Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **505,** 1135–1152 (2021)
Advance Access publication 2021 May 21

https://doi.org/10.1093/mnras/stab1258

Population-based identification of H α-excess sources in the *Gaia* DR2 and IPHAS catalogues

M. Fratta,[1,2]★ S. Scaringi [1,2] J. E. Drew [3] M. Monguió [4,5] C. Knigge,[6] T. J. Maccarone,[2] J. M. C. Court [2] K. A. Iłkiewicz [1,2] A. F. Pala,[7] P. Gandhi [6] and B. Gänsicke [8]

[1]*Centre for Extragalactic Astronomy, Department of Physics, University of Durham, South Road, Durham DH1 3LE, UK*
[2]*Department of Physics and Astronomy, Texas Tech University, Lubbock, TX 79409-1051, USA*
[3]*Department of Physics and Astronomy, Faculty of Maths and Physical Sciences, University College London, Gower Street, London WC1E 6BT, UK*
[4]*Institut d'Estudis Espacials de Catalunya, Universitat de Barcelona (ICC-UB), Martí i Franquès 1, E-08028, Spain*
[5]*Universitat politècnica de Catalunya, Departament de Física, c/Esteve Terrades 5, E-08860 Castelldefels, Spain*
[6]*School of Physics and Astronomy, University of Southampton, University Road, Southampton SO17 1BJ, UK*
[7]*European Southern Observatory, Karl Schwarzschild Strasse 2, Garching bei Munchen, D-85748, Germany*
[8]*Astronomy and Astrophysics Group, Department of Physics, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK*

Accepted 2021 April 27. Received 2021 April 23; in original form 2021 February 10

ABSTRACT

We present a catalogue of point-like H α-excess sources in the Northern Galactic Plane. Our catalogue is created using a new technique that leverages astrometric and photometric information from *Gaia* to select H α-bright outliers in the INT Photometric H α Survey of the Northern Galactic Plane (IPHAS), across the colour–absolute magnitude diagram. To mitigate the selection biases due to stellar population mixing and to extinction, the investigated objects are first partitioned with respect to their positions in the *Gaia* colour–absolute magnitude space, and Galactic coordinates space, respectively. The selection is then performed on both partition types independently. Two significance parameters are assigned to each target, one for each partition type. These represent a quantitative degree of confidence that the given source is a reliable H α-excess candidate, with reference to the other objects in the corresponding partition. Our catalogue provides two flags for each source, both indicating the significance level of the H α-excess. By analysing their intensity in the H α narrow band, 28 496 objects out of 7474 835 are identified as H α-excess candidates with a significance higher than 3. The *completeness* fraction of the H α outliers selection is between 3 and 5 per cent. The suggested 5σ conservative cut yields a *purity* fraction of 81.9 per cent.

Key words: techniques: photometric – catalogues – stars: emission-line, Be – Hertzsprung–Russell and colour–magnitude – novae, cataclysmic variables.

1 INTRODUCTION

H α emission can be observed from both extended sources, such as nebulosities associated with either star-forming regions and/or stellar remnants, and from point-like sources, with no associated extended emission. These latter objects can fall into different source-types and can span various evolutionary stages of stellar populations. The many classes of H α emitting point-like sources include (but are not limited to) a wide range of young stellar-objects (YSOs), classical Be stars, compact planetary nebulae, luminous blue variables (LBVs), hypergiants, Wolf–Rayet stars, and rapidly rotating stars. Furthermore, many interacting binary systems exhibit H α in emission due to accretion (e.g. cataclysmic variables, CVs; symbiotic stars, SySt; or binary systems in which the accreting compact object is a black hole or a neutron star). The H α emitting population is heterogeneous and challenging to identify. Because of this, samples of these objects are plagued by selection biases, which, in turn, prevent stellar evolution models from being adequately tested.

Large, wide-field, high-angular-resolution H α imaging surveys provide the basis to discover and characterize H α-excess sources. Among the previous surveys targeting the ionized diffuse interstellar medium (ISM) that have aimed to increase the sample of known H α sources, we can include, for instance, the H α observations of the Large and Small Magellanic Clouds (Davies, Elliott & Meaburn 1976). However, this survey only observed small patches of the sky, and the limiting magnitude was quite stringent. On the other hand, the Virginia Tech H α and [S II] Imaging Survey of the Northern Sky (VTSS, Dennison, Simonetti & Topasna 1999) and the Southern H α Sky Survey Atlas (SHASSA; Gaustad et al. 2001), covered wider areas of sky, but they suffered from relatively poor angular resolution. Among the imaging surveys that focused on point sources, Kohoutek & Wehmeyer (1999) obtained a list of ∼4000 point-like H α emitters located in the northern Galactic plane ($|b| \leq 10^{\circ}$). Parker et al. (2005), with their Anglo-Australian Observatory/UK Schmidt Telescope (AAO/UKST) SuperCOSMOS H α Survey (SHS), inspected an area of ∼4000 deg^2 in the Southern Milky Way, plus an additional ∼700 deg^2 area around the Magellanic Clouds.

The Isaac Newton Telescope (INT) Photometric H α Survey of the Northern Galactic Plane (IPHAS; Drew et al. 2005) provides photometry with the 2 broad-band r and i filters, as well as with

★ E-mail: matteo.fratta@durham.ac.uk

the narrow-band Hα filter (see also Drew et al. 2005 and Irwin & Lewis 2001). Witham et al. (2008) used the IPHAS pre-publication photometric measurements (without a uniform calibration) to identify candidate Hα emission-line sources. Their method is based on producing two-colour diagrams (TCDs) for each IPHAS field, using $r - \mathrm{H}\alpha$ and $r - i$, respectively, as vertical and horizontal axes. Each Wide Field Camera (WFC) pointing covers an area of 0.22 deg^2 in the sky. Hα line excess source candidates are then selected by iteratively fitting the stellar locus and retaining positive outliers in $r - \mathrm{H}\alpha$. This procedure is performed within pre-defined magnitude ranges to try and mitigate the effect of extinction, which can become substantial when looking through the Galactic plane (Sale et al. 2014). Using their conservative method, Witham et al. (2008) identified in total 4853 Hα emitting candidates. Only a small fraction of these candidates could be confirmed through a comparison with previously developed narrow-line emitters catalogues. A spectroscopic follow-up (presented in Raddi et al. 2013) was then performed on 370 outliers with $r < 18$, and 97 per cent of them did show Hα emission lines.

More recently, Monguió et al. (2020) developed the IGAPS (INT Galactic Plane Survey) catalogue, that includes ∼295 million objects. Of these, 53 234 833 (18 per cent) unblended sources with $r < 19.5$ mag were tested for Hα-excess. IGAPS consists of a cross-match between IPHAS and UVEX (the UV-Excess survey of the Northern Galactic Plane; Groot et al. 2009). With the use of the 2.5-m INT, the latter survey provides photometric measurements for the sources included in a $10^\circ \times 185^\circ$ sky area, centred on the Galactic equator. More specifically, it provides U, g, and r intensities, with a limiting magnitude of 21–22 mag. The g, r, and i magnitudes in IGAPS were calibrated with reference to the 'Pan-STARRS photometric reference ladder' (Magnier et al. 2013), while the Hα narrow-band calibration was based on the methods described in Glazebrook et al. (1994).

In the context of IPHAS, the main metric for Hα-excess is $r - \mathrm{H}\alpha$. However, this colour-index is not quite constant for stars without emission lines, but varies as a function of the spectral type. Without first confining distinct populations, the measured Hα excess of a star in the IPHAS TCD cannot have a consistent relation with the net emission equivalent width, and candidates can remain lost in the main stellar locus. For this reason, population-based Hα-excess selections generally produce more complete results. An example of such study is presented in Mohr-Smith et al. (2017): These authors performed their selection of Hα-excess candidates on a set of previously identified O and early B stars, across the Carina Arm. Their goal was an assessment of the relative frequency of the classical Be (CBe) phenomenon in the VST Photometric Hα Survey of the Southern Galactic plane and Bulge (VPHAS + ; Drew et al. 2014) field of view.

Without any knowledge of the distances, and using only IPHAS measurements, degeneracies may exist in associating a particular object with a specific stellar population. Because of this, the emission-line candidate lists of Witham et al. (2008) and Monguió et al. (2020) are necessarily conservative and incomplete. In our work, Hα line excess candidates are identified from IPHAS survey by using two independent and complementary methods: (a) selecting Hα-excess sources relative to nearby sources in the calibrated *Gaia* colour–absolute magnitude diagram (CAMD), and (b) selecting Hα-excess sources relative to groups of objects that occupy nearby positions in the sky. It is relevant to stress the fact that the objects that are labelled as Hα line excess candidates in this study are not necessarily Hα emitters; the only conclusion that can be reached through this selection process is that their Hα intensity is higher than that associated with objects they are compared to.

The input catalogue used in this work to identify Hα-excess sources is that of Scaringi et al. (2018) (hereafter *Gaia*/IPHAS catalogue), which is the result of a positional sub-arcsec cross-match between the sources in the *Gaia* and IPHAS DR2 fields of view. When performing the cross-match, Scaringi et al. (2018) took into account the proper motions provided by *Gaia* in order to rewind the positions of the objects back to the IPHAS DR2 observation epoch. This catalogue contains a list of approximately 8 million sources, all found in the Northern Galactic plane.

In Section 2, a more detailed description of the input catalogue is provided. Section 3 consists of an explanation of our selection process. The results obtained by our algorithm are presented in Section 4. In Section 5, these results are discussed. Section 6 presents two possible science cases. In Section 7 we draw our conclusions.

2 THE INPUT CATALOGUE

The targets in the *Gaia*/IPHAS catalogue occupy an area of the sky included between $|b| \leq 5^\circ$ and $29^\circ \leq l \leq 215^\circ$, and are mostly found within a distance radius of ∼1.5 kpc from us. These distances are calculated directly as the inverse of *Gaia* parallax measurements, with the caveat that they satisfy the $parallax\_over\_error > 5$ criterion (Scaringi et al. 2018; median parallax uncertainties as well as the systematic parallax offset are discussed in Lindegren et al. 2018). The choice of inferring the distances via parallax inversion is justified by Scaringi et al. (2018) with the introduction of two parameters that quantify the goodness of *Gaia* astrometric fit and the false-positive rate: f_c and f_{FP}, respectively. To compute f_c, they binned the targets according to their *Gaia* G-band magnitudes; f_c corresponds to the percentile assigned to each object in the bin, with respect to the χ^2 of the astrometric fit. On the other hand, f_{FP} reflects the presence of spurious negative parallaxes in *Gaia* measurements, due to poor astrometric fits. To obtain f_{FP}, Scaringi et al. (2018) first produced a mirror sample of their catalogue, including only objects with negative parallaxes, with '$parallax\_over\_error < -5$'. They thus binned the objects in the catalogue (including the mirror sample) with respect to their G-band measurements, and further with respect to the χ^2 of their astrometric fit. They thus define f_{FP} as the fraction of objects from the mirror sample (false positives) in each bin.

To obtain the absolute magnitude for the *Gaia* G band (M_G), Scaringi et al. (2018) used the distances calculated with the parallax-inversion method. Despite the precautions taken, this approximation contributes to the uncertainties on M_G. However, the effects on M_G introduced by the use of parallax-inversion method instead of probabilistic methods (Astraatmadja & Bailer-Jones 2016) to obtain the distances are generally negligible. In fact, 97.2 per cent of the objects in our meta-catalogue fall in the $|\delta_{M_G}| \leq 0.1$ mag range, δ_{M_G} being the difference between the G-band absolute magnitudes obtained with the parallax-inversion defined distances and with probabilistically defined distances.[1]

Besides the errors on *Gaia* photometric measurements and the effects connected to the parallax-inversion defined distances, the location of the sources in the CAMD is also affected by the different extinctions that alter their colours. All these uncertainties may be the causes of stellar population mixing. Our approach to overcome this obstacle is presented in the last paragraph of Section 3.1.1.

2.1 Additional data quality constraints

This work focuses on the subset of targets from the *Gaia*/IPHAS catalogue that pass strict quality criteria, in order to minimize the inclusion of spurious cross-matches. Some quality control selection

[1]The absolute value of the maximum difference is $|\delta_{M_G,\mathrm{max}}| = 0.36$ mag.

cuts have already been applied during the compilation of the *Gaia*/IPHAS catalogue, which are mostly aimed at retaining only those sources with good *Gaia* parallax measurements and good IPHAS photometry. Additional cuts are applied here, in order to

(i) remove sources with low-quality astrometric fits and/or high false-positive probabilities (see Section 2);
(ii) remove targets close to t he saturation limit of IPHAS;
(iii) only retain targets for which we have a valid measurement in each band of interest (r, i, H α, M_G, G_{BP}, and G_{RP}).

The following cuts are thus applied:

(i) retain sources that satisfy both $f_c < 0.98$ and $f_{FP} \leq 0.02$ (as suggested in Scaringi et al. 2018);
(ii) retain sources with $r \geq 13$ mag, $i \geq 12$ mag, and H $\alpha \geq 12.5$ mag;
(iii) retain only sources with measurements in all r, i, H α, M_G, G_{BP}, and G_{RP} bands.

These cuts yield 7474 835 sources out of the original 7927 224. Fig. 1 shows the *Gaia* CAMD (i.e. the *Gaia* Hertzsprung–Russell diagram, HRD) and IPHAS two-colour diagram with the targets that pass the additional quality cuts.

3 SELECTING H α-EXCESS SOURCE CANDIDATES

The aim of this work is to identify H α-excess candidates in a vast sample of objects. This task is achieved by selecting 'positive outliers' in the $r - H\alpha$ versus $r - i$ two-colour space. To mitigate the selection biases due to stellar population mixing and to Galactic extinction, the sources in the master-catalogue are first partitioned with respect to their positions in the *Gaia* CAMD and Galactic coordinates space, respectively. The two $r - H\alpha$ outliers selections, performed on the CAMD-based and on the coordinates-based (or also 'positional'-based) partitions, are independent and complementary. The selection strategy performed on the coordinates-based partitions hinges on the one applied by Witham et al. (2008). We point out that our CAMD-based selection can still be improved, since some populations may overlap in the colour–absolute magnitude space.

It is worth pointing out that other techniques using more novel machine learning approaches could be employed for the selection of H α-excess sources. Our choice of a more rational approach is based on the relative simplicity of the algorithm, which allows to locate exactly in which partition a specific source has been selected from. Furthermore, the approach used here allows us to examine and understand the underlying population used to infer the H α-excess significance values.

The separation of the sources in the two parameter spaces is described in Section 3.1, whilst the proper selection of H α-excess sources is discussed in Section 3.2.

3.1 Partitioning algorithms

3.1.1 CAMD-based partitions

Using the calibrated *Gaia* CAMD shown in the top panel in Fig. 1, subsets (i.e. the partitions) are defined such that they (a) contain a large enough number of sources (500) to be able to statistically identify outliers and (b) are small enough in the colour–absolute magnitude space to make the underlying source population as homogeneous as possible. To balance these two requisites, an iterative method is applied.

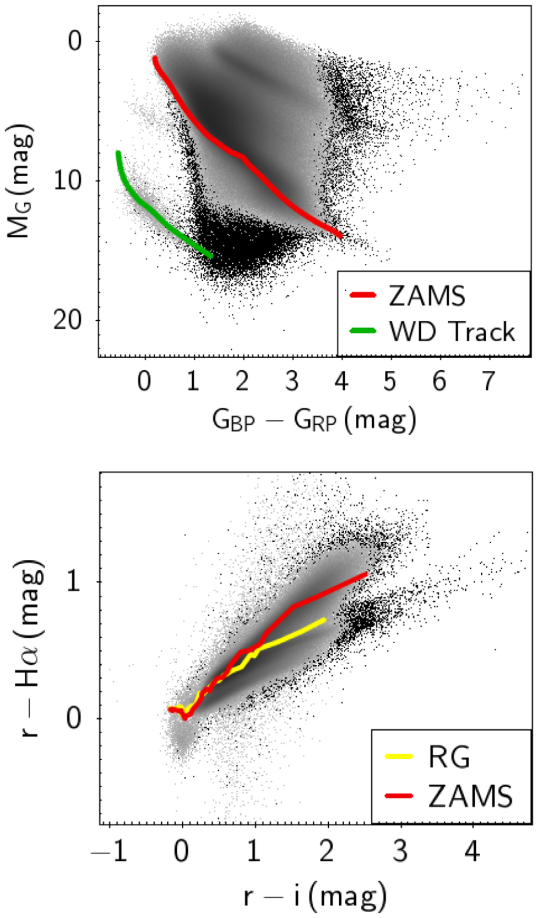


Figure 1. Positions in the *Gaia* M_G versus $G_{BP} - G_{RP}$ CAMD (top panel) and IPHAS $r - H\alpha$ versus $r - i$ TCD (bottom panel) of the sources in the *Gaia*/IPHAS catalogue (Scaringi et al. 2018). The grey dots represent the objects that satisfy the quality constraints described in Section 2.1, while the targets that do not pass this first selection are displayed with the black dots. The red and the green lines in the top panel represent respectively the synthetic zero age main-sequence (ZAMS) track (Bressan et al. 2012) and the synthetic white dwarfs track (Carrasco et al. 2014). The red line and the yellow line in the bottom panel (both taken from Drew et al. 2005) depict, respectively, the synthetic ZAMS track for zero reddening and the synthetic red giant (RG) track, in this parameter space.

First, a fine grid of 840 × 840 equally spaced 'elemental' cells is generated, covering the whole CAMD (each elemental cell with dimensions $l_x \sim 0.007$ mag and $l_y \sim 0.024$ mag, respectively). No elemental cells contain enough objects to be considered a partition. The side lengths of the grid cells are then increased to the next integer divisor of 840, in units of l_x and l_y, respectively. The second iteration produces 420 × 420 cells, 4852 of which satisfy the criteria to become partitions (these belong to the densest regions of the CAMD). These partitions are labelled according to the order by which they are generated during the current iteration (left to right, top to bottom), from 0 to 4851. The iterations carry on for all the integer divisors of 840, between 2 and 60.[2] At the end of the iterative procedure, 204 459 objects are still left without a partition assignment. These 'leftovers' are assigned to the closest partition. 9181 *CAMD − partitions* result from this process, with a maximum density of 1514 sources per

[2]Caveat: each partition must not be completely surrounded by another one; furthermore, all the elemental cells within each partition must be contiguous.

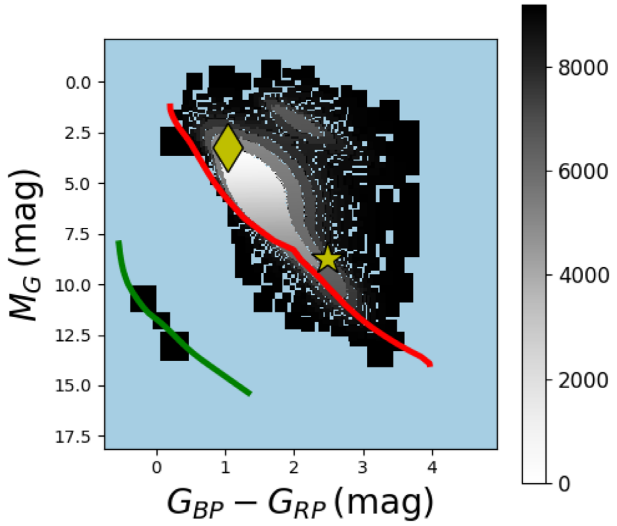


Figure 2. Map of the partitions in the *Gaia* CAMD. The colour code refers to the order in which the partitions were created (no partitions are assigned to the light blue area). The area covered by the single partition increases where the density of sources decreases. The red line represents the synthetic ZAMS track (Bressan et al. 2012), while the green line depicts synthetic white dwarfs track (Carrasco et al. 2014). The yellow star points to partition 7331, while the diamond refers to partition 0, which are discussed in Section 3.2.

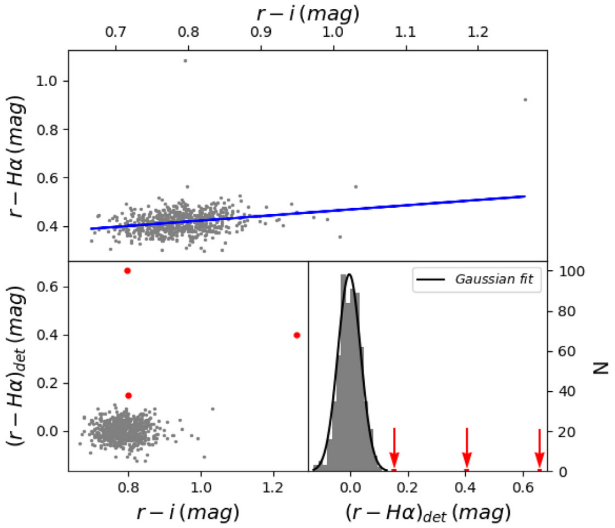


Figure 3. Graphical depiction of the detrending and outliers selection processes performed on CAMD-partition 7331. The top panel shows the corresponding non-detrended IPHAS TCD, in which only one linear trend is clearly visible. The blue line depicts the best-fitting linear model. It was subtracted from the data points to obtain the detrended $r - \mathrm{H}\alpha$ parameter (bottom left-hand panel). The red dots represent the positive outliers of the distribution. The bottom right-hand panel shows the detrended $r - \mathrm{H}\alpha$ distribution of this partition. The Gaussian behaviour of the underlying population is well described by the best-fitting model (the black solid line). The red arrows point to the three outliers of the distribution. The position of the CAMD-partition 7331 in the *Gaia* CAMD is shown in Fig. 2.

partition. The map of the resulting partitions in the CAMD is shown in Fig. 2

To account for the uncertainty on the positions of the objects in the CAMD, this partitioning process is repeated upon change of the side lengths of the elemental cells. As an example, a 20 per cent increase of the side lengths of the elemental cells produces a 0.8 per cent variation in the number of selected outliers, meaning that our selection is independent (to a reasonable extent) on the size of the elemental cells.

3.1.2 Coordinates-based partitions

For this different partitioning algorithm, an evenly spaced grid in the b versus l space is created. The size of each cell is $1.205 \times 1.004\,\mathrm{deg}^2$, i.e. about five times bigger than the 'cell size' used by Witham et al. (2008) (which performed their selection on an IPHAS field-by-field basis), and is chosen so that all the cells are either empty, or contain at least 500 objects. This procedure results in 1674 *positional − partitions*, with a maximum density of 12 604, and a minimum density of 546 objects per partition.

3.2 Detrending and identification of outliers.

The $r - \mathrm{H}\alpha$ versus $r - i$ TCDs are used to identify $\mathrm{H}\alpha$ line excess sources from every partition. First the main stellar population locus is found in each partition by iteratively fitting a line to the data, and applying Chauvenet's criterion. The latter consists of calculating a threshold[3] beyond which only outliers are expected to be found. The outliers are removed from the data at the end of each iteration. In theory, in order to tackle the population-mixing issue, the fit should be forced to the upper branch in the TCD (as done by Witham et al. 2008) by removing only the negative outliers. However, for most of the partitions, the resulting best-fitting line does not deviate sensibly from the model obtained by the direct application of the unmodified Chauvenet's criterion.

Once the stellar locus has been located, it is used as a baseline to identify the outliers: each TCD is detrended by subtracting the corresponding linear model from the data. A second iterative application of Chauvenet's criterion on the detrended TCD enables us to isolate the outliers, and hence to calculate the standard deviation (*rms*) of the remaining sources. The objects that satisfy the following relation are selected as $\mathrm{H}\alpha$-excess candidates, from either the CAMD-based and/or the positional-based partitions:

$$\sigma = \frac{y}{\sqrt{(\delta y)^2 + (m_{\mathrm{fit}} \times \delta x)^2 + rms^2}} \geq 3. \quad (1)$$

Here, y corresponds to the $(r - \mathrm{H}\alpha)_{\mathrm{detrended}}$ intensity, δy is the instrumental uncertainty on this value, δx is the instrumental error on the $r - i$ intensity, and m_{fit} is the slope of the best-fitting line. Thus defined, σ, or *significance*, represents the confidence that each source is an outlier of the corresponding distribution. Since the partitioning process is implemented in two different parameter spaces, two significances are assigned to each source: *CAMD − significance* (σ_{CAMD}) and *POS − significance* (σ_{POS}). Objects that satisfy relation (1), either from the CAMD-based and/or from the positional-based selection, will henceforth be referred to as '3σ outliers'. Fig. 3 provides a graphical depiction of the detrending (top panel) and selection (bottom left-hand panel) processes relative to the CAMD-partition 7331, as an example of a well-behaved partition.

We point out that Chauvenet's criterion assumes an underlying Gaussian population, while a non-negligible amount of our partitions seems to deviate from this (mainly due to population mixing). However, the application of Chauvenet's criterion on non-Gaussian

[3]Chauvenet's threshold depends on the root mean square of the distribution and on the number of objects that constitute such distribution.

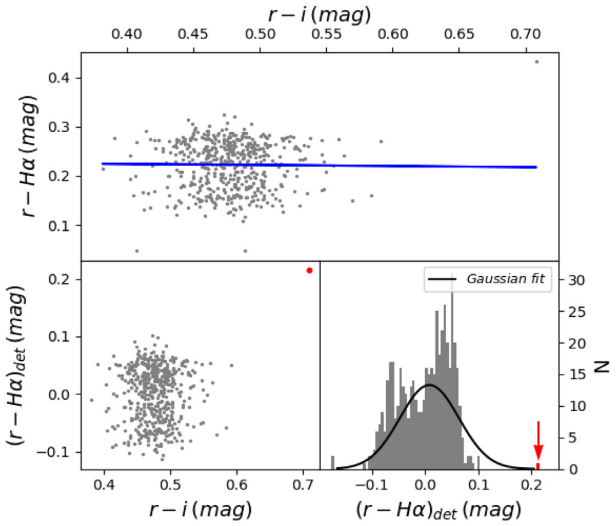


Figure 4. Graphical depiction of the detrending and outliers selection processes performed on CAMD-partition 0. As it stands out clearly, the Gaussian model is not a good fit to the underlying population. The position of partition 0 in the *Gaia* CAMD is shown in Fig. 2.

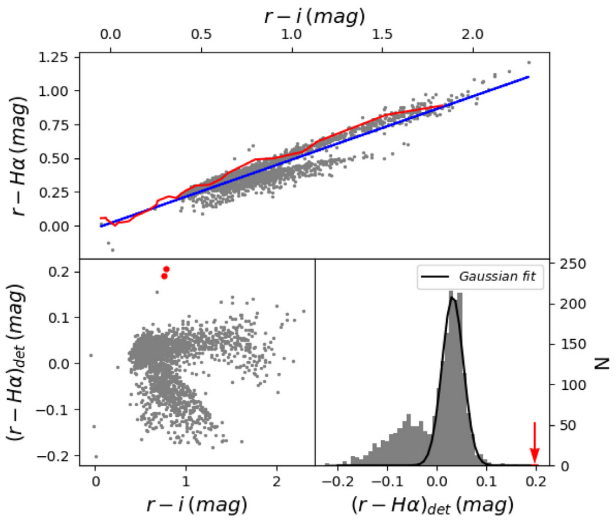


Figure 5. Graphical depiction of the detrending and outliers selection processes performed on positional-partition 154. The red line in the top panel represents the synthetic ZAMS track, for zero reddening (Drew et al. 2005).

partitions provides a more robust H α-excess outlier selection, since the standard deviation of these partitions is overestimated. As can be noticed from the bottom right-hand panel of Fig. 3, the detrended $r - \mathrm{H}\alpha$ distribution relative to the CAMD-partition 7331 constitutes a good example of Gaussian underlying population. On the other hand, Fig. 4 presents an example of partition (CAMD-partition 0) in which the underlying distribution deviates from a standard Gaussian distribution. As a reference, Fig. 5 presents the TCDs (before and after the detrending process) and the histogram of the detrended $r - \mathrm{H}\alpha$ values relative to positional-partition 154. Two trends are identifiable from the TCDs: the top one represents the locus in which unreddened MS stars lie, while the bottom trend corresponds to the reddened RG track. These two trends reflect in the bimodality recognizable in the histogram in the bottom right-hand panel. This effect does not alter the number of outliers selected from partitions

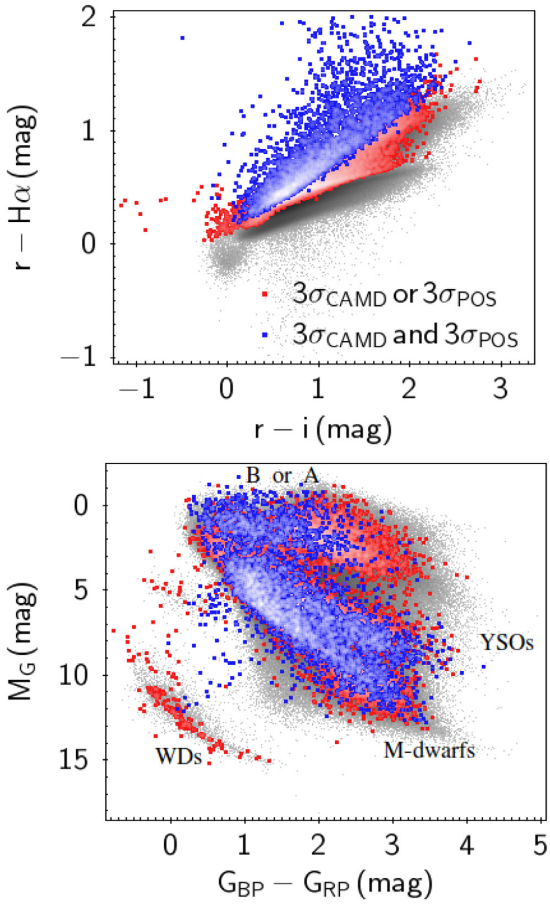


Figure 6. The top panel shows the layout of the $r - \mathrm{H}\alpha$ 3σ outliers in the IPHAS TCD, while their position in the *Gaia* CAMD is presented in the bottom panel. The red dots represent all the 3σ outliers identified by either our CAMD-based or positional-based selection, while the blue dots represent the subset of 5084 outliers selected from both the partition types. The intensity of both these colours scales inversely with the density of objects.

that present it, since the second Gaussian population is always redder than the main one.

Our algorithm selects both positive and negative outliers; however, since our goal is to identify the H α-excess candidates, the term 'outliers' will henceforth refer to only the positive ones.

4 RESULTS

Our selection identifies 28 496 $r - \mathrm{H}\alpha$ 3σ outliers (0.4 per cent of the total data set) above the previously identified stellar loci. More specifically, 25 030 outliers are selected from the CAMD-partitions and 8550 from the positional-partitions.

In Fig. 6, the locations in the IPHAS TCD and *Gaia* CAMD of these outliers are presented. It appears particularly noticeable in the top panel that many of these candidates would have not stood out as outliers, if the chosen statistical analysis had been applied directly in the two-colour domain. From the bottom panel, mainly four regions of the CAMD with a particularly high density of outliers can be highlighted: the white dwarfs (WDs) track, the M-dwarfs area, the YSOs region, and the region where reddened early MS stars of spectral type B or A sit.

Fig. 7 displays the map of the fraction of outliers per CAMD-partition. Mainly two regions with a relatively high fraction of

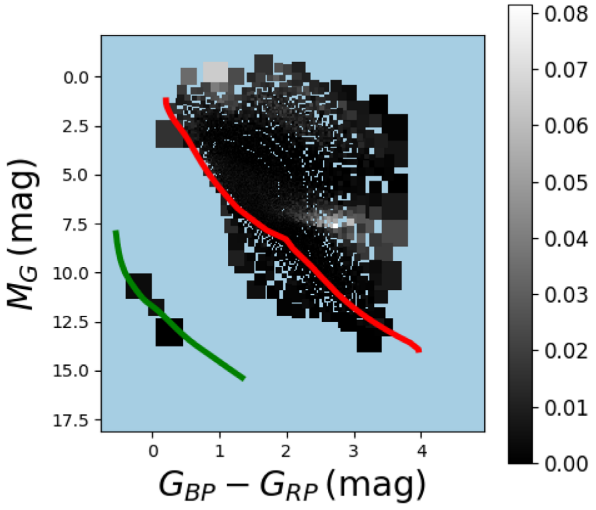


Figure 7. Fraction of outliers per CAMD-partition.

outliers stand out. These regions are: the area commonly associated with YSOs, i.e. to the right of the ZAMS (the red line) and centred at around $M_G = 7.5$ mag, and the region where reddened, bright, B or A spectral types stars lie (i.e. the top area in the CAMD, centred around $G_{BP} - G_{RP} = 1.2$ mag). Further information about the statistical composition of the H α-excess candidates that occupy these areas of the CAMD can be obtained through a cross-match with SIMBAD data base (Wenger et al. 2000). Out of 981 outliers that occupy the former overpopulated region, 608 are classified as YSOs (or candidates), or T-Tauri stars; 154 of them are classified as emission-line objects, while 146 simply as 'Star'.[4] Among the outliers included in the latter group of interest, 131 find a classification in SIMBAD: 55 of them are identified Be stars (or candidates), 34 are emission-line stars, 24 are classified as 'Star', and 10 are red giant branch stars.

In Fig. 8, the map of the fraction of outliers per positional-partition is shown. To rule out systematic effects, an analysis of the relationship between the size of each partition and the fraction of outliers within it was performed; no such correlation was found. The distribution of this ratio in the Galactic coordinate space is consistent with being homogeneous, with no significant trend in either direction. None the less, some areas in the b versus l diagram with a relatively high density of H α-excess candidates can be highlighted. These might correspond, say, to regions with a high rate of star formation, such as molecular clouds, or to open clusters. Two examples are the known open clusters IC1396 (centred at $l = 99.^{\circ}30$, $b = 03.^{\circ}74$; Kharchenko et al. 2013) and NGC 2264 (centred at $l = 202.^{\circ}94$, $b = 02.^{\circ}30$; Dias et al. 2014; Kuhn et al. 2019; Barentsen et al. 2013), which are easily identifiable in Fig. 8. The latter star-forming region has been the subject of previous studies, such as the one presented by Barentsen et al. (2013). They applied the method of the *Bayesian inference* to identify 115 accreting objects in NGC 2264. Positional-partition 1289 (highlighted by the yellow circle in Fig. 8), corresponds to the sky area in which NGC 2264 is located, and is in fact the positional-partition with the highest fraction of H α-excess candidates: out of 2826 objects, our algorithm selects 71 outliers (2.5 per cent). Nine positional-partitions centred around this partition are shown in Fig. 9. The apparently empty areas in the sky are due to the

[4]We point out that the generic 'Star' label in SIMBAD refers to objects that have been identified, but for which there is not enough information for a more specific classification.

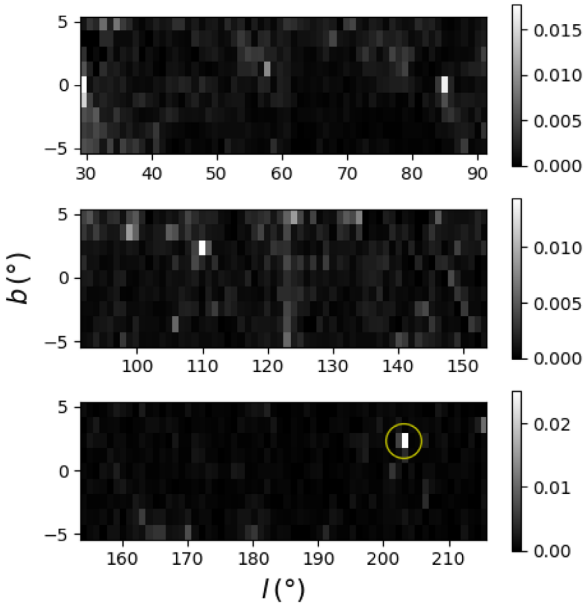


Figure 8. Fraction of outliers per positional-partition. Positional-partition 1289 is highlighted by the yellow circle.

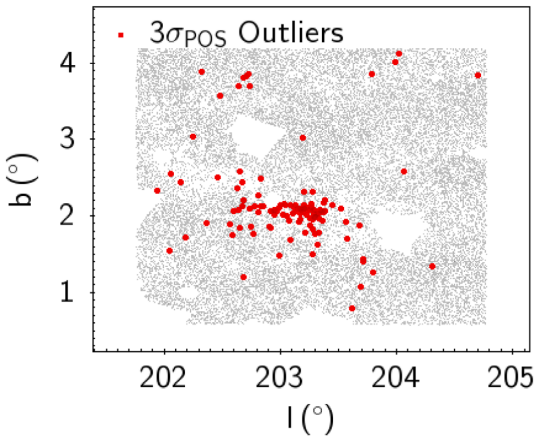


Figure 9. Graphical depiction of the nine positional-partitions centred around positional-partition 1289 in the Galactic coordinates space. This latter partition is the one with the highest fraction of H α-excess candidates (the red dots).

quality cuts applied when compiling IPHAS DR2; because of these cuts, IPHAS DR2 provides photometric measurements for sources covering 92 per cent of its footprint (Barentsen et al. 2014).

Ideally, the concept of 'outliers of a distribution' would be non-arbitrary. However, realistically speaking the definition of 'outlier' is strongly dependent on the chosen threshold. This can be mitigated by the choice of different confidence levels during the selection process. One possibility consists in considering as outliers all the objects that are selected using Chauvenet's criterion: this would be ideal if all partitions were to display Gaussian distributions. On the other hand, one can choose to select as outliers all the objects that satisfy $\sigma \geq 3$ (equation 1); this constitutes a more relaxed threshold, when compared to Chauvenet's one. We point out that by setting this threshold, a certain amount of false-positives in our selection is to be expected. However, this amount is not easily quantifiable, since the distributions are not always

Gaussian, and also they are not equally populated. The suggested threshold is the 5σ one ($\sigma \geq 5$), as a compromise to reduce false-positives fraction, while retaining a robust candidate selection. In fact, all the candidates selected using a 5σ threshold would have been included using Chauvenet's criterion as well. By applying the 5σ cut, 6774 outliers (0.09 per cent of the complete data set, 23.8 per cent if compared to the 3σ sample) are identified: 6455 from the CAMD-partitions and 2209 from the positional-partitions.

For all the sources in the master-catalogue, the two *flagCAMD* and *flagPOS* specifications are evaluated. These entries can assume a value of 0 (if the significance is lower than 3), 1 (if the significance is greater than or equal to 3, but smaller than 5), or 2 (if the significance is equal to or greater than 5). A very similar classification of the significance levels was previously adopted by Witham et al. (2008) and Monguió et al. (2020).

Our results are presented in a *meta-catalogue*, the first 10 rows of which are shown in Table 1. The full set of metrics computed during the catalogue generation is also published. In this full version, the necessary pieces of information to trace back each source to the corresponding CAMD-based and positional-based partitions are provided, as well as the detrending model information for each TCD. Our hope is that this additional information will aid future users of the catalogue to further tune the selection of H α-excess sources to suit a specific task.

5 DISCUSSION

In Fig. 10, the positions of our 3σ outliers in the CAMD (left-hand column) and TCD (right-hand column) are shown. The most evident differences between the selections applied on the CAMD-partitions and on the positional-partitions are: as follows

(i) The CAMD-based selection is less efficient, compared to the positional-based one, in identifying outliers along the WD track. This effect stands out clearly from the comparisons of both the TCDs and the CAMDs. It is due to the constraints set when partitioning the *Gaia* CAMD: the partitions in the least populated regions of the CAMD have to be large enough in size to contain at least 500 sources each. For this reason, our algorithm creates only three large partitions in the WD track and surrounding area of the CAMD. This results in a limited amount of detected outliers.

(ii) The CAMD-based selection identifies more H α-excess candidates in the region of the CAMD where the reddened B and A types emission line stars are, if compared to the positional-based algorithm.

(iii) Most M-dwarf H α-excess candidates are identified mainly through the positional-based selection. We believe that this is related to the fact that M-dwarf stars have an intrinsically more intense $r - \mathrm{H}\alpha$ IPHAS colour, compared to other MS stars. In contrast with the CAMD-partitions, in the positional-partitions these objects are blended with other populations, and hence they stand out as outliers.

(iv) The CAMD-based selection is more efficient at identifying YSOs of various types. This can be observed in the top CAMD in Fig. 10 as the cluster of H α-excess candidates found to the right of the MS track, and centred at around $M_G \sim 7.5$ mag. These systems would be difficult to identify with a positional-based partition, unless they display strong H α emission.

The position of the sources in the CAMD constitutes an indication of the stellar population they most likely belong to. A cross-match between our outliers sample and the SIMBAD data base (Wenger et al. 2000) provides a further statistical representation of the

Table 1. The table shows the first 10 rows of our *meta-catalogue*.

SourceID (*Gaia DR2*)	RA (°)	Dec. (°)	Distance (kpc)	r (mag)	e_r (mag)	i (mag)	e_i (mag)	H α (mag)	e_H α (mag)	G_{BP} (mag)	G_{RP} (mag)	M_G (mag)	Flag CAMD	Flag POS
4299502137354955552	0.011 55	62.530 21	4.911 91	13.217	0.001	12.919	0.002	12.963	0.001	13.418	12.938	–0.196	0	1
4300490967573110848	0.040 97	62.806 30	0.245 88	17.815	0.017	17.868	0.026	17.663	0.018	17.612	17.808	10.816	0	1
4323448083134383336	0.042 68	66.348 93	0.725 53	18.881	0.016	17.163	0.014	17.851	0.019	20.173	16.946	9.068	1	1
4321671684663124448	0.050 05	65.177 39	0.801 85	18.114	0.009	16.901	0.012	17.378	0.013	19.016	16.783	8.374	0	1
4299501793757661144	0.069 64	62.530 50	0.442 58	13.017	0.001	12.419	0.001	12.593	0.001	13.477	12.461	4.819	2	1
4299526876366017728	0.093 28	62.647 48	0.326 89	13.639	0.001	13.100	0.002	13.276	0.002	14.408	13.114	6.252	0	1
4295216060511237712	0.119 78	61.693 65	2.611 61	16.222	0.003	15.723	0.006	15.738	0.006	16.619	15.392	4.000	1	2
4321686502392296696	0.121 83	65.206 98	0.500 34	13.080	0.001	12.362	0.001	12.617	0.001	13.615	12.402	4.591	2	0
4299152321417676768	0.124 47	62.172 43	1.189 93	17.926	0.010	16.656	0.011	17.262	0.015	18.942	16.643	7.368	1	0
4300487231073838808	0.129 39	62.771 23	1.201 74	15.160	0.004	14.507	0.003	14.679	0.003	15.498	14.192	4.520	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Notes. For each source, the following entries are provided: the *Gaia* DR2 SourceID; the *Gaia* DR2 barycentric equatorial coordinates at epoch 2015.5; the distance (calculated in Scaringi et al. 2018); the IPHAS DR2 photometric measurements with corresponding errors; the *Gaia* DR2 photometric measurements; and the two labels *flagCAMD* and *flagPOS*. These two latter parameters express how likely each source is to be an outlier, within its corresponding $r - \mathrm{H}\alpha$ distribution (either in the CAMD-partitions and/or in the positional-partitions).

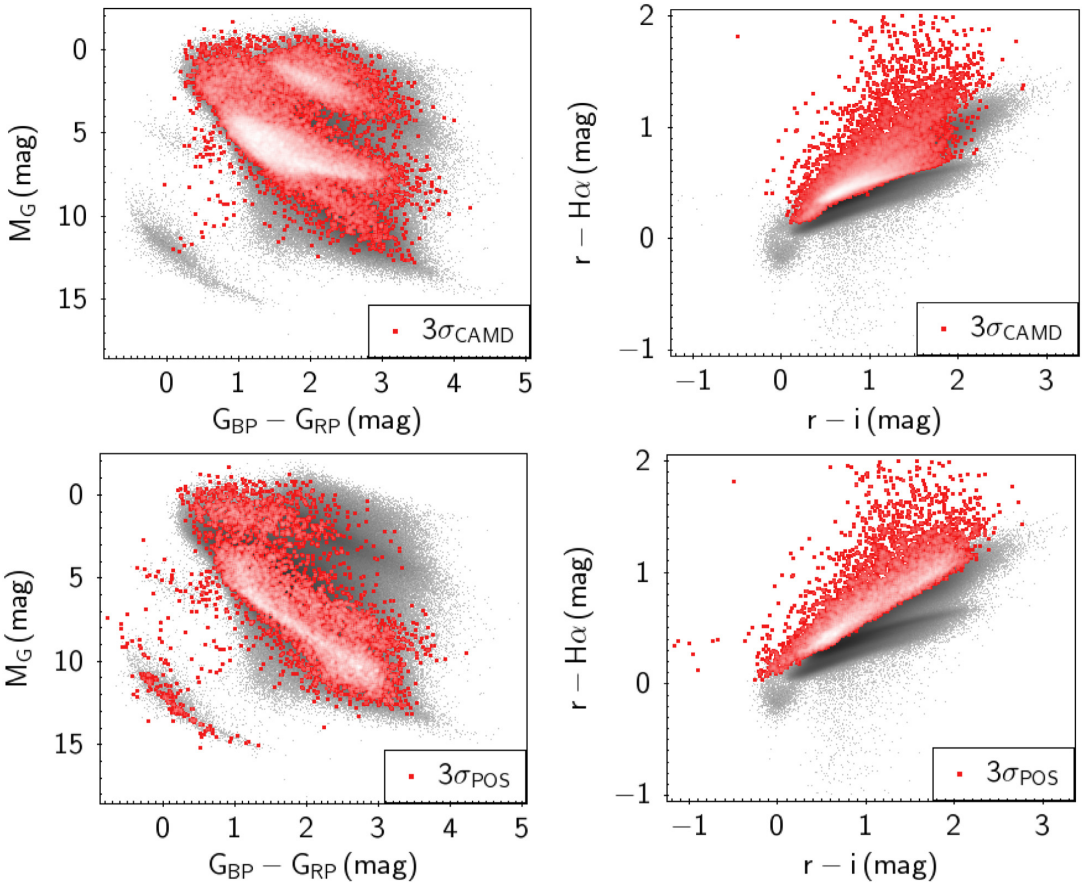


Figure 10. The top row presents the $r - \mathrm{H}\alpha$ 3σ outliers (the red dots) identified from CAMD-partitions, while the ones obtained from positional-partitions are shown in the bottom row.

populations our Hα-excess candidates belong to. A cross-matching radius of 1 arcsec yields 1825 matches. Of them, 822 are classified as YSOs (or candidate YSOs) or T-Tauri stars, 376 sources are classified with the generic epithet of 'Star', 233 are emission-line stars, 113 are classified as Be stars (or candidates), 44 as Orion Variable stars, 13 as WDs (plus 43 WD candidates), and 8 are known CVs (or candidates). The WDs included in our list of outliers and in SIMBAD are further classified, according to their spectral type: six of them are DB white dwarfs, four are DA type, two DC type, one DAB type, and one DBA type. The non-DA type WDs are identified as Hα-excess candidates by our algorithm because they do not present the strong absorption lines, typical of DA type WDs.

The left-hand column in Fig. 11 shows the r magnitude distributions of our 5σ outliers. The bin size for these histograms is 0.2 mag. As can be noticed, both the distributions are bimodal; the peak around the 13th magnitude for the CAMD-outliers, as well as the one around $r = 13.5$ mag for the positional-outliers, is to be partially imputed to an observational bias. The secondary mode for the $5\sigma_{\mathrm{CAMD}}$ outliers is 16.90 mag, and it is very close to the mode of the r intensity of the whole data set (which is $r \sim 16.60$ mag). On the other hand, the most frequent r intensity for the $5\sigma_{\mathrm{POS}}$ outliers is 18.15 mag (i.e. more than 1.5 magnitude fainter than the mode of the whole data set), confirming the fact that, generally speaking, our positional-selection is more efficient in identifying fainter outliers than the CAMD-selection. The blue areas in the histograms represent the most populated bins around $r = 13$ mag (for the CAMD-outliers) and $r = 13.5$ mag (for the positional-outliers), while the red areas indicate the three most populated bins around the respective secondary modes. In the right-hand column of the same figure, the CAMD densities of the sources belonging to these coloured regions are shown. According to their positions in the CAMD, the brightest outliers in both the distributions are active B or A types stars. The CAMD-outliers belonging to the red bins mainly cluster in the region of the CAMD associated with YSOs. Also the positional-outliers with an r magnitude close to the secondary mode mainly occupy the region of the CAMD associated with YSOs; however, some of them lie on the WD track, some in between the WD track and the MS (making them good CV candidates), and some can be found on the M-dwarfs region.

In the following subsections, our results are compared with previous similar studies. More specifically, they are cross-matched with the catalogues developed by Witham et al. (2008) and by Monguió et al. (2020) (IGAPS). Moreover, a further validation of our selection, based on the visual inspection of LAMOST DR5 spectra (Yao et al. 2019) is presented. Since accreting compact objects often show X-ray emission, a fraction of our Hα-excess candidates are expected to be found in X-ray surveys as well. Therefore, in the last subsection, the quantitative results of the cross-matches with three X-ray surveys are briefly discussed. These surveys are: the *ROSAT* All-Sky Survey Faint Source Catalogue ('faint-*ROSAT*' hereafter; Voges et al. 2000), the *ROSAT* All-Sky Survey Bright Source Catalogue ('bright-*ROSAT*' hereafter; Voges et al. 1999), and the *Chandra* Source Catalogue ('CSC' hereafter; Evans et al. 2010).

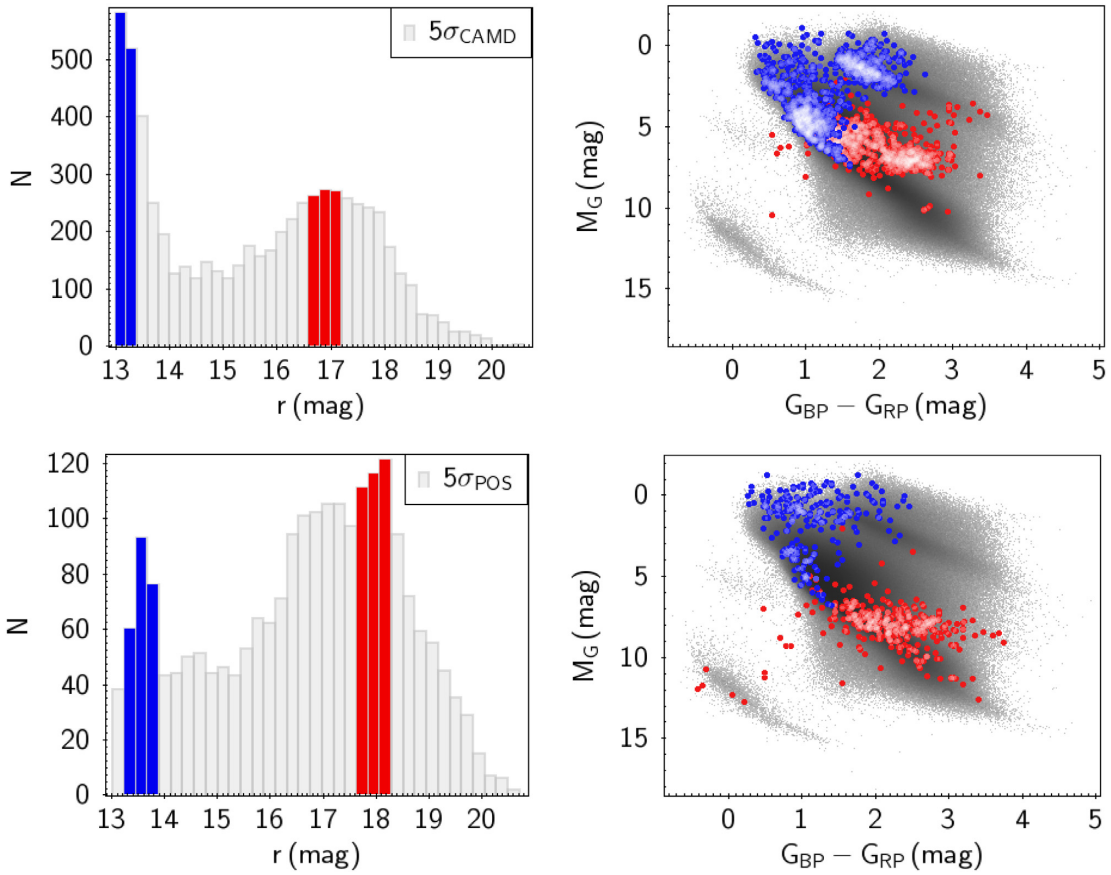


Figure 11. *r* magnitude distributions (left-hand column) of the $5\sigma_{\rm CAMD}$ outliers (top row) and of the $5\sigma_{\rm POS}$ outliers (bottom row), with a bin width of 0.2 mag. Both the distributions are bimodal; the blue areas in the histograms point to the bins around the brightest of the two modes, respectively, while the red areas in the histograms display the three most populated bins around the secondary modes. In the right-hand column, the density in the *Gaia* CAMD of the objects that occupy the areas around the modes in the corresponding histogram are shown.

5.1 Comparison with Witham's catalogue

Comparing the emitters' list in Witham et al. (2008) (which counts 4853 objects) with the full master-catalogue developed by Scaringi et al. (2018) (after the application of the quality cuts described in Section 2.1), 1213 common sources (25.0 per cent of Witham et al. 2008 outliers sample) are found. The cross-match is performed by using a radius of 1 arcsec; however, the number of matches does not change significantly if this parameter is increased up to a generous 5 arcsec. Although the remaining 75 per cent of Witham's outliers can be found in *Gaia* DR2 archive, their astrometric/photometric measurements did not satisfy the quality constraints applied by Scaringi et al. (2018) when producing the *Gaia*/IPHAS catalogue.

Out of the 1213 common targets, 1115 (91.9 per cent) are identified as outliers by our algorithm as well, with a significance (either $\sigma_{\rm CAMD}$ and/or $\sigma_{\rm POS}$) equal to or higher than 3. By comparing the common sources with our CAMD-based outliers, 1053 of common outliers (94.4 per cent) are recovered, whilst the positional-based selection finds 1054 (94.5 per cent) of them (i.e. 992 common outliers are identified by both our selection criteria). A subset of 933 out of 1115 common objects is characterized by $\sigma_{\rm CAMD} \geq 5$ and/or $\sigma_{\rm POS} \geq 5$; 893 of them have a $\sigma_{\rm CAMD} \geq 5$, while 671 of them have $\sigma_{\rm POS} \geq 5$ (hence 631 Witham's emitters are found by both our selection criteria, with a significance equal to or greater than 5). In the two panels of Fig. 12, the positions in the *Gaia* CAMD of the objects resulting from the cross-match with Witham's catalogue are presented. More specifically, the top panel shows the matches between Witham's list and our 3σ CAMD-outliers, while the bottom panel shows an analogous diagram for our positional-outliers. As previously stated in this work, the CAMD-based selection is more efficient in recovering bright objects, such as the ones that lie on the active, reddened, B or A type stars are tracked. On the other hand, the cross-match with our positional-based selection yields more matches in the M-dwarf region of the CAMD, and on the WD track.

By checking the differences in the photometric measurements in Witham et al. (2008) (minus the uncertainties) and in IPHAS DR2 (plus the uncertainties), some variable objects can be spotted. However, we point out that this apparent variability might be due to the different calibrations.[5] Of the 98 3σ objects missing in our list of outliers, 42 showed a stronger $r - {\rm H}\alpha$ emission at the epoch of Witham's study, if compared to IPHAS DR2. This decrease in the $r - {\rm H}\alpha$ intensity might be the reason why those particular sources were identified as H α emitting candidates in Witham et al. (2008), but not by the methods implemented in our study. Fig. 13 shows this apparent $r - {\rm H}\alpha$ intensity drop.

[5]As previously mentioned, the study of Witham et al. (2008) was performed on IPHAS pre-publication measurements, while our selection algorithm is applied to IPHAS DR2 calibrated data.

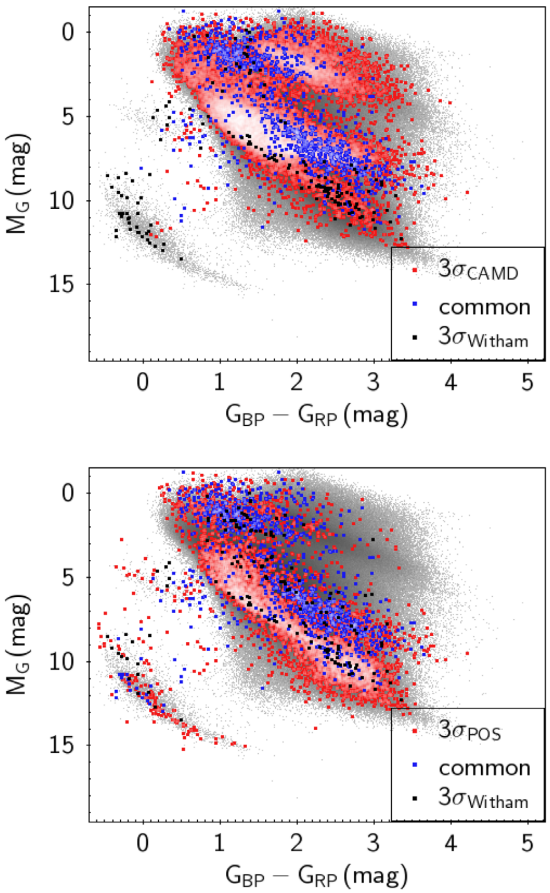


Figure 12. Top panel: position in the CAMD of the matches between the outliers in Witham et al. (2008) and our 3σ CAMD-outliers. Bottom panel: position in the CAMD of the matches between the outliers in Witham et al. (2008) and our positional-outliers (bottom panel). The red dots represent the totality of our CAMD/positional outliers; the blue dots are the common outliers between Witham's list and our CAMD/positional selection; the black dots are Witham's outliers not selected by our CAMD/positional algorithm.

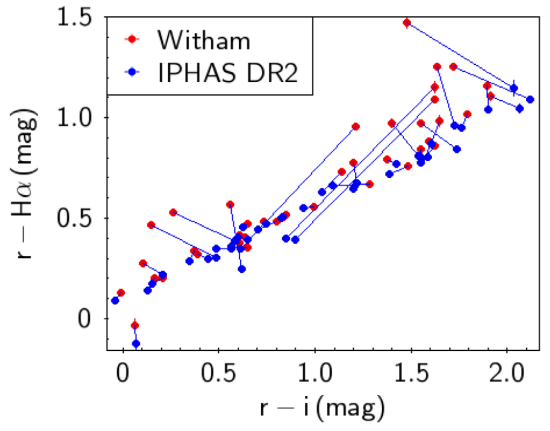


Figure 13. Positions in the TCD of 42 sources that were identified as outliers in Witham et al. (2008), but not my our algorithm. These are the objects the $r - \mathrm{H}\alpha$ intensity of which was higher in Witham's catalogue (the red dots) with respect to the analogous IPHAS DR2 intensity (the blue dots). For most of these targets, the error bars included in the plot are too small to be visible.

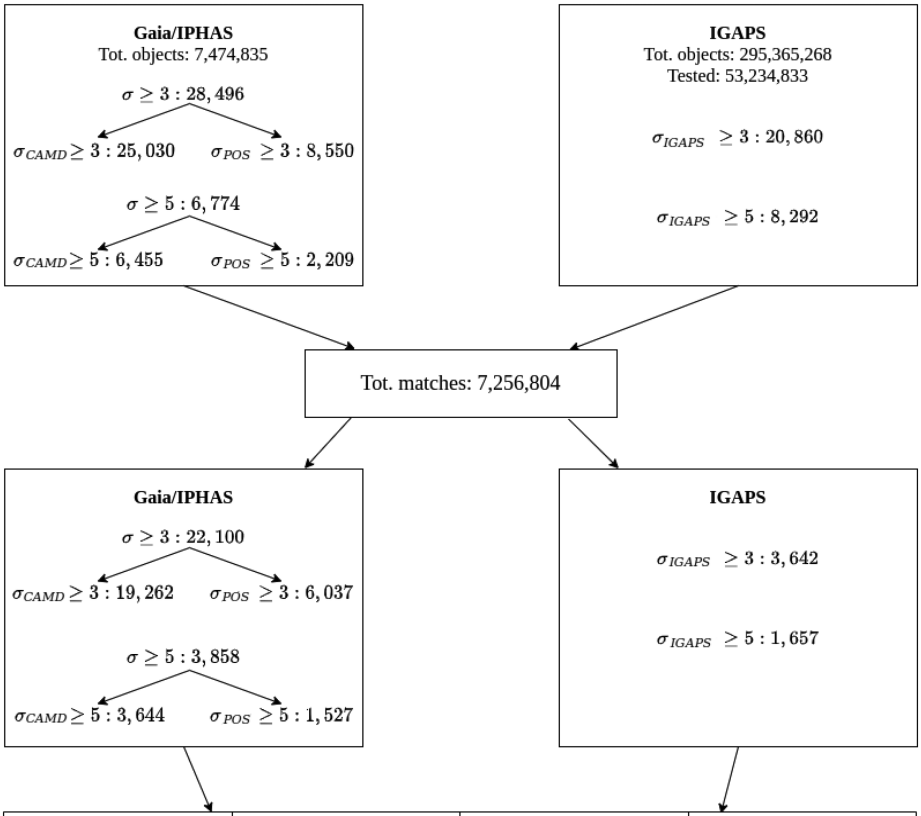
5.2 Comparison with IGAPS

The identification of emission-line objects performed by Monguió et al. (2020) followed a selection strategy that is similar to the one implemented for Witham et al. (2008). The main differences between these two works are related to the data calibration and to the morphology classes being tested (Monguió et al. 2020 only excluded 'morphology class 0' sources, i.e. the 'noise-like sources', from being tested for H α excess; see also Farnhill et al. 2016). Of the 53 234 833 objects in the IGAPS catalogue that were tested for H α excess, Monguió et al. (2020) produced a list of 20 860 excess-line candidates (0.04 per cent of the tested targets). These outliers were selected with a significance higher than 3; a sub-sample of these excess candidates is composed by 8292 objects (0.02 per cent of the tested sample) with significance higher than 5.

A cross-match between the *Gaia*/IPHAS catalogue (after the application of the quality cuts described in Section 2.1) and the ∼53 million IGAPS tested sources yields 7256 804 matches. The cross-matching radius is 1 arcsec. This subset includes 3642 IGAPS outliers, 1657 of which with an associated significance higher than 5. It also includes a subset of 22 100 of our 3σ outliers: 19 262 of them are derived from the CAMD-based selection, and 6037 from the positional-partitions. We point out that our CAMD-based partitioning algorithm is performed on a different parameter space with respect to the one applied in Monguió et al. (2020). Nonetheless, for a more complete discussion, all the results of the possible cross-matches between the two lists are provided. The cross-matching process between these two catalogues and its results are presented in the flow chart in Fig. 14.

The positions in the *Gaia* CAMD and IPHAS TCD of the 843 IGAPS outlier not identified by our algorithm are shown in the top row of Fig. 15 (the red and blue dots). IPHAS DR2 photometric measurements are used to produce the TCD. As can be noticed from the CAMD, the vast majority of these objects can be associated with the M-dwarf region of the CAMD. As previously mentioned, these objects are characterized by significantly different IPHAS colours, with respect to the other MS stars. This appears to fail our selection through the use of CAMD-based partitions. In the bottom row of the same figure, our 3417 positional-outliers that were not identified by Monguió et al. (2020) are placed in the CAMD and TCD. These objects mainly lie in the CAMD on the MS track, on the reddened B and A types stars track, or on the WD track. However, some red dots are placed in between these two tracks, making them good CV candidates.

The mismatches between the results obtained in Monguió et al. (2020) and by us are to be ascribed mainly to two factors: the different definitions used to calculate σ and the different calibrations applied to the photometric measurements in the input data bases. In fact, when producing the *Gaia*/IPHAS catalogue, Scaringi et al. (2018) based their work on IPHAS DR2 calibrated data, while IGAPS calibration (as mentioned in Section 1) relies on the more recent 'Pan-STARRS reference ladder' (Magnier et al. 2013). This latter effect is visible in Fig. 16, where the $\delta\,(r - \mathrm{H}\alpha)$ versus $\delta\,(r - i)$ diagram is presented. The two axes represent the difference between IGAPS and IPHAS DR2 values for $r - \mathrm{H}\alpha$ and $r - i$ parameters, respectively. The grey dots correspond to all the matches between IGAPS and the *Gaia*/IPHAS catalogue, while the blue dots are our positional-outliers that Monguió et al. (2020) did not identify as H α-excess candidates. The $\delta\,(r - i)$ mode for the grey dots is −0.05 mag, while the most common value for the blue dots is -0.06 mag. The mode of the $\delta\,(r - \mathrm{H}\alpha)$ distribution for both the grey and blue points is −0.03 mag. If IGAPS and IPHAS DR2 data had the same



	$\sigma_{IGAPS} < 3$	$3 \leq \sigma_{IGAPS} < 5$	$\sigma_{IGAPS} \geq 5$
$\sigma_{CAMD} < 3$ & $\sigma_{POS} < 3$	7,233,861	807	36
$3 \leq \sigma_{CAMD} < 5$ or $3 \leq \sigma_{POS} < 5$	17,327	760	155
$\sigma_{CAMD} \geq 5$ or $\sigma_{POS} \geq 5$	1,947	418	1,466
$\sigma_{POS} < 3$	7,249,745	946	76
$3 \leq \sigma_{POS} < 5$	3,142	940	428
$\sigma_{POS} \geq 5$	275	99	1,153

Figure 14. The flow-chart describes the cross-matching process between the *Gaia*/IPHAS catalogue and the IGAPS catalogue, as well as its detailed results.

calibration, the points in this diagram would cluster around the (0,0) coordinates. However, almost all our positional-outliers not listed in IGAPS lie below the $\delta\,(r - \mathrm{H}\,\alpha) = 0$ line; this supports our hypothesis that the different calibration is one of the main factors that cause the discrepancy between our positional-selection and IGAPS selection.

5.3 LAMOST spectra[6]

A more direct validation of our selection comes from visually inspecting the spectra of the photometrically identified H α-excess candidates. In order to achieve this validation, a cross-match between our list of outliers with LAMOST DR5 is performed. However, LAMOST DR5 spectra and IPHAS DR2 measurements were acquired at different epochs (IPHAS DR2 observations were implemented between 2003 and 2012, while LAMOST DR5 spectra were collected between 2016 and 2017). Therefore, some transient H α-excess sources selected by our algorithm may not display clear H α emission, and vice-versa.

5.3.1 Purity and completeness

A cross-match with the LAMOST DR5 archive and our H α-excess candidates list (with a 0.5-arcsec cross-matching radius) yields 1873

[6]In the current section, we refer to the sources using their LAMOST DR5 designations, as well.

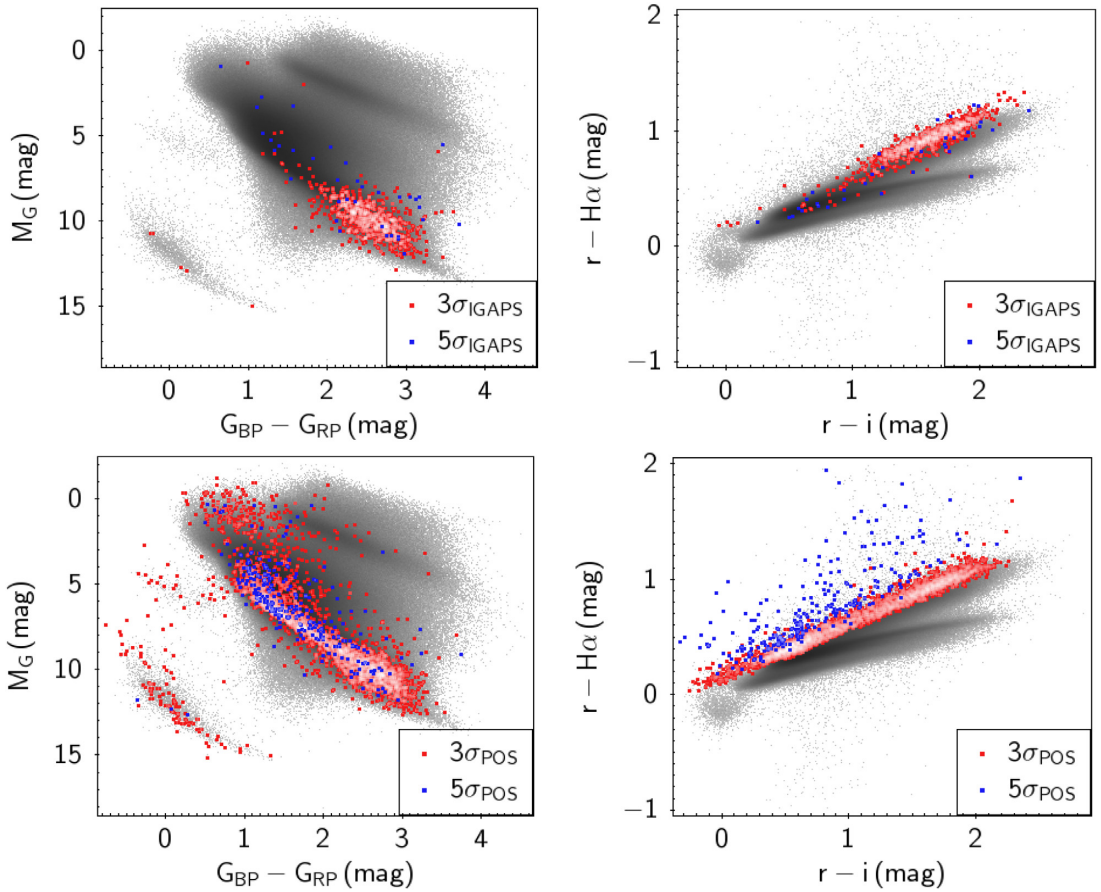


Figure 15. Top row: location in the *Gaia* CAMD (left-hand panel) and IPHAS TCD (right-hand panel) of the 843 IGAPS $r - \mathrm{H}\alpha$ outliers that are not selected by our algorithm. The red dots represent objects that Monguió et al. (2020) identified with a significance included between 3 and 5, while the objects with a higher significance are depicted with blue dots. Bottom row: position in the CAMD (left-hand panel) and TCD (right-hand panel) of our 3417 positional-outliers that are not listed as Hα-excess candidates in IGAPS. The red dots represent the positional-outliers with a significance included between 3 and 5, while the blue dots represent outliers with a higher positional-significance.

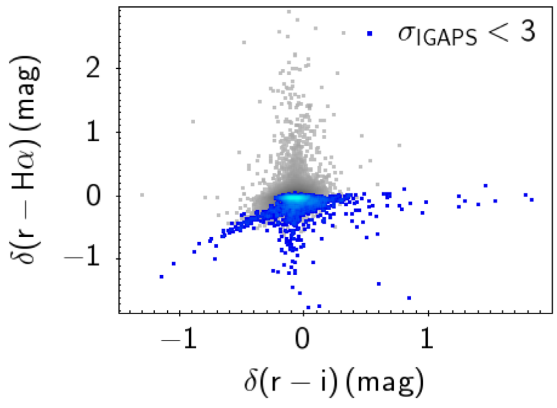


Figure 16. Position in the $\delta(r - \mathrm{H}\alpha)$ versus $\delta(r - i)$ diagram of all the matches between IGAPS and the *Gaia*/IPHAS catalogue (the grey dots), and of our positional-outliers that Monguió et al. (2020) did not select as Hα-excess candidates (the overplotted blue dots).

spectra. These spectra are used to calculate the *purity* of our selection, with the assumption that they constitute a good representation of our 3σ outliers. Of these 1873 objects, 916 (48.9 per cent) are confirmed as reliable H α-excess candidates, while 939 (50.1 per cent) seem to show H α absorption. The remaining 18 spectra do not allow a univocal assessment, due to their low quality. We point out that these relatively low spectral confirmation rates constitute a lower limit for the purity of our selection, since our algorithm does not aim to identify Hα *emitters*, but rather Hα*-excess* sources. Therefore, objects that exhibit excess Hα flux (but not necessarily displaying an Hα emission line) relative to the underlying partition are selected as outliers. This also explains the higher spectral confirmation rate for the positional-outliers, with respect to the CAMD-outliers. Fig. 17 displays two examples of 5σ outliers the LAMOST spectra of which show absorption in the Hα band. These are compared to the spectra of two other objects in the same partitions with an associated significance lower than 3. The ratios between the red and blue fluxes in the top panels, zoomed in around the Hα wavelength, are presented in the bottom right-hand panels. In correspondence with the Hα wavelength, both these ratios are significantly above the mean, which explains the high significance associated with these sources. Hα excess is often accompanied by Hβ excess, as can be seen in the bottom left-hand panels.

Purity does not change significantly, if a more conservative cut on the selection of the outliers is considered: out of 616 spectra relative to sources with either $\sigma_{\mathrm{CAMD}} \geq 5$ and/or $\sigma_{\mathrm{POS}} \geq 5306$ (49.7 per cent) seem to be solid H α-excess candidates. On the one hand, out of the 603 5σ CAMD-outliers for which LAMOST spectra are

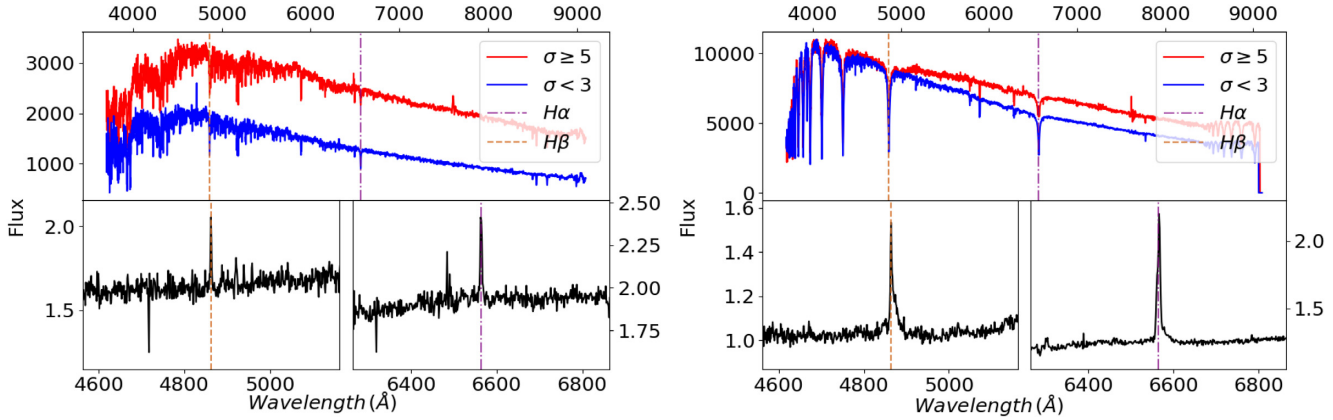


Figure 17. Example of two 5σ *excess* sources that show Hα absorption (the red lines in the top panels). The blue lines represent the fluxes of two objects in the same partitions, with a significance lower than 3. The bottom panels display the ratios between the the fluxes in the top panels, centred around the Hβ wavelength (bottom left-hand panels) and around the Hα wavelength (bottom right-hand panels).

available, 294 (48.8 per cent) show H α-line emission. On the other hand, 128 spectra out of 157 (81.5 per cent) seem to validate our 5σ positional-based selection. By applying a more rigid cut on IPHAS magnitudes, and hence reducing the effects due to saturation, the ratio of spectroscopically confirmed outliers improves significantly. In fact, retaining the sources with $r \geq 13.5$ mag, $i \geq 12.5$ mag, and H$\alpha \geq 13$ mag, 772 spectra out of our 1141 (67.7 per cent) confirm our 3σ outlier selection (either CAMD-based and/or positional-based). Constraining the spectral analysis to our 5σ outliers, 231 LAMOST spectra out of 282 (81.9 per cent) confirm our selection. Thus, these additional quality-cuts are suggested to the users of our meta-catalogue.

The *completeness* parameter (C) relative to our selection (i.e. the ratio between the number of spectroscopically confirmed Hα emitters identified by our algorithm and the total amount of spectroscopically confirmed Hα emitters within our full master-catalogue) is obtained with two different methods:

(i) The first method consists of the evaluation of

$$C = \frac{N_{\sigma\geq 3} \times P}{N_{\sigma\geq 3} \times P + N_{\sigma<3} \times fn}. \quad (2)$$

Here, $N_{\sigma \geq / < 3}$ is the amount of objects with a significance higher/lower than 3, P is the purity fraction relative to the full 3σ outliers sample (48.9 per cent), and fn is the *false-negative* fraction (5.6 per cent). This latter parameter derives from the visual inspection of 1000 spectra belonging to randomly selected sources with $\sigma < 3$, and corresponds to the fraction of these spectra that show Hα emission. This method yields a completeness of around 3 per cent. The positions of the 56 false-negative objects in the CAMD and TCD are shown in Fig. 18 (top and bottom panels, respectively). Most of these sources lie in the M-dwarfs region of the CAMD.

(ii) the second method consists of the visual inspection of 2000 LAMOST spectra belonging to randomly selected objects in our catalogue, 15 of which are identified as Hα 3σ outliers by our selection. The completeness thus obtained is approximately 5 per cent.

Such low completeness values are partially due to the combination of (a) the different epochs between LAMOST DR5 spectra and *Gaia*/IPHAS measurements (and hence the variability of some objects), (b) too conservative thresholds during the Hα outliers selection processes, and (c) too generous definition of 'Hα emitters' during the visual inspection of the spectra. Calibration-related problems are ruled out by the fact that none of our false-negative

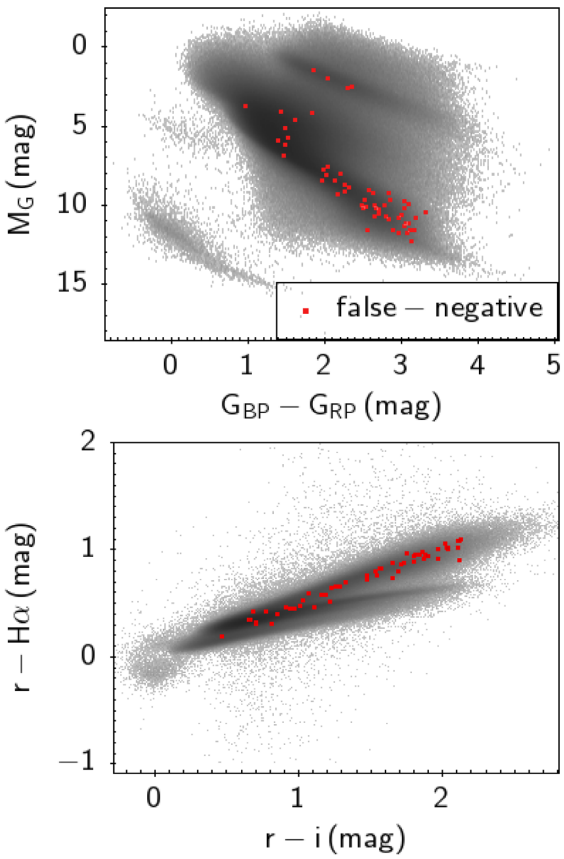


Figure 18. Position in *Gaia* CAMD (top panel) and in IPHAS TCD (bottom panel) of 56 'false-negative' objects.

objects are included in IGAPS list of outliers. However, an exhaustive explanation for this low completeness fractions is still to be found. As a comparison, the same calculations applied on IGAPS catalogue yield a completeness percentage below 1 per cent.

In Fig. 19, four spectra associated with our outliers are shown: two of these spectra belong to sources with an associated σ included between 3 and 5, while the other two belong to objects with a higher significance. For each of these significance-based outliers subsamples, one example of confirmed Hα-excess, and one example

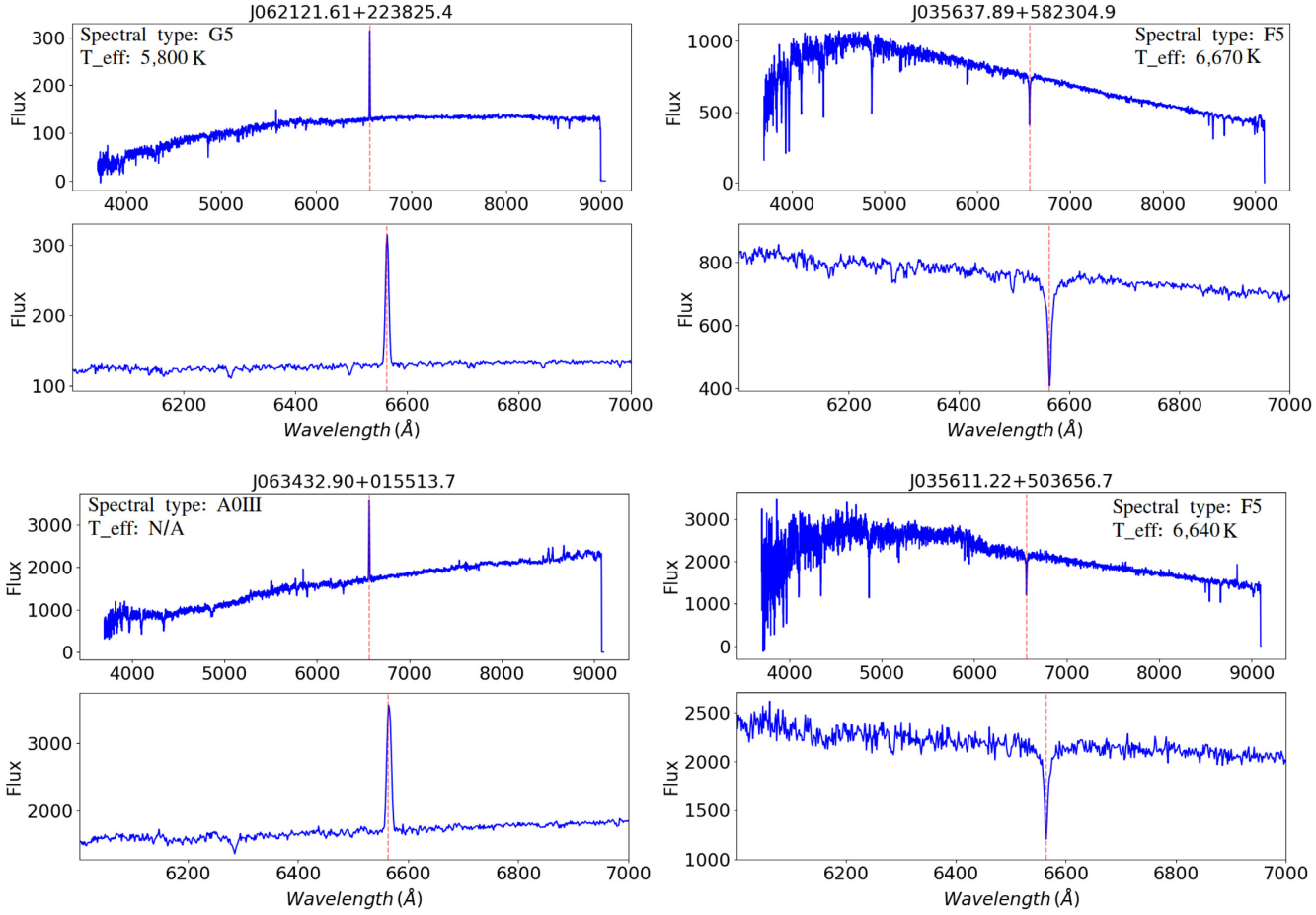


Figure 19. LAMOST spectra of four objects that are selected as H α-excess sources by our algorithm. For each of these spectra, a zoom-in on the region around the H α line is shown. The top two spectra belong to the sources J062121.61+223825.4 (*Gaia* DR2 ID 3377220715714066304) and J035637.89+582304.9 (*Gaia* DR2 ID 470024698144186112), respectively, which have an associated significance (either CAMD-based and/or positional-based) included between 3 and 5. The bottom spectra refer to the objects J063432.90+015513.7 (*Gaia* DR2 ID 3120920947508407808) and J035611.22 + 503656.7 (*Gaia* DR2 ID 250435321081819392), which are characterized by a higher significance. The red dashed line indicates the H α wavelength.

of clear absorption in the H α band are shown. These four sources are located in the *Gaia* CAMD and IPHAS TCD in the left- and right-hand panels in Fig. 20, respectively.

5.3.2 Spectral analysis of the cross-matches with Witham and IGAPS

Out of the 98 Witham's outliers that are not identified by our selection (see Section 5.1), 10 have an associated LAMOST DR5 spectrum. Four of these spectra show clear a H α emission line. Overall, 533 spectra relative to the outliers in Witham et al. (2008) are present in the LAMOST archive, and 481 of them (90.2 per cent) show a clear H α emission line.

Regarding IGAPS 3σ outliers, 543 of them have an associated LAMOST DR5 spectrum, and 491 of these spectra (90.4 per cent) show H α emission. Of the 843 IGAPS 3σ outliers that our algorithm does not identify as H α emitting candidates (see Section 5.2), 21 have an associated LAMOST DR5 spectrum. The absolute majority of these spectra (18/21) shows a clear H α emission line. On the other hand, out of the 3417 3σ positional-based outliers not included in IGAPS outliers list, 149 have an associated LAMOST spectrum. By visually inspecting these spectra, 57 of them (38.3 per cent) belong to clear H α-excess sources. If constraining the subset to our 5σ positional-outliers, 6 out of 9 available spectra show clear H α emission.

5.4 Cross-matches with faint-*ROSAT*, bright-*ROSAT*, and CSC

Accretion on to compact objects is often accompanied by X-rays emission. Table 2 provides the results of the cross-matches (with a radius of 15 arcsec) between our catalogue and three X-ray surveys: the *ROSAT* All-Sky Survey Faint Source Catalogue (faint-*ROSAT*; Voges et al. 2000), the *ROSAT* All-Sky Survey Bright Source Catalogue (bright-*ROSAT*; Voges et al. 1999), and the The Chandra Source Catalogue (CSC; Evans et al. 2010). Among the 972 matches with faint-*ROSAT*, 33 are identified as H α-excess candidates by our algorithm. Almost all of these objects find a classification in SIMBAD (32/33): 30 out of 32 are identified as 'X-ray emitting sources[7]', and the remaining two as CVs. LAMOST spectra are available for three of these targets, and they all present a clear H α

[7]As the label 'Star', the 'X-ray emitting source' generic label in SIMBAD does not provide any further specification on the object being classified.

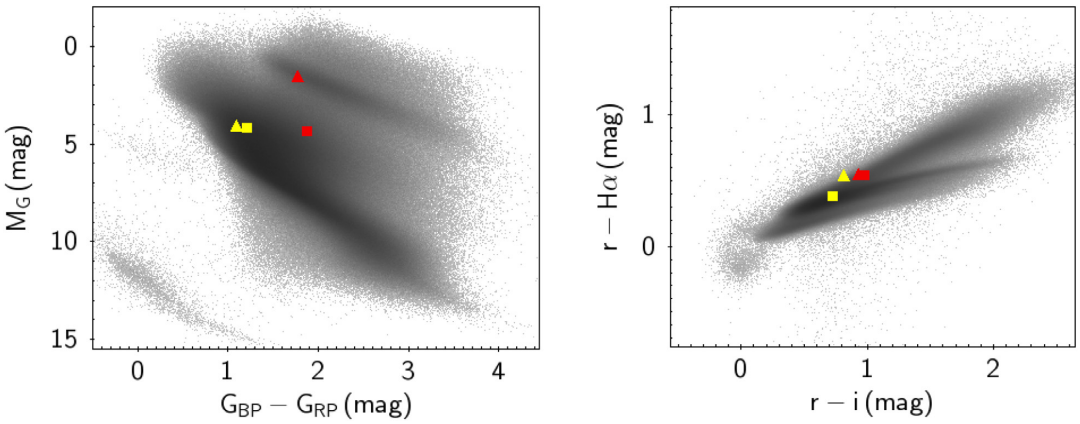


Figure 20. Positions in the *Gaia* CAMD (left-hand panel) and IPHAS TCD (right-hand panel) of the objects in Fig. 19. The squares represent the objects with a significance (either σ_{CAMD} and/or σ_{POS}) included between 3 and 5, while the triangles depict the sources with a higher significance. The colour-code is: yellow for the objects whose spectra show absorption in the H α band, and red for the objects that show H α emission.

Table 2. The table shows the results of different cross-matches between our meta-catalogue and the three X-rays surveys faint-*ROSAT* (Voges et al. 2000), bright-*ROSAT* (Voges et al. 1999), and CSC (Evans et al. 2010).

	Total	*flagCAMD* = 0 *flagPOS* = 0 (per cent)	*flagCAMD* = 1 (per cent)	*flagCAMD* = 2 (per cent)	*flagPOS* = 1 (per cent)	*flagPOS* = 2 (per cent)
Faint *ROSAT*	972	939 (96.6)	26 (2.7)	10 (1.0)	17 (1.7)	8 (0.8)
Bright *ROSAT*	69	59 (85.5)	8 (11.6)	8 (11.6)	9 (13.0)	7 (10.1)
CSC	6667	6,241 (93.6)	177 (2.7)	218 (3.3)	71 (1.1)	72 (1.1)

Notes. The objects in our data set are grouped before the cross-matches, with reference to the corresponding flagCAMD and flagPOS specifications. These entries refer to the significances (either CAMD-based or positional-based) associated with each object in the catalogue.

emission line. Our algorithm assigns a significance (either CAMD-based and/or positional-based) higher than 5–12 of these 32 objects, including the two CVs.

Ten 3σ outliers are included in the 69 matches with bright-*ROSAT*, and all of them find a classification in SIMBAD. Three of them are classified as 'X-ray emitting source', three as CVs, two as Dwarf Novae, one WD, and one T-Tauri star. Three LAMOST spectra are available for this group of sources, as well; they all show the H α line in emission; these spectra belong to the identified WD (of spectral type DA), the CV and the Dwarf Nova. Our algorithm associates a significance higher than 5 to eight of these 10 objects; only the WD and one of the source classified as 'X-ray emitting source' have a lower significance.

The cross-match between our catalogue and CSC yields 6,667 matches, 426 of which are identified as outliers by our algorithm. Out of these 426 objects, 342 are classified in SIMBAD. Among them, 264 are YSOs (or candidates) and T-Tauri stars, 29 are classified as 'Stars', 27 as emission-line stars, and 15 as Orion Variables. Of these 342 objects, 7 find a LAMOST spectrum, and the H α emission line is visible in all of them. Of these 342 targets, 206 have a significance higher than 5, and the vast majority of them (171/206) are classified in SIMBAD as YSOs.

In Fig. 21, these three groups of 342, 32, and 10 sources are located in the *Gaia* CAMD (left-hand panel) and IPHAS TCD (right-hand panel). In agreement with their SIMBAD classification, most of the matches with CSC cluster around the area of the CAMD in which young accreting objects are expected to lie. Some of the sources in the two *ROSAT* surveys are located between the MS track and the WD track, making them robust CV candidates. These are either already identified as such in SIMBAD, or are classified with the general label of 'X-Ray emitting sources'.

6 ENABLED SCIENCE CASES

Two examples of possible science cases for our meta-catalogue are presented here. The first one consists of the identification of previously undetected accreting WD candidates, and it hinges directly on the cross-match between our outliers and the three X-Ray surveys presented in the previous section. In fact, accreting WDs usually occupy a well-known region in the CAMD (between the MS and the WD tracks), and are associated with H α and X-Ray emission. As an example, Lan 23 (Wramdemark 1981; Skrutskie et al. 2006) is a well known WD that is identified as an H α outlier by our algorithm, its LAMOST spectrum shows an H α emission line, and is found in the bright-*ROSAT* catalogue.

Fig. 22 shows the position in *Gaia* CAMD of all our 3σ outliers that also present X-Ray emission. Among the X-Ray emitters that are not yet classified in SIMBAD, the black dots represent good examples of new robust accreting WD candidates. These objects are *Gaia* DR2 4140717539973182 72, *Gaia* DR2 2060626872274773504, *Gaia* DR2 4633503189632106 24, *Gaia* DR2 2203244624288543744, and *Gaia* DR2 2203373473312011008.

Another feature that characterizes accreting WDs (as well as many other stellar populations) is variability. Abrahams et al. (2020) developed a method that enables the calculation of a parameter (ϵ) that quantifies the excess of Poissonian noise relative to the flux of a source. With the use of *Gaia* metrics, this parameter is given by

$$\epsilon = \sqrt{N} \times \frac{\delta f_G}{f_G}. \quad (3)$$

Here, f_G is the mean *G*-band flux obtained with *N* observations, while δf_G represents the corresponding dispersion. The sources in our catalogue are binned with respect to their *G*-band magnitude; the ones with an ϵ larger than five standard deviations above $\epsilon_{mean,i}$ (i.e.

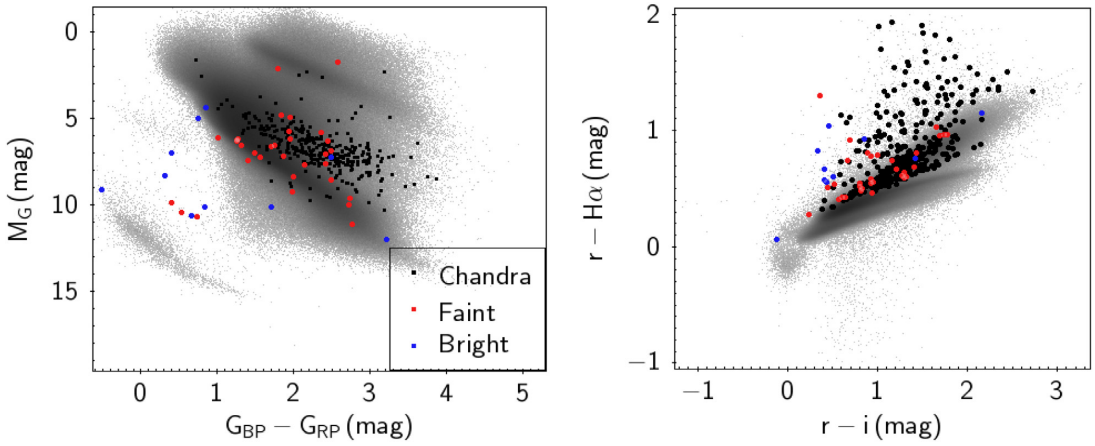


Figure 21. The left-hand panel presents the positions in the *Gaia* CAMD of our $r - \mathrm{H}\alpha$ 3σ outliers that are included in the surveys bright-*ROSAT* (Voges et al. 1999, the blue points), faint-*ROSAT* (Voges et al. 2000, the red dots), and CSC (Evans et al. 2010, the black dots), for which a SIMBAD classification is available. The right-hand panel shows the positions of the same objects in the IPHAS TCD.

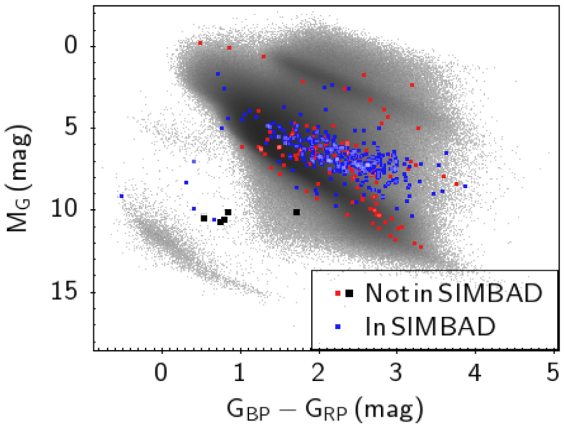


Figure 22. Position in *Gaia* CAMD of our 3σ Hα outliers that show X-Ray emission. The blue dots represent the objects that are already classified in SIMBAD, while the red and black ones are unclassified (or simply classified as 'star' or 'X-Ray emitting source').

the average ϵ value for the ith bin) are selected as variables. With this method, 22 199 variable sources are identified. According to our Hα-excess selection, 2243 of them are also Hα-excess sources. The top panel of Fig. 23 shows the ϵ versus G-band magnitude diagram: The blue dots represent the variable objects, whilst the red ones represent the variable Hα-excess sources. The bottom panel in the same figure shows the location of the variable Hα-excess sources in *Gaia* CAMD. While most of them cluster in the region associated with YSOs, many red dots are found in the region of the CAMD where B or A type stars lie, and between the MS and the WD tracks.

A combination of X-Ray emission and variability makes our accreting WD candidates identification more robust. The bottom panel in Fig. 23 presents the position in the CAMD of the 41 variable objects that show X-Ray emission and are not yet classified in SIMBAD. The five black objects in Fig. 22 are included among them.

7 CONCLUSIONS

In this study, a new method for selecting Hα-excess candidates from a vast photometric survey is presented. Our analysis is performed on the *Gaia*/IPHAS catalogue, produced by Scaringi et al. (2018). It comprises targets included in the $|b| \leq 5^\circ$ and $29^\circ \leq l \leq 215^\circ$ ranges, within a radius of ∼1.5 kpc. *Gaia* photometric measurements and parallaxes play a key role in the development of our selection: by locating the sources in the *Gaia* CAMD, it is possible to associate them to a stellar population. In order to minimize the effects due to stellar population mixing, the targets are partitioned in the *Gaia* CAMD; to mitigate the effects of extinction, they are further (and independently) partitioned in the Galactic coordinate space. For each partition, the main locus in the IPHAS TCD, and subsequently the $r - \mathrm{H}\alpha$ outliers, are found by applying the iterative Chauvenet's criterion twice. The Hα-excess candidates are thus defined as the sources that satisfy the criterion in equation (1).

This process leads to the identification a new set of Hα- excess candidates in the Northern Galactic plane. In fact, the partition of the sources in two different parameter spaces enables the identification of Hα line candidates that would be otherwise hidden among different stellar populations. More specifically, 28 496 Hα-excess candidates (0.4 per cent of the total data set) are identified, with either $\sigma_{\mathrm{CAMD}} \geq 3$ and/or $\sigma_{\mathrm{POS}} \geq 3$. However, a 5σ cut is suggested, as it constitutes a solid agreement between completeness and conservativeness. By applying this latter cut, 6774 objects (23.8 per cent of the 28 500 3σ outliers) are identified as H α-excess sources. The visual inspection of the available LAMOST DR5 spectra of our 3σ outliers shows that 48.9 per cent of them exhibit a clear H α emission line. This purity fraction does not improve significantly if constraining the outliers to the 5σ subset: 49.7 per cent of them are confirmed H α emitters by the available LAMOST spectra. These apparently low percentages are explained by the fact that our algorithm identifies Hα-*excess* sources, rather than Hα *emitters*. This is also consistent with the spectral confirmation rate being systematically higher for our positional-outliers than for our CAMD-outliers. However, by retaining only the outliers that are at least half magnitude fainter than IPHAS saturation limits, 67.7 per cent of our 3σ outliers – and 81.9 per cent of our 5σ outliers – are spectroscopically confirmed as reliable Hα-excess sources. This latter selection cuts are therefore suggested.

Our 3σ selection identifies between 3 and 5 per cent of the H α emitters in the Northern Galactic plane. Despite this constituting an improvement with respect to previous similar studies, it also suggests that our knowledge of the Hα emitters in the Galaxy is still far from being complete.

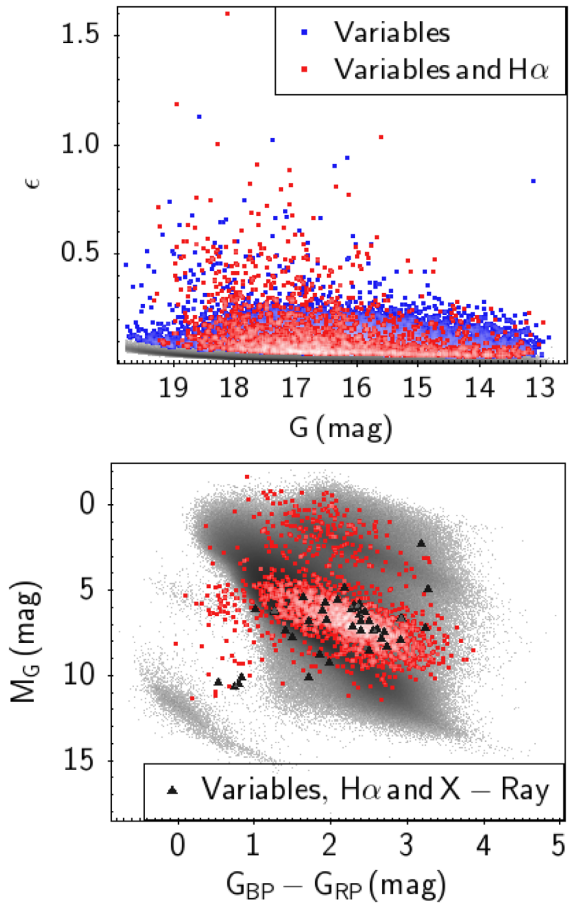


Figure 23. Top panel: variable objects (the blue dots) and variable H α outliers (the red dots) in our meta-catalogue. Bottom panel: position in the CAMD of our variable H α-excess sources (the red dots). The black triangles represent the variable H α outliers that also show X-Ray emission, and are not classified in SIMBAD.

The results of our analysis are presented in our *meta-catalogue* of the *Gaia*/IPHAS H α-excess sources. This includes all the 7474 835 objects in the master-catalogue, and for each of them, the following specifications are provided: the *Gaia* DR2 SourceID, the equatorial coordinates, the distance, IPHAS DR2 and *Gaia* DR2 photometric measurements, as well as the two *flagCAMD* and *flagPOS* labels that specify the confidence level that the source is an H α-excess candidate. Moreover, a full-version of our catalogue is available, in which the whole set of metrics obtained during the H α-excess selection process is added.

A cross-match with SIMBAD (Wenger et al. 2000) shows that 6.4 per cent of our 3σ outliers have been previously identified. However, if followed up spectroscopically, our list of outliers can be used to enhance the census of identified H α emitting point-like sources, such as CVs or SySts. This constitutes a profitable starting point to address, for instance, the problem of the difference between observed and predicted CVs spatial density in the Galactic plane (de Kool 1992; Kolb 1993). Although Belloni et al. (2020) seem to have found a promising way to overcome this impasse, their conclusions are still to be confirmed (Pala et al. 2020). Moreover, the identification of new H α emitting sources can foster population studies, which, by definition, need a vast amount of objects to be performed. In addition, newly classified sources can provide further pieces of the puzzle for a better understanding about the possible evolutionary models of the stellar population they belong to.

With the arrival of *Gaia* early Data Release 3 (*Gaia* eDR3), our intention is to apply our analysis on the list of objects resulting from the cross-match between *Gaia* eDR3 and IGAPS. This will provide more up to date results, compared to the ones in our current meta-catalogue.

ACKNOWLEDGEMENTS

The cross-matches, as well as some of the figures used for this paper, were produced with the use of the astronomy-oriented software 'TOPCAT' (Taylor 2005). MM acknowledges the support by the Spanish Ministry of Science, Innovation and University (MICIU/FEDER, UE) through grant RTI2018-095076-B-C21, and the Institute of Cosmos Sciences University of Barcelona (ICCUB, Unidad de Excelencia 'María de Maeztu') through grant CEX2019-000918-M.

DATA AVAILABILITY

Both the light and the full versions of our meta-catalogue can be found in VizieR as 'The *Gaia*/IPHAS catalogue of H α-excess sources'.

REFERENCES

Abrahams E. S., Bloom J. S., Mowlavi N., Szkody P., Rix H.-W., Ventura J.-P., Brink T. G., Filippenko A. V., 2020, preprint (arXiv:2011.12253)
Astraatmadja T. L., Bailer-Jones C. A. L., 2016, ApJ, 833, 119
Barentsen G., Vink J. S., Drew J. E., Sale S. E., 2013, MNRAS, 429, 1981
Barentsen G. et al., 2014, MNRAS, 444, 3230
Belloni D., Schreiber M. R., Pala A. F., Gänsicke B. T., Zorotovic M., Rodrigues C. V., 2020, MNRAS, 491, 5717
Bressan A., Marigo P., Girardi L., Salasnich B., Dal Cero C., Rubele S., Nanni A., 2012, MNRAS, 427, 127
Carrasco J. M., Catalán S., Jordi C., Tremblay P.-E., Napiwotzki R., Luri X., Robin A. C., Kowalski P. M., 2014, A&A, 565, A11
Davies R. D., Elliott K. H., Meaburn J., 1976, Mem. R. Astron. Soc., 81, 89
de Kool M., 1992, A&A, 261, 188
Dennison B., Simonetti J. H., Topasna G. A., 1999, Am. Astron. Soc. Meeting Abstr. 195, 53.09
Dias W. S., Monteiro H., Caetano T. C., Lépine J. R. D., Assafin M., Oliveira A. F., 2014, A&A, 564, A79
Drew J. E. et al., 2005, MNRAS, 362, 753
Drew J. E. et al., 2014, MNRAS, 440, 2036
Evans I. N. et al., 2010, ApJS, 189, 37
Farnhill H. J., Drew J. E., Barentsen G., González-Solares E. A., 2016, MNRAS, 457, 642
Gaustad J. E., McCullough P. R., Rosing W., Van Buren D., 2001, PASP, 113, 1326
Glazebrook K., Peacock J. A., Collins C. A., Miller L., 1994, MNRAS, 266, 65
Groot P. J. et al., 2009, MNRAS, 399, 323
Irwin M., Lewis J., 2001, New Astron. Rev., 45, 105
Kharchenko N. V., Piskunov A. E., Schilbach E., Röser S., Scholz R. D., 2013, A&A, 558, A53
Kohoutek L., Wehmeyer R., 1999, A&AS, 134, 255
Kolb U., 1993, A&A, 271, 149
Kuhn M. A., Hillenbrand L. A., Sills A., Feigelson E. D., Getman K. V., 2019, ApJ, 870, 32
Lindegren L. et al., 2018, A&A, 616, A2
Magnier E. A. et al., 2013, ApJS, 205, 20
Mohr-Smith M. et al., 2017, MNRAS, 465, 1807
Monguió M. et al., 2020, A&A, 638, A18
Pala A. F. et al., 2020, MNRAS, 494, 3799
Parker Q. A. et al., 2005, MNRAS, 362, 689
Raddi R. et al., 2013, MNRAS, 430, 2169

Sale S. E. et al., 2014, MNRAS, 443, 2907
Scaringi S. et al., 2018, MNRAS, 481, 3357
Skrutskie M. F. et al., 2006, AJ, 131, 1163
Taylor M. B., 2005, in Shopbell P., Britton M., Ebert R., eds, ASP Conf. Ser. Vol. 347, Astronomical Data Analysis Software and Systems XIV. Astron. Soc. Pac., San Francisco, p. 29
Voges W. et al., 1999, A&A, 349, 389
Voges W. et al., 2000, IAU Circ., 7432, 3
Wenger M. et al., 2000, A&AS, 143, 9
Witham A. R., Knigge C., Drew J. E., Greimel R., Steeghs D., Gänsicke B. T., Groot P. J., Mampaso A., 2008, MNRAS, 384, 1277
Wramdemark S., 1981, A&AS, 43, 103
Yao S. et al., 2019, ApJS, 240, 6

SUPPORTING INFORMATION

Supplementary data are available at https://drive.google.com/file/d/1eHDYa_HltSPQjKjPaay7T_sM7YMlqqM2/view?usp=sharing online.

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.