC89 or HM89, and we find 89 (1.4 per cent) sources at high Galactic latitude ($|b| > 8^\circ$) agree with at least one of these two criteria. After eliminating these 89 sources, we have remaining 100 634 sources at high Galactic latitudes, which is used to build the sample of non-H II regions. Despite the incompleteness of WC89 and HM89, we estimate that less than 1.4 per cent of those sources are possibly H II regions.

In the third step, we checked all the possible two-colour combinations of *IRAS* bands, each of which involves at least three bands. For each colour, we require that the shorter wavelength is the denominator so that H II regions will be above the decision boundary. For an *IRAS* point source, if the shorter wavelength band of its colour is marked with an upper limit, the true value will only move this source to the upper right-hand direction on the two-colour diagram. Therefore, we only require that the flux quality at longer wavelengths is better than an upper limit. In total, there are 15 possible two-colour combinations. With the help of SVMs, we determined the decision boundary for each colour combination, and as mentioned above, the penalty parameter (C) is assigned a value of 1, and the effect caused by a small change of C is negligible. Generally, within the range $0.5 < C < 1.5$, the shifts of slopes and y-intercepts of the criteria are less than 0.01.

Before we performed the fourth step, we calculated some statistical measures for the criteria, including sensitivity, specificity, and the F_1 score. The true positives (TP) are the H II region samples agreeing with the criteria, while the false negatives (FN) are the H II region samples rejected by the criteria. The true negatives (TN) are the non-H II region samples rejected by the criteria, while the false positives (FP) are the non-H II region samples agreeing with the criteria. The definition of sensitivity, precision, and the F_1 score are

$$\begin{aligned} \text{sensitivity} &= \frac{\text{TP}}{(\text{TP}+\text{FN})}, \\ \text{precision} &= \frac{\text{TP}}{(\text{TP}+\text{FP})}, \\ F_1\ \text{score} &= \frac{2}{(1/\text{sensitivity}+1/\text{precision})}, \end{aligned} \tag{1}$$

where sensitivity is also called the recall or the true positive rate (TPR), precision is also called the positive predictive value (PPV),

Table 2. All possible two-colour criteria for selecting H II regions from the *IRAS* PSC.

Identity	criterion[a]	Candidates	Association (radio)[b] (per cent)	Sensitivity[c] (per cent)	Precision[d] (per cent)	Tested[e] (per cent)	Score[f]
1[g]	$\log\left(\frac{F_{60}}{F_{12}}\right) > \left(-0.19 \times \log\left(\frac{F_{100}}{F_{25}}\right) + 1.52\right)$, Q_{60} and $Q_{100} > 1$	3041	32.46	90.58	93.32	24 441	0.5707
2	$\log\left(\frac{F_{100}}{F_{60}}\right) > \left(-3.43 \times \log\left(\frac{F_{60}}{F_{12}}\right) + 4.73\right)$, Q_{60} and $Q_{100} > 1$	3077	32.11	90.48	93.70	24 441	0.5674
3	$\log\left(\frac{F_{60}}{F_{12}}\right) > \left(-0.45 \times \log\left(\frac{F_{100}}{F_{12}}\right) + 2.00\right)$, Q_{60} and $Q_{100} > 1$	3126	31.70	90.48	93.41	24 441	0.5628
4[h]	$\log\left(\frac{F_{60}}{F_{100}}\right) > \left(-1.47 \times \log\left(\frac{F_{100}}{F_{12}}\right) + 2.03\right)$, Q_{60} and $Q_{100} > 1$	3128	31.55	90.48	93.31	24 441	0.5612
5	$\log\left(\frac{F_{60}}{F_{12}}\right) > \left(0.19 \times \log\left(\frac{F_{25}}{F_{12}}\right) + 1.18\right)$, Q_{25} and $Q_{60} > 1$	4502	30.16	89.01	92.77	24 941	0.5438
6	$\log\left(\frac{F_{25}}{F_{12}}\right) > \left(-1.25 \times \log\left(\frac{F_{60}}{F_{25}}\right) + 1.47\right)$, Q_{25} and $Q_{60} > 1$	4517	30.04	89.17	92.71	24 941	0.5426
7	$\log\left(\frac{F_{60}}{F_{25}}\right) > \left(-5.53 \times \log\left(\frac{F_{60}}{F_{12}}\right) + 7.90\right)$, $Q_{60} > 1$	4859	29.29	86.11	92.57	60 315	0.5304
8	$\log\left(\frac{F_{100}}{F_{12}}\right) > \left(0.41 \times \log\left(\frac{F_{100}}{F_{25}}\right) + 1.18\right)$, $Q_{100} > 1$	3742	27.15	80.20	91.26	60 349	0.4979
9	$\log\left(\frac{F_{100}}{F_{60}}\right) > \left(-1.47 \times \log\left(\frac{F_{25}}{F_{12}}\right) + 0.86\right)$, Q_{25} and $Q_{100} > 1$	3599	28.59	77.62	78.88	10 714	0.4956
10	$\log\left(\frac{F_{100}}{F_{12}}\right) > \left(-0.31 \times \log\left(\frac{F_{25}}{F_{12}}\right) + 1.66\right)$, Q_{25} and $Q_{100} > 1$	4439	25.21	92.41	91.77	10 714	0.4887
11	$\log\left(\frac{F_{100}}{F_{25}}\right) > \left(-1.29 \times \log\left(\frac{F_{25}}{F_{12}}\right) + 1.65\right)$, Q_{25} and $Q_{100} > 1$	4491	25.01	93.01	91.54	10 714	0.4865
12	$\log\left(\frac{F_{60}}{F_{25}}\right) > \left(-1.83 \times \log\left(\frac{F_{100}}{F_{12}}\right) + 4.01\right)$, Q_{60} and $Q_{100} > 1$	4111	22.70	85.41	88.36	24 441	0.4472
13	$\log\left(\frac{F_{60}}{F_{25}}\right) > \left(-0.02 \times \log\left(\frac{F_{100}}{F_{60}}\right) + 1.02\right)$, Q_{60} and $Q_{100} > 1$	3083	25.43	60.08	79.07	24 441	0.4372
14	$\log\left(\frac{F_{60}}{F_{25}}\right) > \left(-0.03 \times \log\left(\frac{F_{100}}{F_{25}}\right) + 1.06\right)$, Q_{60} and $Q_{100} > 1$	3099	25.20	59.78	79.19	24 441	0.4346
15	$\log\left(\frac{F_{100}}{F_{25}}\right) > \left(0.96 \times \log\left(\frac{F_{100}}{F_{60}}\right) + 1.05\right)$, $Q_{100} > 1$	4600	19.65	58.09	78.30	60 349	0.3710

[a]The slopes and intercepts have been rounded up, the error caused by which is not significant.
[b]The proportion of candidates possessing radio counterparts (radio association).
[c]The sensitivity is defined in equation (1).
[d]The precision is defined in equation (1).
[e]The number of *IRAS* sources tested by the inequality of criteria.
[f]The score is defined by equation (2).
[g]Criterion 1 is the optimal criterion.
[h]In order to make sure H II regions are above the decision boundary, we use F_{60}/F_{100} instead of F_{100}/F_{60} for criterion 4.

and the F_1 score is the harmonic mean of precision and sensitivity. In the fourth step, we modify the F_1 score to include the proportion of candidates matching to radio sources (radio association).

In the fourth step, we applied all criteria to 234 261 *IRAS* point sources, identifying H II region candidates. After checking the quality of fluxes, we filter the *IRAS* point sources with the inequality of each criterion (see Table 2), yielding the H II region candidates. In order to check the quality of criteria, we matched the H II region candidates to radio continuum emission within a radius of 1 arcmin (following Walsh et al. 1997), and the proportions (radio association) are listed in Table 2.

Because the radio association is also an important indicator of the quality of the criteria, we modified the F_1 score and adopt a new type of score that is

$$\text{score} = \frac{3}{(1/\text{sensitivity} + 1/\text{precision} + 1/\text{radio association})}, \quad (2)$$

where score is the harmonic mean of sensitivity, precision, and radio association.

In Table 2, we list the parameters of all criteria, and from left to right, the columns are the identity, the criterion, the number of selected candidates, the proportion of candidates associated with radio continuum (radio association), sensitivity, precision, the number of *IRAS* sources tested by the inequality of the criteria, and the score. We sort the criteria according to their scores.

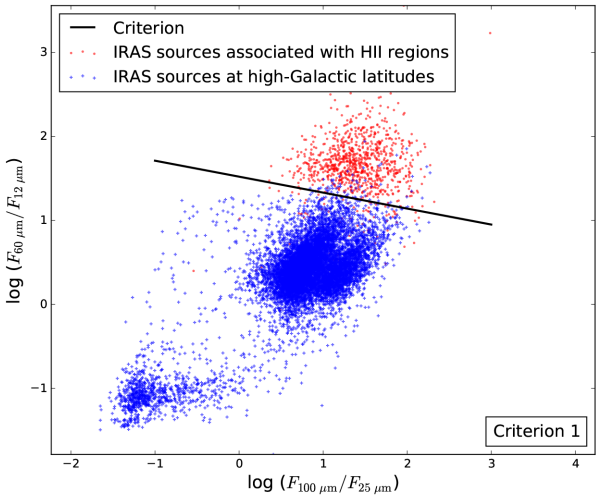
We display details of the top six criteria of Table 2 in Fig. 1, where the criteria provided by HM89 and WC89 are delineated by green lines. Fig. 2 shows the distribution of *IRAS* sources in terms of distances to the decision boundary, with negative values signifying sources rejected by criteria.

In the final step, we determine the optimal criterion. Because criterion 1 possesses the highest score, we adopt criterion 1 as the optimal criterion. In Table 2, criterion 2, 3, and 4 require $Q_{60} > 1$ and $Q_{100} > 1$ and they all use 12, 60, and 100 μm bands, which means they share the same data. The resemblance of results between criterion 2, 3, and 4 indicates the robustness of SVMs, and as expected, criterion 1 shows slightly better results because it uses all four bands.

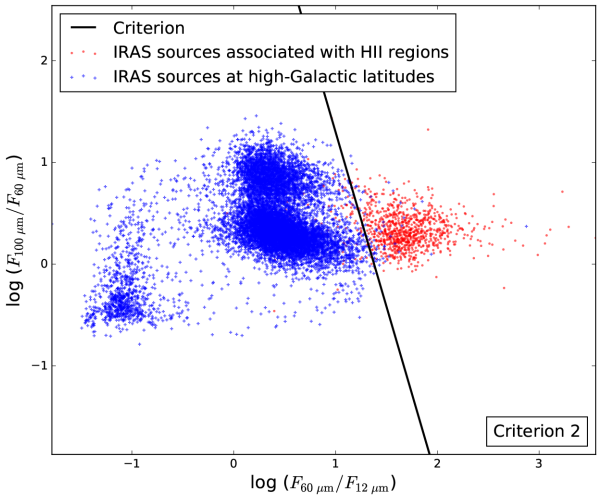
The sensitivity of criterion 1 indicates that ∼91 per cent of H II regions have been identified and its specificity (see Fig. 2) shows that less than 1 per cent of high Galactic latitude sources are H II regions. The radio association of criterion 1 is approximately 2 per cent higher than the proportion (32 per cent) in *WISE* H II region catalogue, which gives us a high confidence in the reliability of H II region candidates.

2.4 SVMs versus LDA

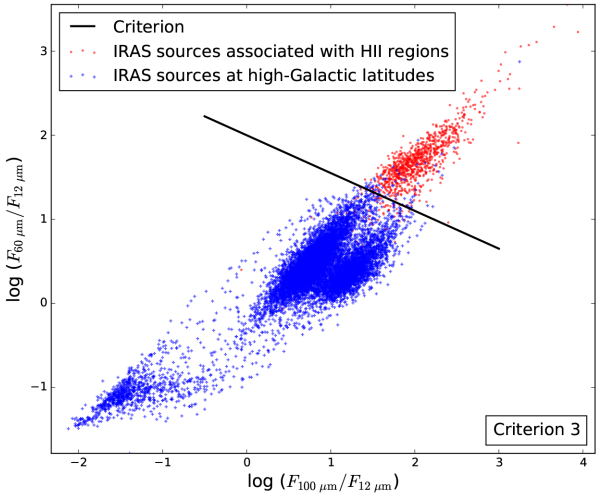
In this subsection, we compare the decision boundary determined by SVMs to that produced by linear discriminant analysis (LDA), which is another method to decide a linear decision boundary. In two-dimensional cases, LDA attempts to determine a decision boundary by maximizing the separation between samples and minimizing their scatter for the projections on an axis perpendicular to the decision boundary. However, SVMs primarily focus on the points near the boundary of samples.



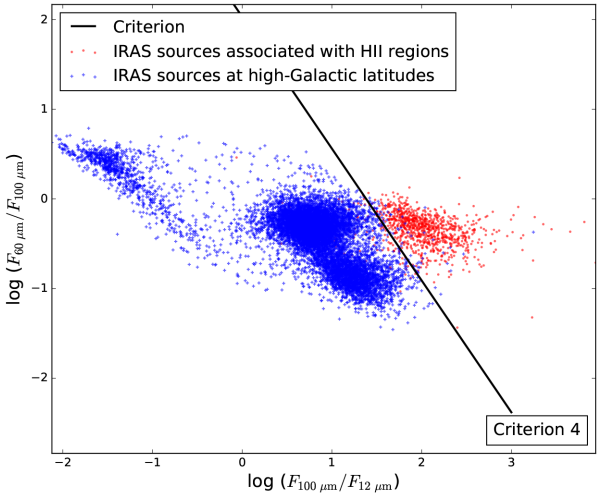
(a) Criterion 1.



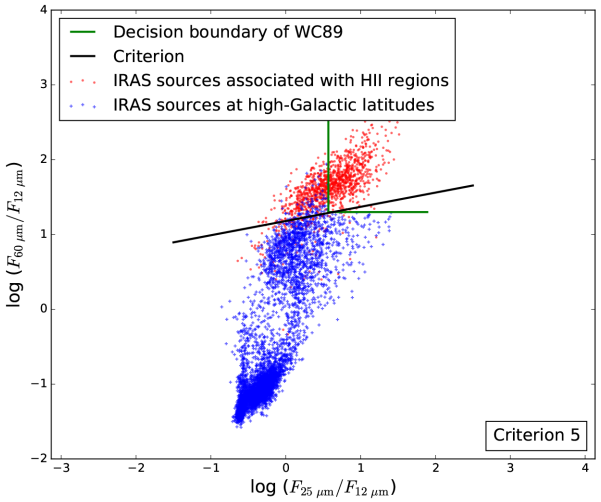
(b) Criterion 2.



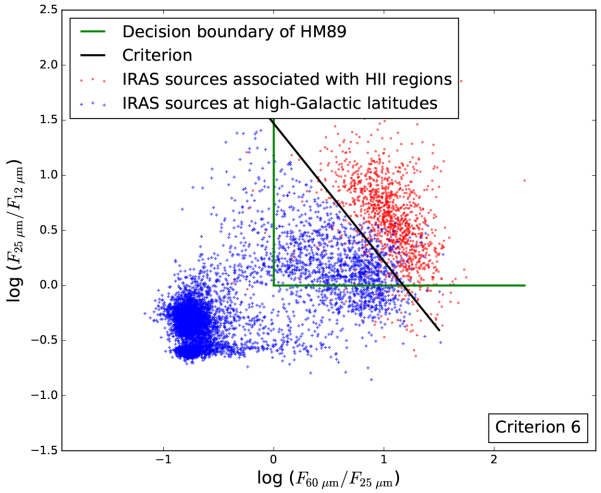
(c) Criterion 3.



(d) Criterion 4.



(e) Criterion 5.



(f) Criterion 6.

Figure 1. Top six criteria of Table 2. The criteria are criterion (a) 1, (b) 2, (c) 3, (d) 4, (e) 5, and (f) 6. The red and blue markers represent H II regions and non-H II regions, respectively. The black lines denote the decision boundary (given by SVMs) of two groups of points. The green lines in criterion 5 and 6 delineate the decision boundary of WC89 and HM89.

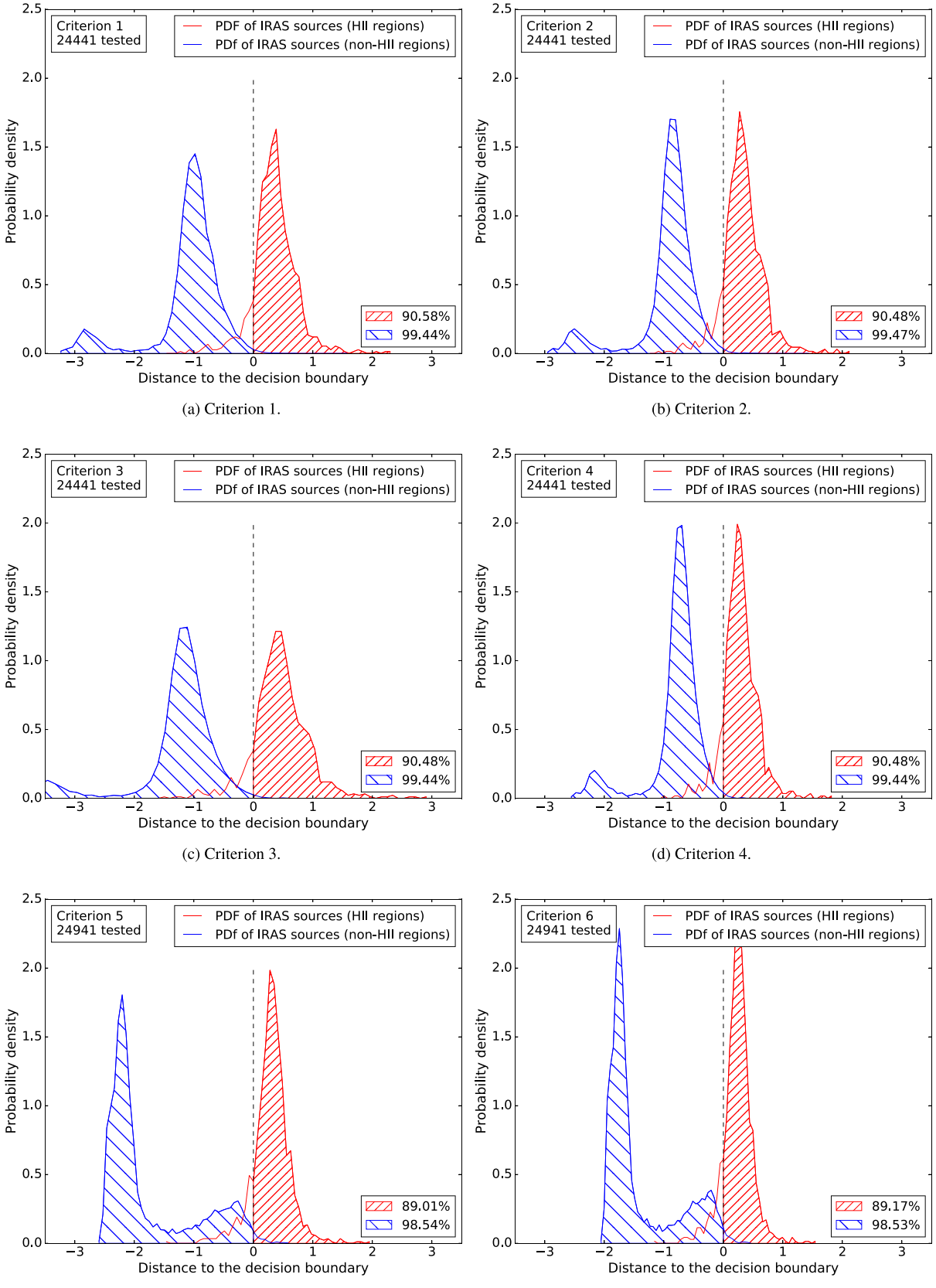


Figure 2. Probability density functions (PDFs) of the top six criteria in Table 2. The red (H II regions) and blue (non-H II regions) lines represent the probability density of *IRAS* sources with respect to the distance to the decision boundaries. Negative distances mean that *IRAS* sources disagree with the criteria. The red and blue shadowed areas represent the sensitivity (see Table 2) and specificity (TN/(TN + FP)) of criteria, respectively.

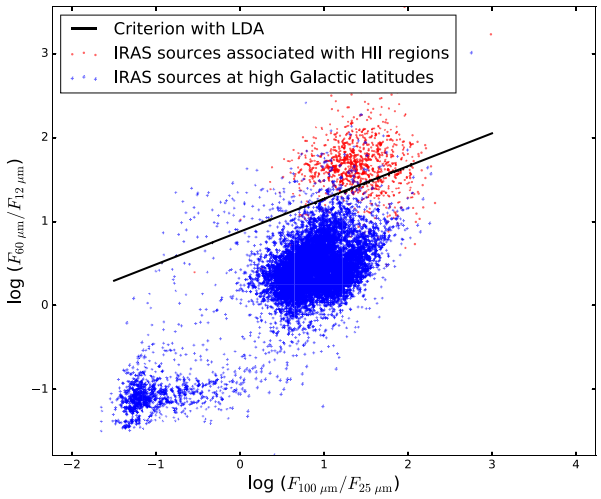


Figure 3. Decision boundary determined by LDA. The red and blue markers represent H II regions and non-H II regions, respectively.

We performed LDA with the SKLEARN package with default parameters. As an example, we display the results of colour combination of criterion 1. In Fig. 3, we illustrate the decision boundary determined by LDA. Evidently, LDA misclassifies many more H II regions and includes many more non-H II regions than the SVM algorithm (see Fig. 1a).

Consequently, the decision boundary determined by SVMs gives better results than that determined by LDA for selecting H II regions from the *IRAS* PSC.

2.5 Three-colour criteria

Because the *IRAS* PSC provides fluxes for four bands, we also examined three-colour criteria. The process of creating three-colour criteria is identical to that of two-colour criteria. We also require that the flux quality of the numerator is better than an upper limit (above the detection threshold).

The most important factor that affects the results of three-colour criteria is which bands are required to be better than an upper limit (>1). In each three-colour criterion, we already used all four bands, and the rearrangement of the colour combinations of these four bands does not lose or gain information. Consequently, if two three-colour criterion requires the qualities of same bands to be better than an upper limit (>1), they essentially give the same results, and two-colour criteria (which involve three bands) show similar results, for instance, the resemblance of criterion 2, 3, and 4 in Table 2.

We derived three-colour criteria for all 16 possible three-colour combinations. The three-colour criteria possessing the highest score (0.565) is given by

$$\begin{cases} 4.93 \times F_{60}/F_{12} + 0.17 \times F_{60}/F_{25} + 1.42 \times F_{100}/F_{60} > 6.95, \\ Q_{60} > 1, \\ Q_{100} > 1, \end{cases} \quad (3)$$

which identifies 3093 candidates. The radio association, sensitivity, and precision are 31.88, 90.48, and 93.70 per cent, respectively. Those parameters are close to that of criterion 2. However, the score is lower than criterion 1. Intuitively, three-colour criteria should do better than two-colour criteria, but the four bands are not entirely independent and the decision boundary in three-dimensional space may not be well constrained. Consequently, we only consider two-colour criteria for selecting H II regions from the *IRAS* PSC.

3 RESULTS

In this section, we examine the distribution of H II candidates selected from *IRAS* PSC, the total number of H II regions in the Milky Way, and the evolution of H II regions on the two-colour diagram.

3.1 Source distribution

We display the all-sky distribution of those sources selected by criterion 1 in Fig. 4. As illustrated in Fig. 4, the Galactic plane is clearly delineated by those 3061 H II region candidates, and the sources at high Galactic latitudes are sparse. As expected, the LMC and the SMC are evident, as well as the Orion nebula. We examined 17 selected high Galactic latitude sources ($|b| > 30^\circ$, not in SMC or LMC) and found that 14 of them have been identified as galaxies or extragalactic H II regions. The remaining three sources are one planetary nebula (PN), one post-asymptotic giant branch (AGB) star, and one unknown infrared source (IRAS 13458−0823).

In Fig. 5, we plot histograms of these sources along the Galactic latitude and Galactic longitude axes, the bin sizes of which are 1° and 2°, respectively. The distribution of H II region candidates along the Galactic latitude is approximately Gaussian, despite two prominent groups at high Galactic latitudes, associated with the Orion nebula and the LMC. In the right-hand panel of Fig. 5, two peaks are evident near 80° and 280 °, corresponding to the Local Arm and Carina Arm (Avedisova 1985; HM89; Xu et al. 2013), respectively.

After eliminating the sources in LMC and SMC, we have 2940 H II region candidates left. We further reject those *IRAS* sources that have been identified as PNe, post-AGB stars, supernova remnants, and extragalactic objects via SIMBAD catalogues, and 2805 (=2940 − 135) H II region candidates remain. 68 per cent of those remaining (2805) H II candidates are in the first and fourth Galactic quadrants (within 90° of the Galactic Centre), and 47 per cent are within 60° of the Galactic Centre. Compared with the proportion of *WISE* catalogue (86 and 76 per cent), those proportions are smaller. This is because the spatial resolution of the *IRAS* is not sufficiently good to resolve many small-angular-size H II regions in the first and fourth Galactic quadrants. This can be demonstrated by the fact that the proportions of large (radii > 1 arcmin) and small (radii < 1 arcmin) angular size *WISE* H II region candidates in the first and fourth Galactic quadrants are 76 and 98 per cent, respectively The asymmetric distribution of H II regions in Galactic latitude is also present in *IRAS* H II candidates. 53 per cent of *IRAS* H II candidates are at negative latitudes, and this is close to the value of *WISE* catalogue (56 per cent).

3.2 The total number of H II regions in the Milky Way

We estimate the total number of H II regions in the Milky Way by combining the *WISE* H II region catalogue with the H II region candidates in the *IRAS* PSC identified by criterion 1 or by WC89.

In Fig. 6, we show a Venn diagram with respect to criterion 1, WC89, and *WISE* H II region candidate catalogues. Of the 2805 H II region candidates selected according to criterion 1, 1421 candidates are positionally associated with the *WISE* H II region catalogue (within 1 arcmin). After eliminating non-H II regions, WC89 identify 1615 H II region candidates, and 993 of them agree with criterion 1. As shown in Fig. 6, the total number of H II regions in the Milky Way is approximately 10 156.

The number of ∼10 200 is a lower limit of H II regions in the Milky Way. On one hand, the sensitivity of criterion 1 suggests

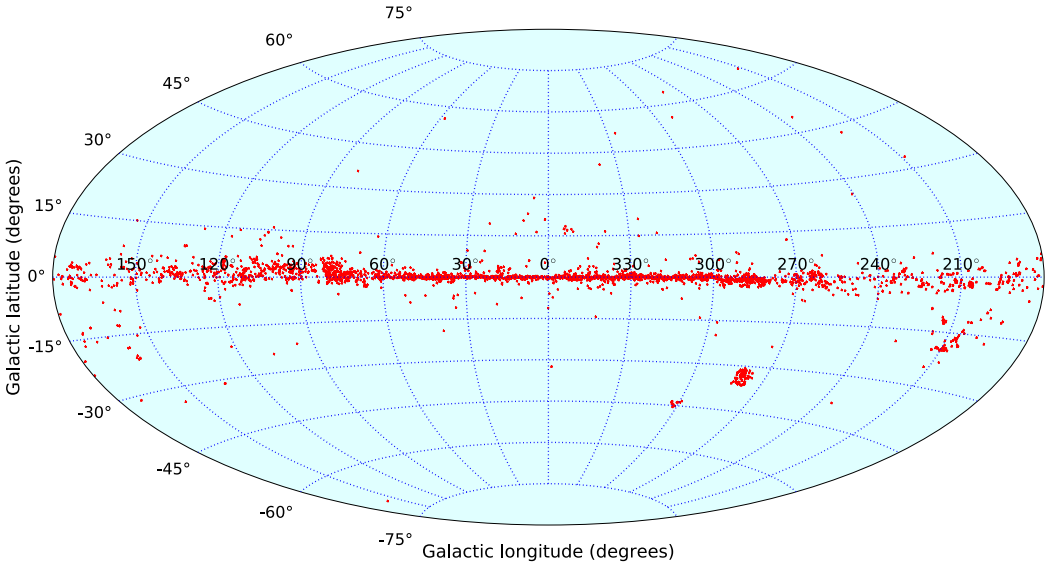
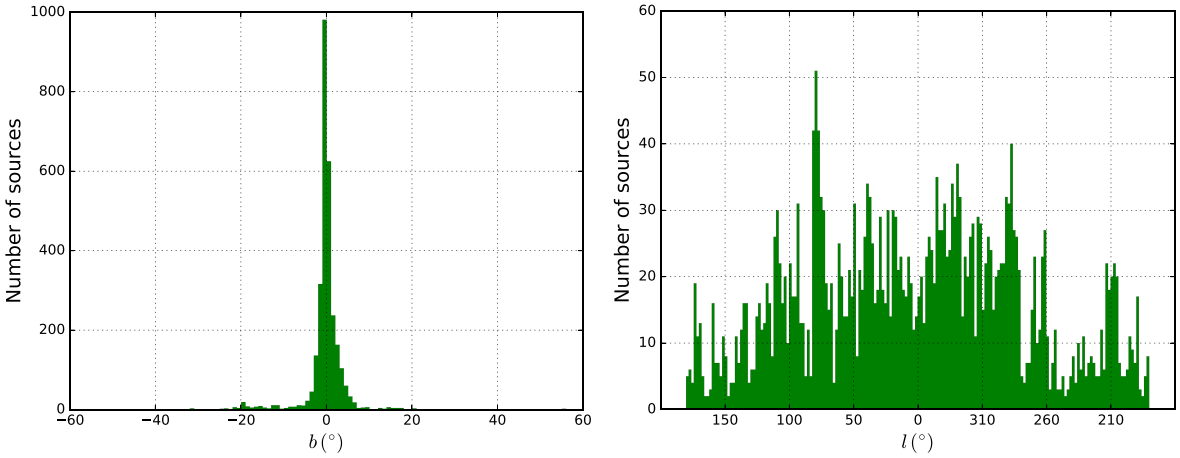


Figure 4. All-sky distribution of 3041 *IRAS* H II region candidates selected by criterion 1. After eliminating sources in LMC and SMC and that have been identified as PNe, post-AGB stars, supernova remnants, and extragalactic objects via SIMBAD catalogues, 2805 H II region candidates remain.



(a) The histogram along the Galactic latitude.

(b) The histogram along the Galactic longitude.

Figure 5. Histograms of the distribution of H II region candidates identified by criterion 1 along the Galactic latitude and longitude, the bin sizes of which are 1° and 2°, respectively.

~91 per cent of H II regions in the *IRAS* sources have been identified, missing a small amount of H II regions; on the other hand, due to the lower spatial resolution and sensitivity, the *IRAS* catalogue cannot resolve many H II regions whose angular sizes are small and may have missed some H II regions whose infrared emissions are beyond the detectability of *IRAS*.

3.3 Evolution of H II regions on two-colour diagrams

In this section, we examine the evolution of H II regions on two-colour diagrams. A three-dimensional simulation of the expansion of H II regions performed by Tremblin et al. (2014), involving the effect of the internal turbulence in surrounding gases, enables us to determine the age of *IRAS* point sources that are tracing H II regions.

The age calculation for H II regions requires the value of the distance and radio continuum flux. Anderson et al. (2014) provide distances for 1413 H II region candidates, 627 of which are associated with both *IRAS* and radio continuum sources. In order to assure the purity of H II regions, we require that the fluxes of all four bands are better than upper limits, and we calculated the ages for the remaining 305 H II regions, based on the results of the 3D simulation of Tremblin et al. (2014).

The typical age of high-mass stars is about 5 Myr (Zinnecker & Yorke 2007), and given a time-scale of 15 per cent at embedded phase (Churchwell 2002; Zinnecker & Yorke 2007), the age of H II regions should be substantially less than 5 Myr. However, considering the uncertainties of calculation, for instance, caused by inaccurate assumptions of initial densities of surrounding gases or large errors in distances, we only rejected those H II regions whose ages are greater than 10 Myr, and adopted those sources younger than 0.5 Myr as the young sample and those sources older than 6 Myr as the evolved sample. These divisions make the number of young and old samples approximate.

The infrared colours of *IRAS* point sources tend to be bluer with the evolution of H II regions. We display the result of the colour combinations of criterion 1 (left) and WC89 (right) in Fig. 7. For

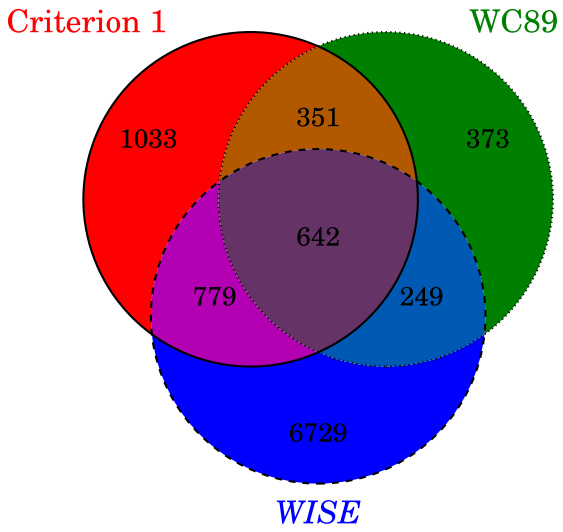


Figure 6. The Venn diagram of H II region candidates in the Milky Way. The red, green, and blue colours represent *IRAS* sources selected by criterion 1, *IRAS* sources identified by WC89 (Wood & Churchwell 1989b), and *WISE* H II region candidates (Anderson et al. 2014), respectively.

both cases, the colours of old samples are generally bluer than that of young samples. As a comparison, we plotted colours of blackbodies with different temperatures. The difference in the colour of blackbodies and H II regions indicates that a single blackbody cannot model the whole infrared spectral energy distribution (SED) of an H II region, and this agrees with the result of Walsh et al. (1999), suggesting a two-component blackbody model. This is mainly caused by the distribution of dust grain size (Natta & Panagia 1976; MacLeod & Hughes 1991; Ochsendorf & Tielens 2015), accompanied by the silicate absorption and the PAH emission (Walsh et al. 1999).

Nevertheless, the evolutionary trend is consistent with a blackbody, whose temperature is increasing with time.

4 DISCUSSION

4.1 Planetary nebulae

Dust surrounding or inside of ionized regions potentially has similar colours even if the ionized gas arises from a PN instead of an H II region. In order to check this possibility, we matched the H II region candidates (3041) to a Galactic PNe catalogue (Frew, Bojičić & Parker 2013) that contains 1258 sources. The match radius is 1 arcmin, and in total, we find 22 H II region candidates are near PNe.

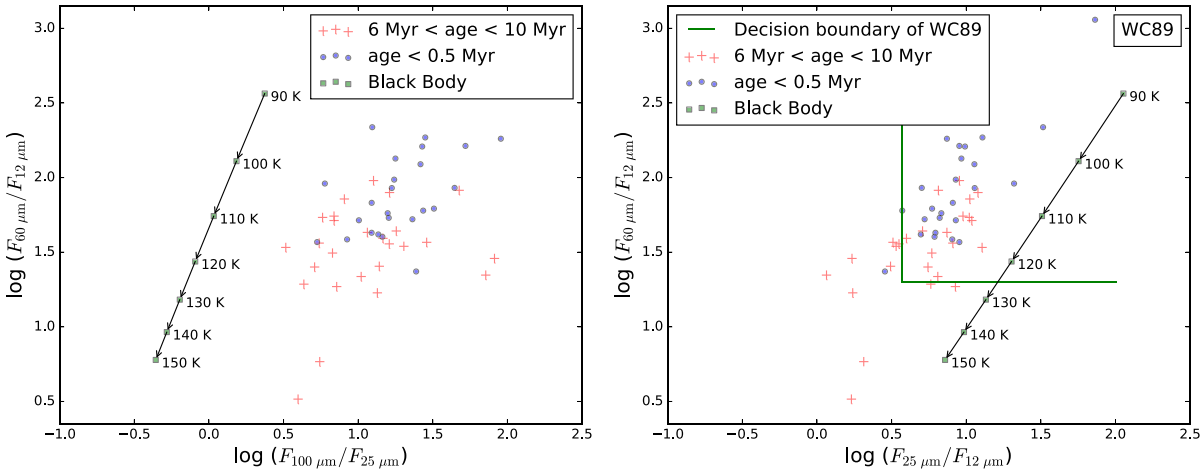
Consequently, less than 1 per cent of H II regions are located nearby PNe, indicating H II regions are well separated from PNe. 20 of those 22 *IRAS* sources are excluded in the final 2805 H II region candidates identified by criterion 1. The left two *IRAS* sources remain in the candidate list, because they are still possibly tracing H II regions despite neighbouring PNe.

4.2 The evolution of H II regions

The evolutionary trend of *IRAS* sources is consistent with the result of Xu, Zheng & Jiang (2003). Xu et al. (2003) investigated 482 6.7-GHz methanol masers, 361 of which are associated with *IRAS* sources, and they find that on the two-colour diagram, most of those *IRAS* sources concentrate in a small area, $0.57 \leqq \log(F_{25}/F_{12}) \leqq 1.30$ and $1.30 \leqq \log(F_{60}/F_{12}) \leqq 2.5$. They suggest that the infrared colours of UC H II regions move towards blue colours, which is consistent with the trending revealed by our results, as illustrated in the right-hand panel of Fig. 7.

4.3 WC89 and HM89

As shown in Figs 1(e) and 7(b), the criterion of WC89 has missed many evolved H II regions. This is because the known H II region samples used by WC89 are UC H II regions, making WC89 sensitive to UC H II regions. About half (∼900) of UC H II region candidates selected by WC89 have counterparts in the *WISE* H II region catalogue, meaning that the *WISE* H II region catalogue indeed has missed some UC H II regions.



(a) Criterion 1.

(b) WC89 (criterion 5).

Figure 7. The evolution of H II regions on two-colour diagrams. The colour combinations of (a) and (b) correspond to criterion 1 and WC89 (criterion 5), respectively. The blue points and red crosses represent young and old samples of *IRAS* sources, respectively. The age is estimated according to a 3D simulation for the expansion of H II regions (Tremblin et al. 2014). The positions of blackbodies possessing different temperatures are marked with solid green squares.

As mentioned in Section 3.2, 622 *IRAS* sources selected by WC89 disagree with criterion 1. However, this does not mean criterion 1 is not compatible with WC89, because most (544) of those 622 sources, having $Q_{100} = 1$, are not tested by the inequality of criterion 1. We conclude that the data quality of *IRAS* PSC may have affected the completeness of H II region candidates. Consequently, those 622 *IRAS* sources are counted in the total number of H II region candidates in the Milky Way.

As illustrated in Fig. 1(f), the area bounded by the decision boundary of HM89 includes a large area that is dominated by non-H II regions. The decision boundary given by HM89 is $F_{60}/F_{25} \geqq 0$ and $F_{25}/F_{12} \geqq 0$, resulted in the selection of many sources at high Galactic latitudes. However, extra constraints on F_{100} and the Galactic latitudes may have compensated for this inaccuracy.

In Fig. 6, criterion 1 independently identities 1033 *IRAS* sources, 659 of which disagree with HM89. Those 659 *IRAS* sources are those H II region candidates missed by HM89, WC89, and the *WISE* H II region catalogue.

5 CONCLUSIONS

We present new criteria for identifying H II regions from the *IRAS* PSC, and the criteria are determined by the distribution of two samples. One sample (H II regions) is produced by matching the *WISE* H II region catalogue to the *IRAS* PSC; the other sample (non-H II regions) is constructed by filtering *IRAS* sources at high Galactic latitudes ($|b| > 8^\circ$). We determined the decision boundary of the two samples using SVMs, which are efficient classifiers.

We find that the optimal selection criterion is $\log(F_{60}/F_{12}) \geqq (-0.19 \times \log(F_{100}/F_{25}) + 1.52)$, $Q_{60} > 1$, and $Q_{100} > 1$, identifying 3041 H II region candidates from *IRAS* sources, ∼660 of which are new. The known H II regions in our method are more complete than that used in HM89, and we have improved the criterion of HM89 significantly. A high proportion of *IRAS* H II region candidates have radio counterparts, providing a high confidence for those *IRAS* H II region candidates. Combining with the *WISE* H II region catalogue, we find that the lower limit of H II regions in the Milky Way is ∼10 200.

We estimate the age of some H II regions based on a 3D simulation involving internal turbulence of surrounding gases, and find that on two-colour diagrams, younger H II regions have redder *IRAS* colours and with the evolution, their infrared colours become bluer.

The SVM is an efficient classifier to identify H II regions from infrared surveys. With our methodology and future sensitive infrared (*Herschel*, for instance) and radio (recombination lines) observations, we will be able to confirm those H II region candidates and detect more H II regions, approaching the completeness of H II regions in the Milky Way.

ACKNOWLEDGEMENTS

We would like to thank an anonymous reviewer and J. R. Dawson for careful proofreading of the manuscript and constructive comments. This work was partly sponsored by the 100 Talents Project of the Chinese Academy of Sciences, the National Science Foundation of China (Grant Numbers: 11673066, 11233007, 11673051, and 11590781), the Natural Science Foundation of Shanghai under grant 15ZR1446900, and the Key Laboratory for Radio Astronomy.

REFERENCES

Anderson L. D., Bania T. M., Balser D. S., Rood R. T., 2011, ApJS, 194, 32
Anderson L. D., Bania T. M., Balser D. S., Cunningham V., Wenger T. V., Johnstone B. M., Armentrout W. P., 2014, ApJS, 212, 1
Assef R. J. et al., 2013, ApJ, 772, 26
Avedisova V. S., 1985, Soviet Astron. Lett., 11, 185
Bai Y., Zou H., Liu J., Wang S., 2015, ApJS, 220, 6
Benjamin R. A. et al., 2003, PASP, 115, 953
Bock D. C.-J., Large M. I., Sadler E. M., 1999, AJ, 117, 1578
Churchwell E., 2002, ARA&A, 40, 27
Churchwell E. et al., 2006, ApJ, 649, 759
Condon J. J., Cotton W. D., Greisen E. W., Yin Q. F., Perley R. A., Taylor G. B., Broderick J. J., 1998, AJ, 115, 1693
Cook D. O. et al., 2014, MNRAS, 445, 881
Cortes C., Vapnik V., 1995, Machine Learning, 20, 273
Frew D. J., Bojičić I. S., Parker Q. A., 2013, MNRAS, 431, 2
Fullmer L., Lonsdale C. J., 1989, Cataloged Galaxies and Quasars Observed in the IRAS Survey. JPL D-1932, Version 2, part no. 3, Jet Propulsion Laboratory, Pasadena
Garay G., Lizano S., 1999, PASP, 111, 1049
Helou G., Walker D. W. eds, 1988, Infrared Astronomical Satellite (IRAS) Catalogs and Atlases. Vol. 7, The Small Scale Structure Catalog. GPO, Washington, DC, NASA RP-1190
Hughes V. A., MacLeod G. C., 1989, AJ, 97, 786 (HM89)
Immer K. et al., 2011, ApJS, 194, 25
Karachentsev I. D., Makarov D. I., Kaisina E. I., 2013, AJ, 145, 101
Kraan-Korteweg R. C., 1986, A&AS, 66, 255
Kurtz S., Churchwell E., Wood D. O. S., 1994, ApJS, 91, 659
MacLeod G. C., Hughes V. A., 1991, AJ, 102, 658
Mauch T., Murphy T., Buttery H. J., Curran J., Hunstead R. W., Piestrzynski B., Robertson J. G., Sadler E. M., 2003, MNRAS, 342, 1117
Murphy T., Mauch T., Green A., Hunstead R. W., Piestrzynska B., Kels A. P., Sztajer P., 2007, MNRAS, 382, 382
Natta A., Panagia N., 1976, A&A, 50, 191
Neugebauer G. et al., 1984, ApJ, 278, L1
Ochsendorf B. B., Tielens A. G. G. M., 2015, A&A, 576, A2
Protheroe R. J., Ott J., Ekers R. D., Jones D. I., Crocker R. M., 2008, MNRAS, 390, 683
Rice T. S., Goodman A. A., Bergin E. A., Beaumont C., Dame T. M., 2016, ApJ, 822, 52
Simpson R. J. et al., 2012, MNRAS, 424, 2442
Tremblin P. et al., 2014, A&A, 568, A4
Vapnik V. N., 1995, The Nature of Statistical Learning Theory. Springer-Verlag, New York
Walsh A. J., Hyland A. R., Robinson G., Burton M. G., 1997, MNRAS, 291, 261
Walsh A. J., Burton M. G., Hyland A. R., Robinson G., 1998, MNRAS, 301, 640
Walsh A. J., Burton M. G., Hyland A. R., Robinson G., 1999, MNRAS, 309, 905
Wilson B. A., Dame T. M., Masheder M. R. W., Thaddeus P., 2005, A&A, 430, 523
Wolfire M. G., Churchwell E., 1994, ApJ, 427, 889
Wood D. O. S., Churchwell E., 1989a, ApJS, 69, 831
Wood D. O. S., Churchwell E., 1989b, ApJ, 340, 265 (WC89)
Wright E. L., 1981, ApJ, 250, 1
Wright E. L. et al., 2010, AJ, 140, 1868
Xu Y., Zheng X.-W., Jiang D.-R., 2003, Chin. J. Astron. Astrophys., 3, 49
Xu Y., Reid M. J., Menten K. M., Zheng X. W., 2006, ApJS, 166, 526
Xu Y. et al., 2013, ApJ, 769, 15
Zinnecker H., Yorke H. W., 2007, ARA&A, 45, 481

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