Identification and classification of H alpha-excess candidates in the northern Galactic Plane.

MATTEO FRATTA

Identification and classification of Hα-excess candidates in the northern Galactic Plane.

Matteo Fratta

A thesis presented for the degree of
Doctor of Philosophy

Centre for Extragalactic Astronomy
Department of Physics
Durham University
United Kingdom

January 2023

Identification and classification of Hα-excess candidates in the northern Galactic Plane.

Matteo Fratta

Abstract: The key-role of studying Hα emission from astronomical sources has been well established in the past decades. Its relevance is set by the impressively wide range of Hα-emitting populations, the inspection of which can significantly contribute to understanding our Universe in terms of its evolution and composition. Therefore, finding and classifying new emission-line sources has become a fundamental task in modern astronomy, fueled by the availability of a constantly increasing number of wide field surveys.

State-of-the-art techniques to identify Hα emission line sources in narrow-band photometric surveys consist of searching for Hα excess with reference to nearby objects in the sky (position-based selection). However, while this approach usually yields very few spurious detections, it may fail to select intrinsically faint and/or rare Hα-excess sources. In order to try and obtain a more complete representation of the heterogeneous emission line populations, I develop a technique to find outliers relative to nearby objects in the colour-magnitude diagram (CMD-based selection). By leveraging astrometric and photomeric information from *Gaia* and by combining position-based and CMD-based selections, I build an updated catalogue of Hα-excess candidates from data obtained in the INT Photometric Hα Survey of the Northern Galactic Plane (IPHAS). Out of 7,474,835 inspected targets, 28,496 Hα-excess candidates are identified. I further perform spectroscopic follow-up observations and classification of 114 objects from this catalogue, that enable to test the novel selection method. Out of the 69 spectroscopically confirmed Hα emitters in the sample, 14 are identified only by the CMD-based selection, and would be thus missed by the classic position-based technique.

In addition, I explore one of the possible applications of the CMD-based selection. More specifically, I describe how accreting white dwarf candidates can be identified among the Hα-excess sources, with the introduction of information regarding their X-ray emission and variability amplitude. A specific application of these criteria is provided by the target HBHA 4204-09. Spectroscopically confirmed as a cataclysmic variable, this source is part of the Hα-excess candidates that also show a significantly large variability amplitude. The analysis

of its light curve supports its further classification as a nova-like variable.

Moreover, I show how the CMD-based selection can be augmented with the introduction of additional colours. Especially with this latter improvement, the CMD-based technique can be used to support the pre-selection and classification of emission line sources in large spectroscopic surveys.

Contents

List of Figures

List of Tables

Declaration

The work in this thesis is based on research carried out in the Centre for Extragalactic Astronomy, Department of Physics, Durham University, United Kingdom. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text.

The content presented in Chapter 2, as well as part of the content presented in Chapter 4 have been published in:

Fratta, M., Scaringi, S., Drew, J.E., et al. *Population-based identification of Hα-excess sources in the Gaia DR2 and IPHAS catalogues.* Monthly Notices of the Royal Astronomical Society, **505**(1) (2021), 1135. 2105. 00749.

The content presented in Chapter 3 has been published in:

Fratta, M., Scaringi, S., Monguió, M., et al.*Spectroscopic follow-up of a sub-set of the Gaia/IPHAS catalogue of Hα-excess sources.* arXiv e-prints (2022), arXiv:2211.01694. 2211.01694.

The author of the thesis was primarily responsible for all aspects of this publication.

Acknowledgements

First of all, I would like to thank Dr. Simone Scaringi for supporting and encouraging me throughout my Ph.D. It has been an amazing journey, that made me grow professionally, but most importantly, from a personal point of view.

Thanks to all the people that I was lucky enough to meet during these five incredible years. Being more than simple friends, they constantly remind me that I have a family all around the World.

Thanks to my friends in Italy, the newly found ones, and the ones that have been there since I have memory. They made me feel missed in the periods I spent away, and every time I came back home it has been like I never left.

Last but not least, thanks to my family, for unconditionally supporting me in any possible way regardless on how far my choices would have taken me. Even while being in opposite sides of the Planet, we have never really been apart. Thanks for believing in me, even in the times I myself did not.

To all of you, and to those I did not have the chance to mention here, thanks. Perhaps, I could have made it through, perhaps not. But you are definitely the ones that made it worth it.

The cross-matches, as well as some of the figures used for this thesis were produced with the use of the astronomy-oriented software "TOPCAT" (Taylor, 2005).

Part of the work presented in this thesis is based on observations obtained at the international Gemini Observatory, a program of NSF's NOIRLab, which is managed by the Association of Universities for Research in Astronomy (AURA) under a cooperative agreement with the National Science Foundation, on behalf of the Gemini Observatory partnership: the National Science Foundation (United States), National Research Council (Canada), Agencia Nacional de Investigación y Desarrollo (Chile), Ministerio de Ciencia, Tecnología e Innovación (Argentina), Ministério da Ciência, Tecnologia, Inovações e Comunicações (Brazil), and Korea Astronomy and Space Science Institute (Republic of Korea). The

analysed spectra were acquired through the Gemini Observatory Archive at NSF's NOIRLab* and processed using the Gemini IRAF package. Program IDs: GN-2019B-Q-401 and GS-2019B-Q-403.

"Dang it, Jim, I am an astronomer, not a doctor. I mean, I am a doctor, but I am not that kind of doctor. I have a doctorate, but it's not the same thing. You can't help people with a doctorate, you just sit there and you are useless!" (Dr. Delbert Doppler).

Chapter 1

Introduction

The abundance of Hydrogen in our Universe makes Hα emission (also called "Balmer α" emission) a fundamental feature in astronomical studies at optical wavelengths. Due to the diversity of sources that present this hallmark (either permanently or during circumscribed evolutionary stages), population studies of Hα-bright objects can substantially improve our comprehension of the structure and evolution of the nearby Universe, including our own Galaxy. In order to produce unbiased results, these analyses require reliable and complete representations of the many Hα-emitting populations. Historically, inquiries to identify and classify these types of objects have prioritised low rates of spurious detections, sometimes at the expenses of more complete selections. I here describe a novel algorithm to identify Hα-excess candidates in large photometric catalogues, with reference to nearby objects in calibrated color-magnitude diagrams (CMDs). Providing a "population-based" selection, the outcomes of this approach are particularly valuable, especially if inserted in the context of a pre-selection of Hα-excess candidates in large spectroscopic surveys.

In this Chapter, I first briefly describe the physical processes behind the emission in the Hα band, and give an overview of some point-like objects that present it (Section 1.1). Section 1.2 lists some of the most relevant wide field surveys (mainly photometric and spectroscopic). Section 1.3 outlines the modern photometry-based selection techniques of Hα-excess candidates.

1.1 Hα-emitting sources

Balmer emission lines are produced by the de-excitation of Hydrogen. In particular, the Balmer α line is released when an electron transits from principal quantum number n=3 to

the energy level n=2 (Fig. 1.1).

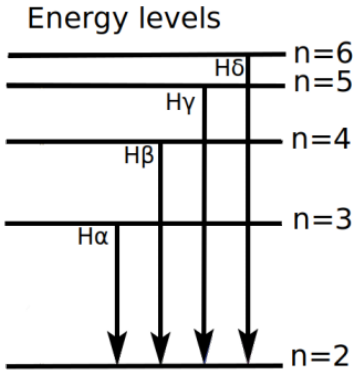


Figure 1.1: Schematic representation of the de-excitation channels of atomic Hydrogen that are responsible for the emission of Balmer lines.

In gases at sufficiently high temperature and concentration, electrons detach from the H nuclei, rather than being excited. Ionised H atoms recombine with the electrons at high energy states, which then cascade towards lower levels. During this process, optical radiation is emitted, including at the Hα wavelength (6564 Å). Alternatively, said gases might reprocess light from an external source. If the incident light is energetic enough, it may be (partially or totally) absorbed by the H atoms, to promote electrons to higher energy levels. A continuum spectrum with discrete absorption features would result from the spectroscopic observation of this reprocessed light.

The de-excitation channel of Hydrogen from n=3 to n=2 is associated to a relatively high branching ratio, with respect to other channels. Therefore, Hα emission is generally more intense than other Hydrogen lines, and hence easier to detect. For this reason, Hα emission (rather than other lines) has been extensively studied in optical astronomy as a direct indicator of high concentrations of ionised Hydrogen. Since Hydrogen is the most abundant element in the Universe, the suitable conditions for the production of Balmer radiation are met by a broad range of extremely diverse objects. The technique of Hα-imaging has been largely used to identify and delineate extended nebulosities (Kennicutt et al. 2008, Gavazzi et al. 2003), mostly linked to stellar remnants or star-forming H_{II} regions. There, a hot central object, such as O-B spectral type star, ionises the surrounding Hydrogen gas. As an example, Fig. 1.2 shows an image of the compact star-forming region Sh 2-106 in the Cygnus constellation, as observed by the Hubble telescope. The recombination at high energies of the electrons with the nuclei and the following de-excitation cause the production of Hα radiation. Hα photometry is often inspected to quantify the local star-formation rates (SFRs) up to redshift $z \sim 3$ (Kennicutt 1983, Bechtold et al. 1997, Lee et al. 2009, Vučetić et al. 2015,

Figure 1.2: Image of the compact star-forming region Sh 2-106, observed by the NASA/ESA Hubble Space Telescope. The Hα emission in this picture is represented with the blue colour. Credit: NASA & ESA

Tacchella et al. 2022). However, these values might be affected by dust attenuation and/or by external sources.

On the other hand, unresolved Hα emission can be associated with very heterogeneous populations of both single and binary stars, spanning a wide spectrum of evolutionary stages. While some of these objects are frequently observed, some others might be intrinsically very faint and/or belong to the rarest types of stellar populations. Isolated point-like sources include, but are not limited to: classical Be (CBe) stars, chromospherically active M-Dwarfs, luminous blue variables (LBVs), and Wolf-Rayet (WR) stars. Among the point-like Hα-emitting populations, accretion driven line-emitters can also be included. This list contains the young stellar objects (YSOs) population, but also interacting binaries, such as cataclysmic variables (CVs), symbiotic stars (SySts), and X-ray binaries (XBs) like accreting neutron stars (NS) or stellar mass black holes (BH).

In the next Sections, I briefly describe some of the Hα-emitting populations, with specific focus on point-like sources.

1.1.1 Young stellar objects

This class includes stellar objects that have not yet reached the Main Sequence (MS) phase. They are formed in the densest regions of molecular clouds, and can be surrounded by an IR-

bright protoplanetary disk. This population is further divided into T-Tauri stars and Herbig Ae/Be stars, based on their masses: YSOs with an initial mass lower than $2\,M_\odot$ are classified as T-Tauri stars, whilst the masses of Herbig stars are approximately included in the 2-10 $M_\odot$ range. More massive objects become visible only when they reach the conditions to trigger Hydrogen fusion in their cores, thus concluding their (relatively short) pre-Main-Sequence (PMS) cycle (Waters & Waelkens, 1998).

In low/mid-mass YSOs, the inner rings of the circumstellar disk are disrupted by the magnetosphere of the protostar. Matter is bound to follow the magnetic field lines, and to accrete onto the central star. This process is (mainly) responsible for the production of Balmer emission lines, in this type of sources (Mortier et al., 2011). The equivalent width (EW) of the Hα line is also used to further classify T-Tauri stars, with respect to their evolutionary phase. Classical T-Tauri stars (CTTS) are relatively young systems, and are accreting matter from the circumstellar disk. They are distinguished by relatively large EWs. On the other hand, Weak-line T-Tauri stars (WTTS; also called "Naked T-Tauri stars, or NTTS) are older, and are surrounded by little circumstellar material (Walter, 1986). The accretion rate is lower in this phase, and the typical EWs are much smaller. The standard EW cutoff between CTTS and WTTS is 10 Å (Strom et al., 1989).

Theoretically, the location of PMS stars in the Hertzsprung-Russel diagram (HRD) is described by a best-fit isochrone. However, observed YSO clusters unexpectedly spread around this locus (Marton et al., 2019). A possible explanation for this effect, given by Audard et al. (2014), involves variable infall rates: long-lasting phases in which matter collapses at a rate of $\sim 10^{-7}\,M_\odot yr^{-1}$ alternate with rare, shorter high-accretion rate ($\sim 10^{-4}\,M_\odot yr^{-1}$) phases. This is in contrast with models that predict stable infall rates of $\sim 10^{-6}\,M_\odot yr^{-1}$. This variability would also explain the observed luminosity of Class I YSOs, which is 10-100 times lower than that predicted for an accretion rate of $\sim 10^{-6}\,M_\odot yr^{-1}$. However, a larger representation of the diverse YSO population is required, in order to unambiguously confirm this explanation.

1.1.2 Classical Be stars

CBe stars are defined as rapidly rotating[1] MS stars of early B spectral type, that present (or presented) Balmer emission lines in their optical spectra (Collins, 1987). According to the

[1] On average, the rotational velocity of this objects is very close to the centrifugal limit. The origin of such a high angular velocity is not yet completely understood.

study presented in Keller et al. (1999), the fraction of CBe stars among the MS stars in a cluster seems to be related to the ellipticity of the cluster itself. Included between the 10-34% percentage range, the authors observe the highest ratio of CBe stars in highly elliptical and fast rotating clusters. On the other hand, no relation between the age/metallicity of the cluster and the CBe fraction is observed.

These stars are surrounded by a "decretion" Keplerian disk, the formation mechanism of which has not been uniquely unravelled, yet (Rivinius et al. 2013, Pringle 1992). Said disks do not contain dust, and therefore the spectral energy distributions (SED) of CBe stars present a much smaller infrared (IR) excess, compared to that produced by the thermal emission from dust-dominated disks, in Herbig Be stars. This criterion is commonly used to discern these two stellar populations.

While free of dust, decretion disks in CBe stars are mainly made of H gas. These disks are the main responsible for Hα emission in CBe stars, as their inner rings are heated up and ionised by the central objects. This feature is highly variable, in this type of objects; for this reason, it is not unusual to refer to them as "in the CBe phase". The CBe phase can last from ~ 10 to ~ 50 years, with increasing duration for later spectral types (Hubert & Chambon, 1976).

1.1.3 M-dwarfs

Chromospherically-active M-dwarfs are late-type, low-mass, MS stars that present Hα emission. As in most of the solar-like stars, their Hα intensity is used as an indicator of chromospheric activity (Young et al., 1989). The chromosphere of this type of objects is heated up by the inwardly directed X-rays from the corona (Cram 1982, Giampapa et al. 1982), as well as by the intense magnetic activity of the star (Gunning et al., 2014). Besides in the Balmer lines, their spectra often exhibit emission in the Mg II and Ca II lines (Linsky et al., 1982). Objects belonging to this stellar population are often searched for (especially with the availability of *TESS* light curves) in relation to the discovery of new possible transiting planets. In fact, planets orbiting around an M-dwarf might belong to the habitable kind. However, the intense magnetic fields associated to coronally active M-dwarfs hinder the habitability of orbiting planets. Therefore, studies of the fraction of Hα-emitting M-dwarfs (Kiman et al., 2021) can support models that predict the distribution of habitable planets.

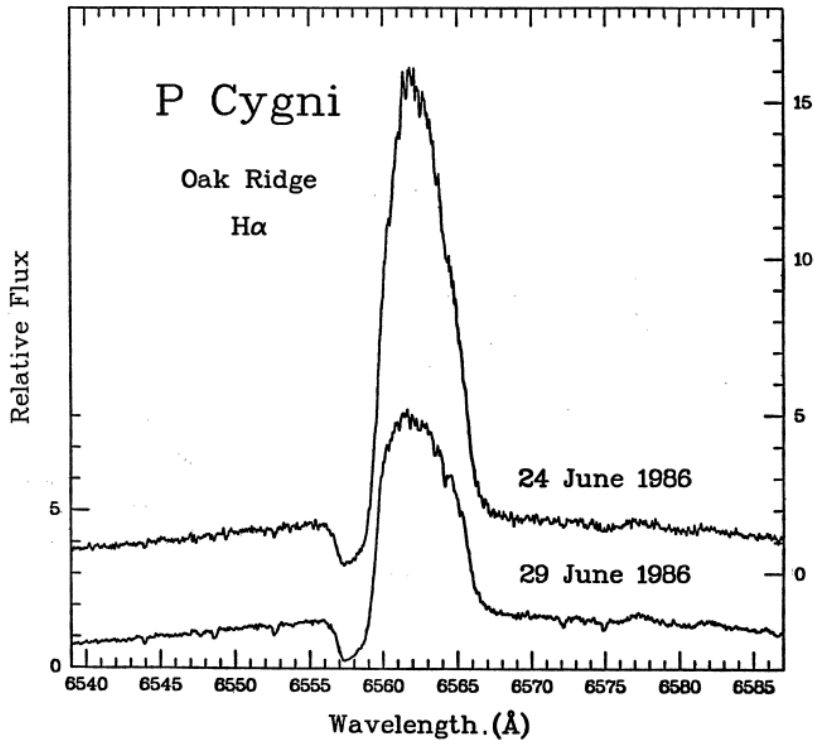


Figure 1.3: Two Hα line profiles of the source P Cygni, observed at different epochs. Observed at the Oak Ridge Observatory (Massachusetts), this image is taken from Fig. 3 of de Groot (1988).

1.1.4 Luminous blue variables

As their name suggests, LBVs (also called S Doradus stars) are very bright supergiants ($10^5 - 10^7\,L_\odot$), with masses larger than $\sim 21\,M_\odot$. Additionally, these objects are characterised by the emission of very intense winds (Weis & Bomans, 2020). The first LBV to be identified as such was P Cygni. This source is well known, as it was the first to show the hallmark now referred to as "P Cygni profile" (Fig. 1.3) in its spectrum. This involves the observation of emission lines at the rest wavelengths, along with absorption at immediately shorter wavelengths (de Groot, 1988). This phenomenon is caused by the interaction between the isotropic stellar radiation and the portion of the winds emitted in direction of the line of sight. Fig. 1.4 shows the geometry of the physical process behind the shape of the P Cygni profile. It is impossible for an observer (O) to collect photons that are emitted in the direction of region B (i.e., "away" from the observer). The collection of photons that were emitted specifically in the direction of the regions E would produce a spectrum with broad and symmetrical emission lines. However, a portion of these photons have their trajectories diverted by scattering with the outflow in region A, preventing them from reaching the observer. Since this matter is moving in the direction of the observer, the scattered photons are

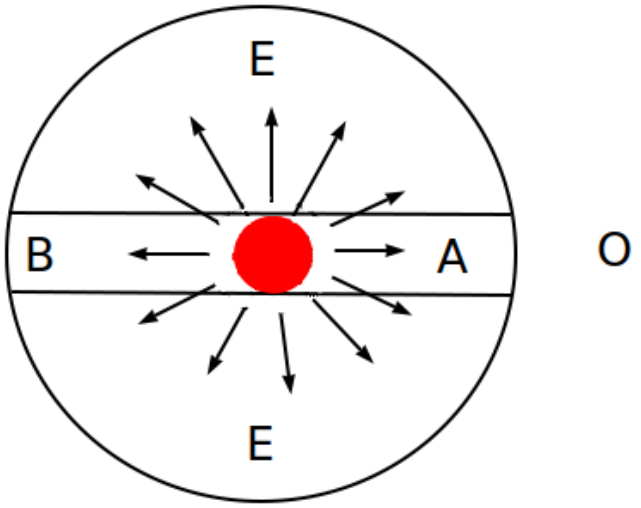


Figure 1.4: Visual description of the geometry behind the P Cygni profile phenomenon. The red full circle is the stellar source, while the larger concentric ring represents the surrounding shell. The arrows depict the radiation spherically emitted from the central source. The observer is denoted with the letter O, while the letters A and B highlight the regions in which the outflow moves in the direction of the lines of sight, respectively towards and away from the observer.

blue-shifted, resulting in the typical emission line shape shown in Fig. 1.3. Since its discovery, the P Cygni profile has been observed in many other spectra belonging to massive sources.

Due to their very large mass and luminosity, LBVs are very short-lived in comparison with other massive sources, and hence are relatively rare. In addition, their spectral features are not very dissimilar to those of OB supergiants. This makes LBVs even harder to efficiently identify and classify. However, they can be distinguished from other evolved stars for the high variability. Historically described as an *eruptive* behaviour, LBVs alternate optically faint (hot) phases with bright (cold) phases, with a difference of 1-2 mag between the two states. These photometric (and spectroscopic) changes are larger in more luminous LBVs; this effect is called *amplitude-luminosity-relation* (Wolf, 1989). Some LBVs also show additional, more energetic eruptions, during which their optical intensity can increase by up to ~ 5 mag (Humphreys et al., 1999). The time interval between two successive *giant eruptions* can be on the order of centuries. For these reasons, in the past these phenomena have been often mistaken for Supernovae explosions. The origin of the giant eruptions has not been established yet, as well as their possible connection with "normal" S Doradus eruptions. However, the most credible scenarios to try and explain this sudden increase in brightness involve pressure and dynamical instabilities (Humphreys & Davidson, 1994a). During these giant outbursts,

Figure 1.5: Hubble image of the luminous blue variable AG Carinae. Credit: NASA, ESA, STScI.

LBVs can expel up to several $M_\odot$, with mass loss rates of $10^{-5} - 10^{-4}\,M_\odot yr^{-1}$. This matter accumulates in circumstellar ring nebulae or shells Humphreys & Davidson (1994b). The interaction between successive outbursts and the surrounding shell produces the emission in the Hα band. As an example, a Hubble image of the LBV AG Carinae, taken with the UV and optical filters, is shown in Fig. 1.5.

1.1.5 Wolf-Rayet stars

One of the possible outcomes of LBV evolution, mainly driven by the intense mass loss, is the formation of a WR star, surrounded by a nebula. They are considered to be Supernovae (SN) progenitors, but the most massive of them might skip this step and directly turn into a BH via gravitational collapse. Since these evolved objects represent the final stages of very massive ($M > 25\,M_\odot$) MS stars, they are generally rich in heavy elements. However, while most of the spectra of MS stars are dominated by narrow absorption lines, WR stars present strong, broad emission lines. The production of these lines (including in the Hα band), as well as their widths, can be explained by the interaction between the intense winds ($800 - 3000\,km\,s^{-1}$) that regulate the evolution of these objects[2], and the surrounding nebula. While these sources are intrinsically rare, their high luminosity, as well as the presence of strong emission lines in their spectra, make them easy to discern from other evolved giants, especially at short distances. Around 200 WR stars have been identified in our Galaxy, but

[2] with mass loss rates comparable to those of LBVs (Hillier, 1996, 2000)

many more might be obscured by interstellar dust; from this follows the necessity of the investigation of new methods for identifying WR stars.

Based on the relative intensity of the observed spectral lines, WR stars are divided into three classes, named WN, WC and WO, that reflect their evolutionary stages. As these sources evolve, their spectra are dominated by emission from heavier elements: spectra of WN stars are commonly dominated by N and He emission lines, although some of them also present contribution from C and H in emission. The spectra of the other two classes mainly show C and He emission, but no contribution from H and N is seen. The most relevant difference between WC and WO classes is that WO stars show very prominent O lines, while WC stars do not.

1.1.6 Cataclysmic variables

CVs are binary systems in which a white dwarf (WD) is accreting mass from a (typically low-mass) secondary MS star. The dynamical interaction between the two components of the system is described by *Roche potential* Φ

$$\Phi = -\frac{GM_1}{|\mathbf{r}-\mathbf{r_1}|} - \frac{GM_2}{|\mathbf{r}-\mathbf{r_2}|} - \frac{1}{2}\cdot(\mathbf{\Omega}\times\mathbf{r})^2, \tag{1.1.1}$$

that owes its name to Edouard Roche, the French mathematician who first studied it. Roche's description assumes that the components of the system trace a circular orbit[3] around a common centre of mass (CM), and that their masses are concentrated in their centres. In this context, G is the gravitational constant; $\mathbf{r}$ is a generic position, with respect of the CM, in which the potential is probed; M_1 and M_2 are the masses of the WD and of the companion star, respectively; $\mathbf{r_1}$ and $\mathbf{r_2}$ are the positions of the WD and of the MS star, with respect of the CM; $\mathbf{\Omega}$ is the orbital angular frequency of the system. The first two terms in Equation 1.1.1 represent the gravitational potential of the two components of the system, whereas the third one is the centrifugal term, that takes into account the non inertial frame of reference. Roche geometry is illustrated in Fig. 1.6. The surfaces in which Roche potential is constant (i.e., where the vectorial sum of the forces is zero) are called *equipotentials*, and they meet in the five *Lagrangian points* ($L_{1\text{-}5}$, the red dots in Fig. 1.6). L_1 is also called *inner Lagrangian point*, due to its position in between the two bodies. The "8-shaped" equipotential surface that passes through this point is called *Roche Lobes* (RLs; one lobe for each component of the system).

[3]This assumption is justified by the fact that the timescale for the accretion process is much longer than the circularisation timescale.

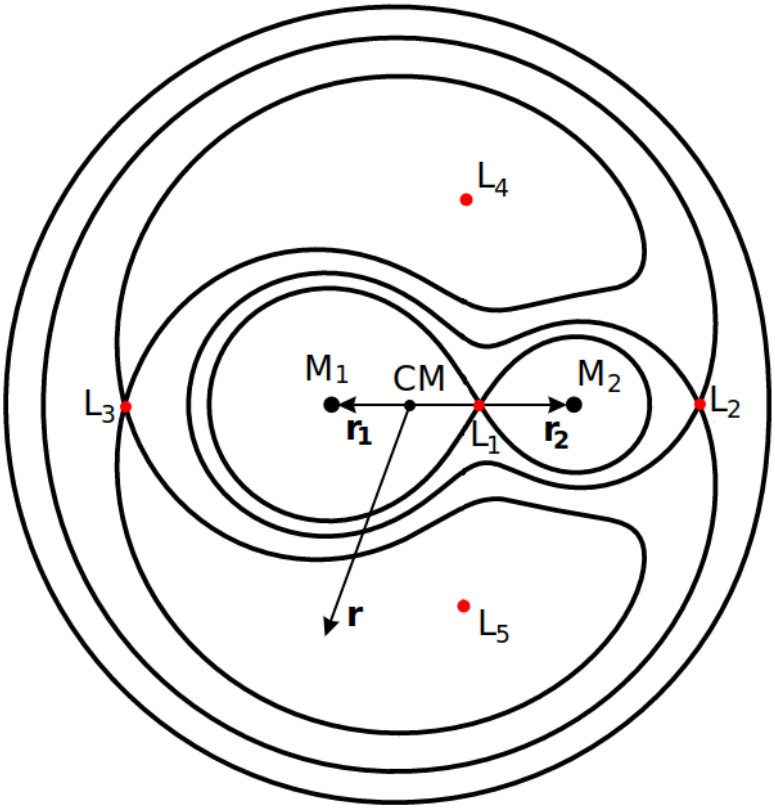


Figure 1.6: Schematic representation of the dynamic interaction between the WD and the companion MS star in a binary system. The potential of this interaction takes into account by the gravitational attraction of the two bodies (of masses M_1 and M_2, respectively), as well as the centrifugal term related to the rotation of the system around the centre of mass (CM). The arrow marked by the letter **r** points to a generic position in which the potential is evaluated. The closed lines represent the surfaces in which Roche potential is constant, i.e. the *equipotentials*. These surfaces meet in the five Lagrangian points ($L_{1\text{-}5}$), depicted by the red dots. To produce this picture, I adapted Fig. 16.9 from Tauris & van den Heuvel (2006).

Matter within a lobe is gravitationally attracted to the corresponding star, predominantly, while in coincidence of the inner Lagrangian point it is in dynamical equilibrium. The sizes of the RLs are proportional to the ratio of the masses of the components of the binary system, as well as to the orbital separation. At some points in the evolution of the system, the photosphere of companion star may fill its RL, and matter may therefore flow through L_1 towards the WD. This process is known as *Roche Lobe overflow*, and its occurrence marks the beginning of the CV phase. Typical mass transfer rates for CVs span from 10^{-11} to $10^{-8}\,M_{\odot}yr^{-1}$ (Avilés et al., 2018).

The most common scenario in which the donor star fills its RL involves the contraction of the binary system, with the subsequent decrease in the radii of both RLs. The variation of the separation between the components of the system is regulated by angular momentum losses (AML). The main factors that contribute to these AML are magnetic braking and gravitational radiation emission. While gravitational waves are produced by the space-time deformation related to the revolution of the two bodies around the CM, magnetic braking is connected to the interaction between the winds outflowing from the secondary star and its magnetic field. AML cause the orbital period of CVs to shorten from initial values of the order of few hours, to tens of minutes. The observed values for this parameter are included between ~ 80 min and ~ 10 hr (Avilés et al., 2018), with some exceptions in the shorter and longer ends of the distribution. However, very few CVs are found with orbital periods included between 2 and 3 hours; for this reason, this interval is called "period gap" (Rappaport et al., 1983). The lack of CVs in this period range can be attributed to the secondary star entering a fully convective phase of its evolution. The theory described in Kolb (1993) predicts that, after the system has reached the minimum period, the donor has lost enough mass to stop fusing H in its core, and to become a degenerate brown dwarf. In this scenario, since the orbital period of the CV scales with the inverse of the square root of the density of the donor star, the system starts expanding again (Paczynski, 1981), with the mass transfer proceeding at much lower rates. CVs in this phase are addressed as "period bouncers". According to Kolb (1993), $\sim 70\%$ of the CVs should be period bouncers. Nonetheless, despite these predictions, just a limited amount of period bouncers have been identified. Gänsicke et al. (2009) recently studied the orbital periods distribution of 137 faint CVs in the Sloan Digital Sky Survey (SDSS; Gunn & Knapp 1993), and found that these mostly cluster between 80 min and 86 min. While this discovery seems to point to the reliability of the theoretical predictions, it does not provide conclusive evidence to rule out selection biases as a cause for the discrepancy between predicted and observed density of period bouncers (Pala et al.,

2018). However, it highlights the relevance of developing new techniques to identify CVs.

Matter that flows into the RL of the WD through the inner Lagrangian point still carries its angular momentum, and therefore it cannot directly fall onto the surface of the WD. Instead, it starts orbiting around the WD to form a hot accretion disk. From a kinetic point of view, this disk can be described as formed by infinitesimally thin concentric rings, each rotating around the WD with its own Keplerian velocity. The velocity difference between two adjacent rings causes friction, that in turn increases the temperature of the orbiting gas. This effect is responsible for the optical emission from the disk, including at the Hα wavelength. Moreover, it causes the dissipation of angular momentum towards outer rings. As the angular momentum is transmitted to larger radii, the matter in the disk slowly spirals in until it reaches the magnetospheric radius of the WD. This represents the distance from the centre of the WD at which the pressure of the magnetic field of the WD (directed radially outwards) equals the pressure of the free-falling matter (directed inwards). At this point, the ionised atoms are bound to follow the magnetic field lines, and to accrete onto the magnetic poles of the WD. However, if the magnetic field of the WD is so intense that the magnetospheric radius exceeds the radius of the RL, no disk can be formed. These types of CVs are called *polars*. On the other hand, if the magnetospheric radius of the WD is smaller than its physical radius, matter in the disk keeps on slowly spiraling in, until it reaches the surface of the WD. These CVs are the canonical "non-magnetic" CVs. Unsurprisingly, CVs in which the accretion disk can form, but is truncated to the magnetospheric radius are named *intermediate polars*. Another classification, based on the intensity of the mass-accretion-rate, divides the CV population in *dwarf novae* and *nova-like variables* (or simply *nova-likes*). The low mass-accretion rates in dwarf novae reflect in the temperature of the accretion disk, which is low enough to trigger thermal instabilities. These instabilities produce in turn repeated dwarf nova outbursts. On the other hand, nova-likes are characterised by higher mass-accretion rates, which mirror in a stable accretion disk, and considerably less frequent outbursts in their light curves (Wargau, 1985; Knigge et al., 2011).

1.1.7 Symbiotic stars

As CVs, SySts are interacting binary systems, with a central WD. However, the donor star in SySts is a red giant (RG), that transfers matter to the WD primarily via stellar winds (Bondi & Hoyle, 1944), but also via RL overflow. Such intense winds also cause the formation of H_{II} regions around the system. As in the CVs, in most of the SySts the transferred matter forms

an accretion disk around the WD. The number of known SySts in the Galaxy is relatively low, considering the large number of RGs that populate it. Similarly to the CV case, the low empirical density of SySts in the Galaxy is mainly imputed to observational biases, and pushes towards the achievement of new methods for the identification of this type of systems.

Due to the sizes of the RG, SySts are very wide binary systems, with typical orbital separations up to tens of AU (Eyres et al., 2001). As a consequence, their orbital periods are among the longest measured in binary systems, spanning from hundreds to thousands of days (Belczyński et al., 2000). Such long periodicities are hard to observe, and would require years of data acquisition. Based on their near-IR colours, SySts are split into two classes (Webster & Allen, 1975): "S-type" (or "stellar" type), where the IR colour is compatible with a G-type or M-type donor giant; "D-type" (or "dusty" type), the IR emission of which implies a Mira variable companion. The orbital periods of these two groups differ substantially, as the ones relative to S-type SySts are about ten times shorter than those of D-type SySts (Whitelock, 1987).

Optical spectra of SySts commonly show the absorption features typical of G-type (or later types) stars. These result from the contribution to the SED from the companion star. However, relatively stable Oxygen, Helium and Hydrogen emission lines (including in the Hα band), originated from the ionised nebula that surround the system, are also observed (Luna et al., 2013). The innermost rings of the accretion disk (if one is formed) are the origin of X-ray emission from SySts, with the hardness of the X-ray spectrum depending on the mass of the WD: more massive WDs produce harder X-ray spectra than less massive ones. This is explained by the fact that WDs are degenerate objects; therefore, the larger their mass is, the smaller they are in size. The accreting matter that falls into the deeper potential wells associated to more massive WDs emits more energetic radiation, in the X region of the electromagnetic spectrum.

1.1.8 X-ray binaries

In these interacting binary systems, a donor star is accreting matter onto a NS or a stellar mass BH. Since the accreting objects are considerably more compact than a WD, their gravitational potential wells are much deeper. Consequently, emission due to mass accretion onto a BH or a NS is dominated by X-rays, rather than by radiation at longer wavelengths. However, Hα emission is observed in this type of binary systems, as well. Produced in the hot accretion disks, this can either be emitted as thermal radiation from the hot Hydrogen gas, or can

result from X-ray scattering onto the inner rings of the disk (Charles & Coe, 2003).

As in CVs, the interaction between the two components of these binaries can be described by Roche geometry. Observed orbital periods range from few tens of minutes to hundreds of days, depending on the masses involved. X-ray binary systems are divided in two classes, based on the mass of the donor star: *low-mass X-ray binaries* (LMXBs), in which the mass of the companion star is less than $1\,M_{\odot}$, and *high-mass X-ray binary* (HMXB), in which the mass of the MS component is larger than $10\,M_{\odot}$. Systems belonging to the former class are generally older, and more compact than the latter. The typical mass-ratios between their components are similar to those observed in CVs, and for this reason they have comparable orbital separations and periods. Moreover, as in CVs, the main mechanism for mass-transfer in LMXBs is RL overflow. On the other hand, HMXBs are much wider systems, considering their larger mass-ratios, and the typical orbital periods last from few days to hundreds of days (Chou, 2014). The main contribution to mass transfer process in HMXBs is provided by the strong wind outflowing from the companion star. These objects are listed among the progenitors of Gravitational Waves (GWs) emitting sources. After the first GWs detection by the Laser Interferometer Gravitational-Waves Observatory (Abramovici et al. 1992); Abbott et al. 2016), the discovery of new HMXB has become an even more compelling task in modern astronomy.

1.2 Wide field surveys

Population studies are enabled by the availability of large amounts of data. The following Subsections supply a brief overview of some wide field surveys necessary to collect these data.

The development of wide surveys stems from the necessity to study large areas of the sky, rather than specific single targets. Other astronomical surveys, on the other hand, concentrate on sources that belong to the same stellar population or show similar characteristics. Surveys can be roughly classified based on the type of data they are designed to collect[4]: instruments that use broad-band and/or narrow-band filters to measure the intensity of the sources in an area of the sky are called *photometric* surveys; as their name suggests *spectroscopic* surveys are meant to collect the spectra of the sources they observe; surveys that repeatedly image the sky in search for variability are called *time-domain* surveys. Some surveys present hybrid configurations, and can provide multiple pieces of information. The following sections provide

[4]Which in turn reflects in very different instrumentation set-ups and characteristics.

brief descriptions of a few wide field surveys, mainly focusing on the optical and IR ranges of the electromagnetic spectrum. In fact, while surveys in other bands are available - such as *Fermi* (Michelson et al., 2010) in the γ range, the extended ROentgen Survey with an Imaging Telescope Array (*eROSITA*; Predehl et al. 2006) in the X-ray, and the NRAO VLA Sky Survey (NVSS; Condon et al. 1998) in the radio regime - the analyses outlined in this thesis mostly focus on the optical and IR regimes.

1.2.1 *Gaia*

Gaia is a space-based observatory launched in 2013 by the European Space Agency (ESA) with the main task of understanding the structure of our Galaxy, as well as its evolution, from kinematical and chemical points of view (Gaia Collaboration et al., 2016). While its main achievement is to provide parallax measurements with micro-arcsecond precision (unachievable from ground-based observatories; Lindegren et al. 2018), *Gaia* is also equipped with the instrumentation to perform photometric measurements. Inheriting said tasks from the ESA HIPPARCOS mission (ESA, 1997), *Gaia* started its observations in 2014, and obtained astrometric, photometric and spectro-photometric measurements for about one billion targets in the Milky Way (roughly 100 times the size of HIPPARCOS database).

Based on almost two years of observations, the second data release (*Gaia* DR2; Gaia Collaboration et al. 2018) augmented the previous data release by including parallaxes, proper motions and broadband photometry in the G (< 20 mag), G_{BP} and G_{RP} bands for 1.3 billion sources. These performances have been further enhanced with the third data release (*Gaia* DR3): not only have photometric measurements been obtained (in all *Gaia* bands) for 1.5 billion sources, but also low and high-resolution spectra for a subset of these targets are included (Gaia Collaboration et al., 2022).

1.2.2 IPHAS

The main purpose of the Isaac Newton Telescope (INT) Photometric Hα Survey of the Northern Galactic Plane (IPHAS, Drew et al. 2005), is to study the structure of the disk of our Galaxy. The value of IPHAS stems from its sensitivity, that is about 1000 times better than those of previous surveys (such as Kohoutek & Wehmeyer 1999). This allows IPHAS to observe many faint objects (up to 7 mag fainter; Barentsen et al. 2014) that were not included in previous surveys, and hence have a more complete view on the composition of our disk.

Operated in La Palma and initiated in 2003, IPHAS observes the optical sky with its Wide Field Camera (WFC) in the $-5°<b<+5°$ latitude range (for a total of $\sim$1,800 sq^2). It provides photometric measurements with the two broad-band r (<21 mag) and i (<20 mag) filters, as well as with the narrow-band Hα (<20 mag) filter (see also Irwin & Lewis 2001), for more than 200 million sources. Beyond the discovery of new point-like Hα-excess sources, as well as extended nebulosities, IPHAS colours enabled the exploration of new applications such as the production of 3D extinction maps across the Galactic Plane (Sale et al., 2014).

1.2.3 UVEX

Based on La Palma, and with the use of the 2.5 m INT, the UV-Excess Survey of the Northern Galactic Plane (UVEX; Groot et al. 2009) observes a sky area of 10° × 185°, centred on the Galactic Equator. The combination of UVEX, the VPHAS+ survey (VST Photometric Hα Survey of the Southern Galactic Plane and Bulge; Drew et al. 2014) and IPHAS constitutes the European Galactic Plane Surveys (EGAPS; Groot et al. 2006), that encompass the entire Galactic Plane. UVEX provides photometric measurements in the U, g, r and He$_{I5875}$ bands, with an upper limiting magnitude of 21-22 mag, using the same set-up implemented for IPHAS. However, the Helium filter became defective in a relatively early stage of this survey.

UVEX was designed with the main task of identifying stellar remnants, both single and in binaries. Moreover, since U band measurements are particularly sensitive to dust attenuation, the collaboration between IPHAS and UVEX might also contribute to better constrain a dust map of the Galactic Plane. More specifically, it may support the distinction between unreddened stars of late types and reddened stars of the early types (Groot et al., 2009).

1.2.4 *WISE*

The Wide-field Infrared Survey Explorer (*WISE*) was an all-sky survey funded by NASA, that started in December 2009. In six months, its 40 cm telescope scrutinised the IR properties of the entire sky in four bands (*W1*, *W2*, *W3*, and *W4*), centred at 3.4, 4.6, 12, and 22 μm (Wright et al., 2010). With its four million pixels, this survey achieved a remarkable sensitivity: more than 1,000 times better than IRAS (InfraRed Astronomical Satellite; Neugebauer et al. 1984) in the 12 μm band, and 50,000 times better than DIRBE (Diffuse InfraRed Background Experiment; Smith et al. 2004) in the two mid-infrared bands. Thanks to its features, *WISE* was particularly efficient in detecting (even old, faint) Brown Dwarfs

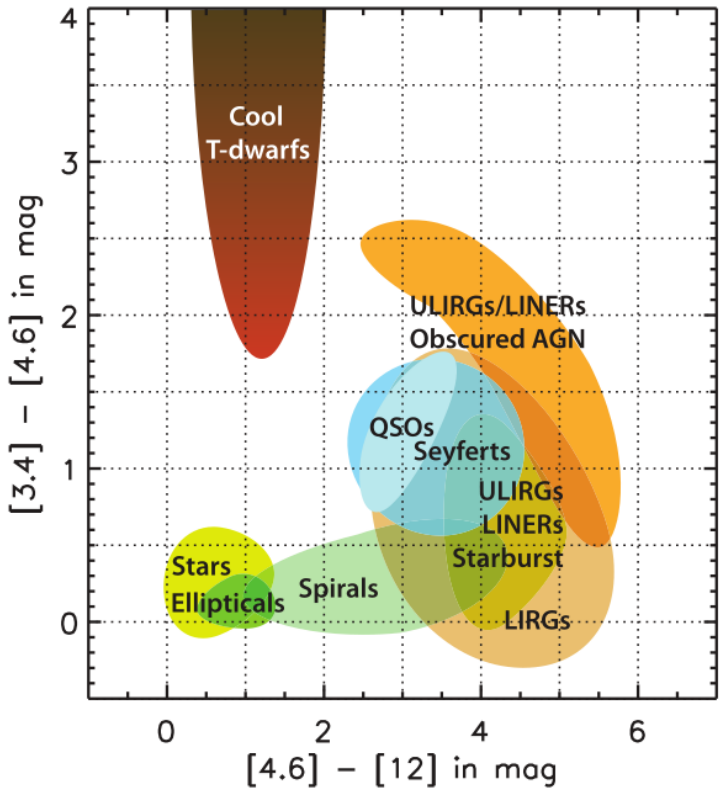


Figure 1.7: Graphic representation of the varied populations investigated by *WISE*, located in one of the *WISE* CCDs. More specifically, this figure display the locations in the *WISE* CCD of: brown T-dwarfs and other stars; elliptical and spiral galaxies; Seyfert galaxies; Quasi Stellar Objects (QSOs); Ultra-Luminous InfraRed Galaxies (ULIRGs) and Low-Ionization Nuclear Emission-line Regions (LINERs) obscured Active Galactic Nuclei (AGNs) and starbursts; Luminous InfraRed Galaxies (LIRGs). This image is taken from Fig. 12 in Wright et al. (2010).

(BDs), Ultra-Luminous InfraRed Galaxies (ULIRGs), Active Galactic Nuclei (AGNs) and Quasi Stellar Objects (QSOs). The distribution in one of the *WISE* colour-colour diagrams (CCDs) of these classes of objects is shown in Fig.1.7 (Fig. 12 from Wright et al. 2010):

1.2.5 *TESS*

As its name suggests, the Transiting Exoplanet Survey Satellite (*TESS*) is a NASA time-domain mission the main task of which is the identification of transiting nearby exoplanets (Ricker et al., 2015). Inheriting its duty from *Kepler* (Borucki et al., 2010), *TESS* searches for hundreds of Neptune-like or smaller planets, that are transiting bright stars with orbital periods down to few hours. Given the homogeneous distribution of these stars in the sky, *TESS* was designed as an all-sky survey. Considering the orbital period range that *TESS* aims at detecting, it collects brightness measurements with a 2 min cadence, or shorter. While the

principal focus of this photometric survey is clearly stated by its name, the wide sky coverage combined with the short cadence and the frequent data release make this telescope a perfectly suited tool to detect an incredibly large range of different variability features (e.g.: new CVs can be discovered by the analysis of the typical variabilities detected in their optical light curves).

1.2.6 SDSS

The 2.5 m telescope of the Sloan Digital Sky Survey, based at the Apache Point Observatory (New Mexico), began observing in 2000. The photometric system, retired in 2009, collected images with its five filters in the u, g, r, i, and z bands, with a limiting magnitude of ~ 22 mag. On the other hand, its two spectrographs cover the $4,000 - 9,000$ Å wavelength range; with the aid of their optical fibers, these coupled devices can simultaneously observe more than 600 spectra. With a π steradians field of view in the direction of the Northern Galactic cap, the main task of this photometric and spectroscopic redshift survey is to provide a detailed three-dimensional map of the nearby Universe, up to a redshift of $z = 0.2$. However, a particular focus was put on studying quasars and galaxies, up to a redshift of $z \sim 6$.

Among the many contributions from SDSS to the understanding of the structure of our Universe, this survey helped proving the key-role portrayed by dark matter and dark energy. Its spectra from (more than) 46,000 quasars helped proving that these objects evolve very slowly, compared to the age of the Universe. Moreover, they enabled the identification and census of a significant amount of otherwise dust-obscured quasars. Furthermore, studies on weak gravitational lensing helped understanding the effects of the interaction between the spectroscopically observed galaxies and the surrounding dark matter. Investigations of our own Galaxy can also benefit from SDSS spectroscopy. Besides the example of Gänsicke et al. (2009) (mentioned in Section 1.1.6), that identified faint Galactic CVs in SDSS with orbital periods close to the minimum value, Jurić et al. (2008), Ivezić et al. (2008), Bond et al. (2010), Berry et al. (2012), and Loebman et al. (2014) leveraged SDSS spectroscopy to present a set of papers about Galactic tomography. They thus modelled the stellar density distribution, metallicity, and kinematics of our Galaxy, and mapped the Galactic dust extinction and dark matter halo.

1.2.7 LAMOST

Beginning its operations in 2013, the Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST; Cui et al. 2012) is a reflecting Schmidt telescope with a field of view of 5 sq° and two effective aperture set-ups: 3.6 m and 4.9 m. Thanks to its 4,000 fibers, it can observe 4,000 spectra with a single exposure, with a resolution of 1,800 for an 19th r magnitude.

LAMOST survey can be nominally divided in two smaller surveys, based on their distance-scale of interest: the LAMOST Experiment for Galactic Understanding and Exploration (LEGUE) survey of Milky Way stellar structure, and the LAMOST ExtraGAlactic Survey (LEGAS). As its name suggests, LEGUE focuses on the description of Galactic structures by observing the optical spectra of several millions of stars. On the other hand, with the spectroscopic observations of millions of galaxies, LEGAS fosters the study of extragalactic structures and their evolution, as well as the density and distribution of dark matter and energy.

1.2.8 J-PAS

The Javalambre-Physics of the Accelerated Universe Astrophysical Survey (J-PAS) will shortly present itself as a *photo-spectrograph*[5] (Benitez et al., 2014). With an effective field of view of 4.7 sq°, and with its 56 narrow-band optical filters, the 2.5 m telescope will observe an overall sky area of 8,500 sq° above the Javalambre Observatory in approximately five years. Due to its coverage, J-PAS can be listed in the group of very large field astrophysical surveys (i.e., that cover a larger area than 5,000 sq°). This includes, among the others, SDSS and the Vera C. Rubin Observatory (VRO; Thomas et al. 2020), formerly known as the Large Synoptic Survey Telescope (LSST; LSST Science Collaboration et al. 2009).

One of the main purposes for this extremely large cosmological survey is to measure photometric redshifts for galaxies up to $z = 1.3$, with a precision of $0.003(1 + z)$. This is possible thanks to a combination of the large field of view with the multi-wavelength narrow-band photometry (see also Ilbert et al. 2006, Lee & Chary 2020, et similia). J-PASS will deliver low-resolution photo-spectroscopy (R$\sim$ 50), based on the narrow-band multi-filters imaging. The advantage of this technique over standard spectroscopy is that the filters are observationally independent, and hence systematic uncertainties in flux calibration are excluded.

[5] Although it is worth pointing out that the "miniJPAS" precursory survey has already been released (Bonoli et al., 2021).

Thanks to these features, J-PASS will help constraining the Dark Energy equation of state, by measuring the mass of hundreds of thousands of galaxy clusters. Furthermore, thanks to its wide field of view, it will identify and study the varied stellar populations in our Galaxy and in nearby ones, by producing 3D images of the Northern sky.

1.2.9 WEAVE and 4MOST

The William Herschel Telescope (WHT) Enhanced Area Velocity Explorer (WEAVE; Dalton et al. 2012) and the 4-metre Multi-Object Spectroscopic Telescope (4MOST; de Jong et al. 2019) are two forthcoming next generation multi-object spectrographs based at Roque de los Muchachos (Canary Islands, Spain) and at Paranal (Atacama desert, Chile), respectively. With its 1,000 fibers and 20 Integral Fields Units (IFU), WEAVE will provide spectroscopic follow-up observations for *Gaia* and for the Low Frequency Array (LOFAR; Kassim et al. 2000) surveys, in the 370-1000 nm wavelength range, with intermediate (R$\sim 5,000$) and high resolution (R$\sim 20,000$). Fed by its 2,436 fibers, 4MOST will be able to simultaneously obtain spectra for $\sim 2,400$ objects, with a resolution range comparable to WEAVE.

Complementing *Gaia* astrometry and photometry, WEAVE spectra will provide a great contribution in unravelling the chemical content and dynamical status of our Galaxy. On a larger scale, its measurements of galaxy redshifts will promote the study of the properties of dark matter (for instance, by measuring the Baryon Acoustic Oscillations, or through the comparison of the degree of clustering for several types of galaxies). The 4MOST science program is divided into ten surveys, that will be completed in five years. By observing the Southern sky, they will enable Galactic archaeology studies (five surveys), analyses on the formation of BHs and galaxies (four surveys), and of the detection of SN transients and variations of quasar luminosity (one survey, in integration of LSST time domain data).

1.3 Photometry-based Hα-excess selections

The diversity of the various Hα-emitting populations makes the study of emission line sources particularly important to understand the composition and evolution of our Galaxy. However, as stated in Section 1.1, some of the Hα-emitting populations belong to the rarest/faintest kinds, and are therefore hard to identify and further classify. Even more challenging is the classification of those objects that may not display clear Hα-emission line properties, but rather display Hα-excess relative to their parent population. Even though these types

of sources do not present evidence of Hα emission, their intensity in said band might still significantly exceed the ones of the objects they are compared to. An example is provided by the elusive population of low mass transfer rate CVs that have evolved beyond the so-called period minimum (Podsiadlowski et al. 2003; Breedt et al. 2012; Pala et al. 2020). Among the several ways to identify and classify stellar objects, the most telling is a moderate-resolution spectroscopy. However, the collection of spectra requires considerable time and efforts, in particular when attempting to classify large ensembles of targets. Similar issues also affect classification studies based on light curve analysis. In order to alleviate the burden, several pre-selection methods have been developed and applied over the years. A popular approach to pre-select targets is based on optical photometry.

Among the imaging surveys that focused on point sources, Kohoutek & Wehmeyer (1999) obtained a list of ∼4000 point-like Hα emitters located in the northern Galactic plane ($|b| \leq 10°$). Parker et al. (2005), with their Anglo-Australian Observatory/UK Schmidt Telescope (AAO/UKST) SuperCOSMOS Hα Survey (SHS), inspected an area of $\sim 4000\,\mathrm{deg}^2$ in the Southern Milky Way, plus an additional $\sim 700\,\mathrm{deg}^2$ area around the Magellanic Clouds.

1.3.1 Selections based on IPHAS colours

In the context of IPHAS, the main metric for Hα-excess is $r-$Hα. 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 colour-colour diagrams (CCDs) for each IPHAS field, using $r-$Hα and $r-i$, respectively, as vertical and horizontal axes. Each WFC pointing covers an area of $0.22\,\mathrm{deg}^2$ in the sky. Hα line excess source candidates are then selected by iteratively fitting the stellar locus and retaining positive outliers in $r-$Hα. 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 4,853 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% of them did show Hα emission lines.

More recently, Monguió et al. (2020) combined IPHAS and UVEX photometric data to develop the IGAPS (INT Galactic Plane Surveys) catalogue. The g, r and i magnitudes 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)[6]. On the other hand, the astrometric calibration was performed against *Gaia* DR2 astrometric measurements. The catalogue includes ~ 295 million objects. Of these, 53,234,833 (18%) unblended sources with $r < 19.5$ mag were tested for Hα excess. Following a similar strategy to that described in Witham et al. (2008), Monguió et al. (2020) selected almost 21,000 Hα-excess candidates in the Galactic plane.

1.3.2 Population-based vs. position-based selections

Challenges emerge from these past *position-based* (i.e., with respect to nearby objects in the sky) selections, if considering that the r-Hα 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 colour-colour space 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 a 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 CBe phenomenon in the VPHAS+ field of view.

Without any knowledge of the distances, and using only IPHAS measurements, degeneracies may exist in associating a particular object to 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 order to attempt to include these elusive outliers, I developed a novel technique that prioritises a larger representation of the potential Hα-emitting populations, sometimes at the expense of a higher false-positive rate. Taking advantage of the *Gaia* DR2 parallaxes, this method consists in augmenting the classic position-based selection by identifying Hα-excess candidates relative to nearby sources in the calibrated M_G vs. G_{BP} - G_{RP} CMD (*CMD-based* selection). However, it is to be pointed out that the absolute magnitudes in the G band, as well as the colour indices, are calculated here with the assumption of zero extinction, and therefore they should be treated with caution. In fact, without any correction for extinction the sources appear to be fainter and redder than what they intrinsically are, and their location in the CMD results to be shifted along the extinction vector. The effects of extinction are greater for bluer and further objects, and

[6] We point out that U band measurements in the IGAPS catalogue were not uniformly calibrated.

are especially relevant when inspecting crowded areas of the sky, such as the Galactic plane. This contributes to increase the rate of population mixing in the colour-magnitude parameter space, and might lead to the detection of bogus Hα-excess candidates. One way to approach this would be to correct the magnitudes with the use of extinction maps (Sale et al., 2014). Alternatively, the effects of extinction can be accounted for by complementing the CMD-based selection with a position-based one. In fact, nearby objects in the sky, that are found in a restricted distance-range from us, are similarly affected by extinction. Considering the amount of data involved in this work, the latter approach is preferred, here. Moreover, 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 is that their Hα intensity is higher than those of the objects they are compared to.

The next sections of this thesis are organised as follows: Chapter 2 describes the CMD-based (or population-based) identification of Hα-excess sources (Fratta et al., 2021), performed on the *Gaia*/IPHAS catalogue, introduced in Scaringi et al. (2018). In Chapter 3, this selection is validated through a set of spectroscopic follow-up observations of a sub-sample of Hα-excess sources (Fratta et al., 2022). An example of how the Hα-excess selection outlined in Chapter 2 can be relied on to identify accreting WD candidates is presented in Chapter 4. Chapter 5 illustrates how said algorithm can be augmented with the introduction of additional photometric parameter spaces, to pre-select promising Hα-excess candidates in large surveys; finally, a summary of the analyses presented in this thesis is provided.

Chapter 2

Population-based identification of Hα-excess sources

2.1 *Gaia*/IPHAS catalogue

The Hα-excess candidates identified in this work are selected from the catalogue described in Scaringi et al. (2018) (hereinafter, the *Gaia*/IPHAS catalogue). Said catalogue is the outcome of a positional sub-arcsecond cross-match between the sources in *Gaia* and IPHAS DR2 fields of view. It contains a list of approximately 8 million sources that occupy an area of the sky included between $|b| \leq 5^{\circ}$ and $29^{\circ} \leq l \leq 215^{\circ}$, and are mostly located in the northern Galactic plane[1].

2.1.1 Catalogue generation

When performing the cross-match, we 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 (which is included between 2003 and 2012). In an ideal scenario, the recalculation of *Gaia* astrometry for all the sources involved in the cross-matching process would be minimum-uncertainty approach. However, this task is complicated by the substantial proportions of the catalogues being studied. In order to get around this obstacle, we divided the sources in IPHAS into 46 groups, based on the epoch of their observation. Each group included one month of observation epochs. We then rewound *Gaia* astrometry to the mid-epoch of each

[1] All the sources in the catalogue are visible from the northern Terrestrial hemisphere.

monthly batch, which resulted in reasonably small positional uncertainties (0.08 arcsec, for an object with a 2 arcsec yr^{-1} proper motion).

The targets in the *Gaia*/IPHAS catalogue are found mainly within a distance of ~ 1.5 kpc from us. These distances were 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 was justified with the introduction of two parameters that quantify the goodness of *Gaia* astrometric fit and the false-positive rate: $\mathrm{f_c}$ and $\mathrm{f_{FP}}$, respectively. To compute $\mathrm{f_c}$, the targets were binned according to their *Gaia* G band magnitudes; $\mathrm{f_c}$ 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, $\mathrm{f_{FP}}$ reflects the presence of spurious negative parallaxes in *Gaia* measurements, due to poor astrometric fits. To obtain $\mathrm{f_{FP}}$, we first produced a mirror sample of the catalogue, including only objects with negative parallaxes, with `parallax_over_error > -5`. We then binned the objects in the whole catalogue (including the mirror sample) with respect to their G band measurements, and further with respect to the χ^2 of the astrometric fit. Hence, we defined $\mathrm{f_{FP}}$ as the fraction of objects from the mirror sample (false positives) in each bin. Fig. 2.1 and Fig. 2.2 show the retained (top rows) and discarded (bottom rows) targets, if different $\mathrm{f_c}$ and $\mathrm{f_{FP}}$ cuts respectively are applied. These two images correspond to Fig. 9 and Fig. 10 of Scaringi et al. (2018).

The effects on the absolute magnitude in the *Gaia* G band (M_G) introduced by the use of parallax-inversion instead of probabilistic methods (Astraatmadja & Bailer-Jones, 2016) to obtain the distances are generally negligible. In fact, 97.2% 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[2].

2.1.2 Additional quality constraints

The current selection focuses on the subset of targets from the *Gaia*/IPHAS catalogue that pass strict quality criteria, in order to minimise the inclusion of spurious cross-matches. Some quality control selection cuts have already been applied during the compilation of the

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

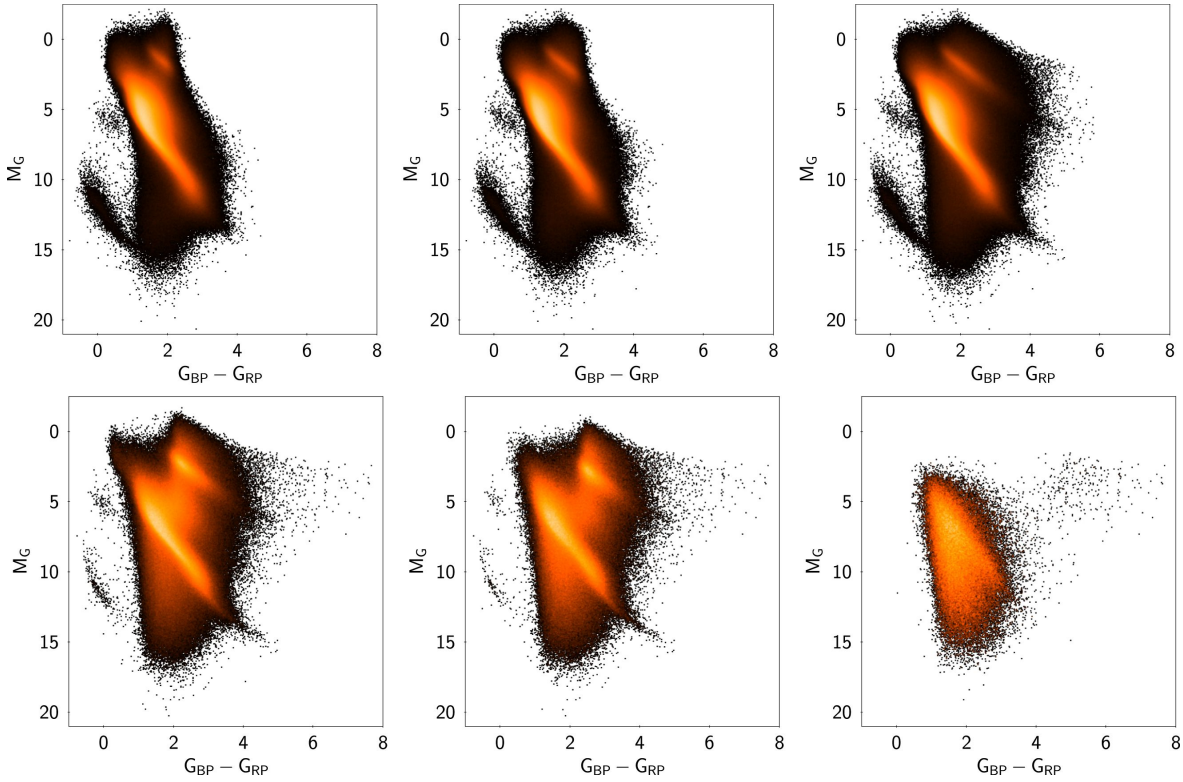


Figure 2.1: Location in the *Gaia* CMD of the targets in the *Gaia*/IPHAS catalogue, after the application of the $f_c < 0.8$, $f_c < 0.9$ and $f_c < 0.99$ "completeness" cuts (left-hand, middle, and right-hand columns, respectively). The top row shows the objects that satisfy those cuts, whilst the discarded objects are displayed in the bottom row. This image is taken from Fig. 9 of Scaringi et al. (2018).

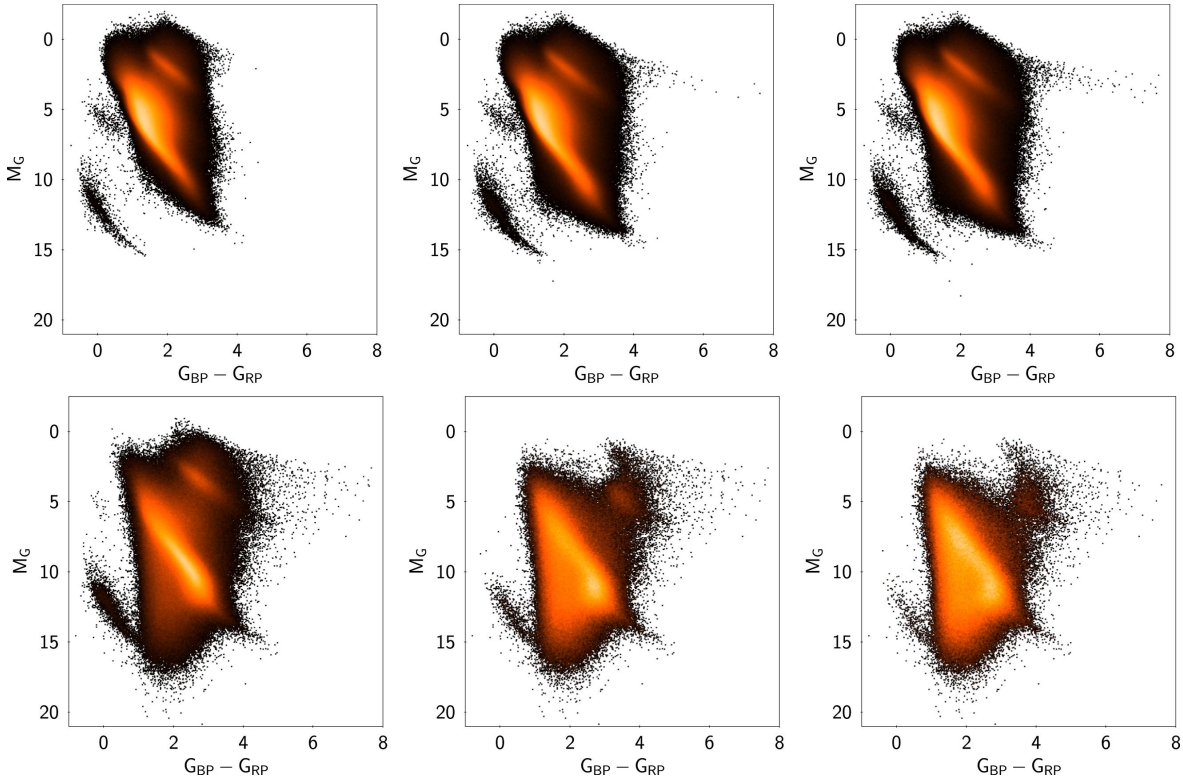


Figure 2.2: Location in the *Gaia* CMD of the targets in the *Gaia*/IPHAS catalogue, after the application of the $f_{FP}= 0$, $f_{FP}\leq 0.01$ and $f_{FP}\leq 0.02$ "false positive" cuts (left-hand, middle, and right-hand columns, respectively). The top row shows the objects that satisfy those cuts, whilst the discarded objects are displayed in the bottom row. This image is taken from Fig. 10 of Scaringi et al. (2018).

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:

1. remove sources with low-quality astrometric fits and/or high false-positive probabilities (see Sec. 2.1);
2. remove targets close to the saturation limit of IPHAS;
3. only retain targets for which we have a valid measurement in each band of interest (r, i, Hα, M_G, G_{BP} and G_{RP}).

The following cuts are thus applied:

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

These cuts yield 7,474,835 sources out of the original 7,927,224. Fig. 2.3 shows the *Gaia* CMD and IPHAS CCD with the targets that pass the additional quality cuts.

2.2 Partitioning processes

The identification of Hα-excess candidates is achieved here by selecting "positive outliers" in the $r-$Hα vs. $r-i$ colour-colour space. To mitigate selection biases due to stellar population mixing and Galactic extinction, the sources in the master-catalogue are first partitioned with respect to their positions in *Gaia* CMD and in the Galactic coordinate space, respectively. The two $r-$Hα outliers selections, performed on the CMD-based and on the position-based (or also "coordinates-based") partitions, are independent and complementary. The selection strategy applied on the position-based partitions ($POS-partitions$) hinges on the one applied by Witham et al. (2008). I point out that the CMD-based selection can still be improved, since some populations may overlap in the colour-magnitude space. Other techniques that use more novel machine learning approaches could be employed to select Hα-excess sources. The 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,

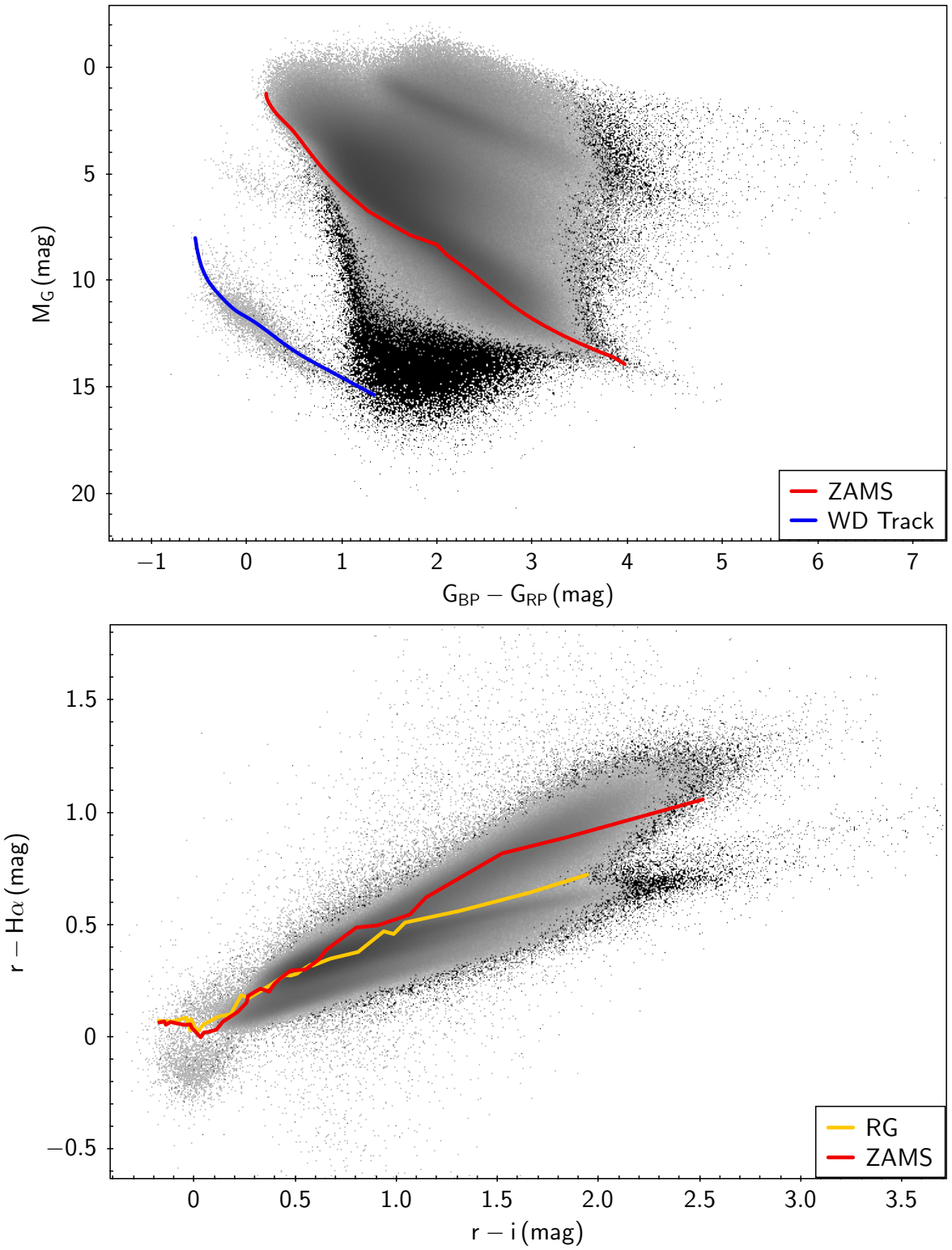


Figure 2.3: Positions in the *Gaia* M_G vs. G_{BP}-G_{RP} CMD (top panel) and in the IPHAS $r-$Hα vs. $r-i$ CCD (bottom panel) of the sources in the *Gaia*/IPHAS catalogue (Scaringi et al., 2018). The grey dots represent the objects that satisfy the quality constraints described in Sec. 2.1.2, while the targets that do not pass this first selection are displayed with the black dots. The red and the blue 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 orange line in the bottom panel (both taken from Drew et al., 2005) depict respectively the synthetic ZAMS track for zero reddening and the synthetic RG track, in this parameter space.

the approach used here allows to examine and understand the underlying population used to infer the Hα-excess significance values.

2.2.1 CMD-based partitions

Using the calibrated *Gaia* CMD shown in the top panel in Fig. 2.3, 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 colour-magnitude space to make the underlying source population as homogeneous as possible. To balance these two requisites, an iterative method is applied. First, a fine grid of 840×840 equally spaced "elemental" cells is generated, covering the whole CMD (each elemental cell with dimensions $l_{\rm x} \sim 0.007$ mag and $l_{\rm y} \sim 0.024$ mag, respectively). No elemental cells contain enough objects to be considered as a partition. The side lengths of the grid cells are then increased to the next integer divisor of 840, in units of $\rm l_x$ and $\rm l_y$, respectively. The second iteration produces 420×420 cells, 4,852 of which satisfy the criteria to become partitions (these belong to the densest regions of the CMD). 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 4,851. The iterations carry on for all the integer divisors of 840, between 2 and 60, with the following additional caveats: each partition must not be completely surrounded by another, bigger, partition; all the elemental cells within each partition must be contiguous. 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. 9,181 $CMD-partitions$ result from this process, with a maximum density of 1,514 sources per partition. The map of the resulting partitions in the CMD is shown in Fig. 2.4 To account for the uncertainty on the positions of the objects in the CMD, this partitioning process is repeated upon change of the side lengths of the elemental cells. As an example, a 20% increase of $\rm l_x$ and $\rm l_y$ produces a 0.8% variation in the number of selected outliers, meaning that this selection is independent (to a reasonable extent) on the size of the elemental cells.

2.2.2 Position-based partitions

For this different partitioning algorithm, an evenly spaced grid in the b vs. l space is created. The size of each cell is $1.205^\circ \times 1.004^\circ$, 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

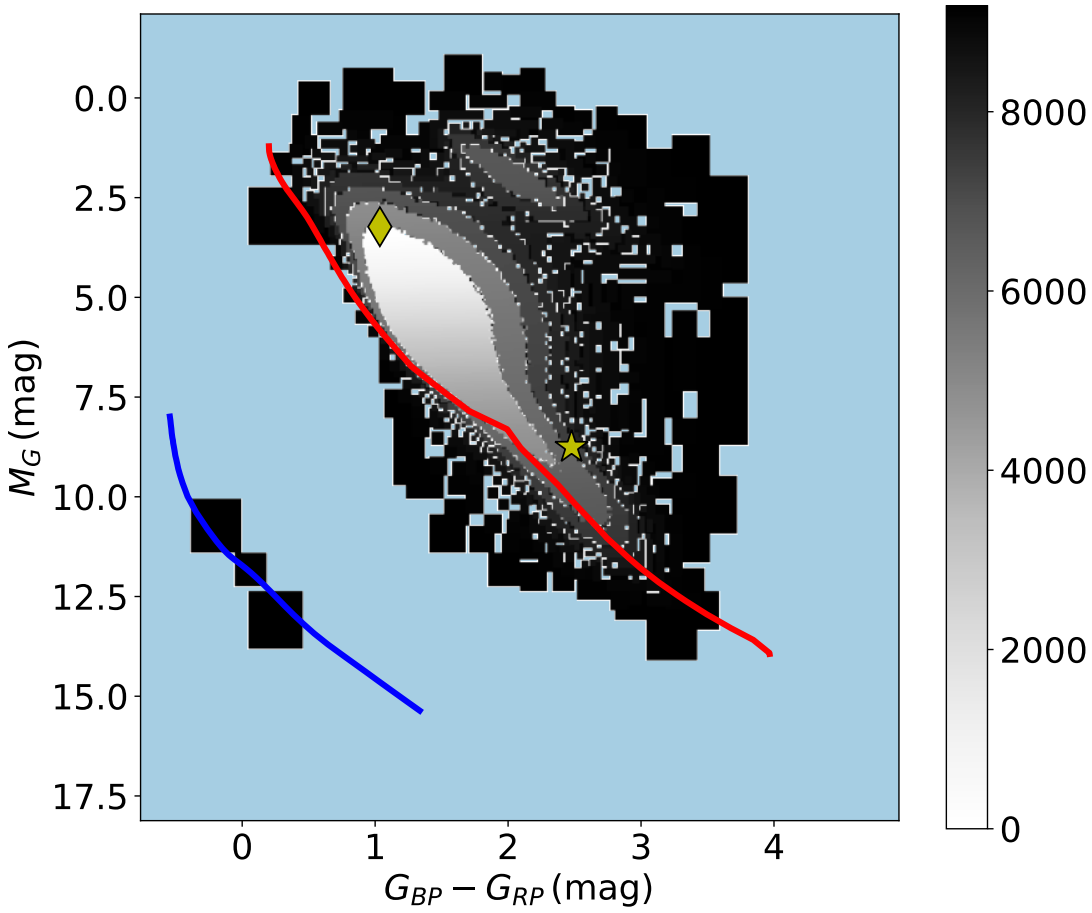


Figure 2.4: Map of the partitions in *Gaia* CMD, before the "leftovers" assignment. The colour code refers to the order in which the partitions were created (no partitions are assigned to the light blue background). The area covered by the single partition increases where the density of sources decreases. The red line represents the synthetic ZAMS track (Bressan et al., 2012), while the blue 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 2.3.

results in 1,674 POS-partitions, with a maximum density of 12,604, and a minimum density of 546 objects per partition.

2.3 Detrending and outliers identification

The Hα line excess candidates are identified by inspecting the $r-$Hα vs. $r-i$ CCDs. The method for selecting outliers from each partition is divided into two steps. First, the main stellar locus is found by iteratively fitting a line to the data, and applying Chauvenet criterion to identify the outliers. These are removed from the data at the end of each iteration. Chauvenet criterion consists of calculating a threshold beyond which only outliers are expected to be found. Assuming a Gaussian distribution, this threshold is proportional to the root mean square of the distribution, and inversely proportional the number of objects that constitute it. In theory, in order to tackle the population-mixing issue, the fit should be forced to the upper branch in the CCD (as done by Witham et al., 2008) by removing only the negative outliers. However, for most of the partitions the resulting best-fit line does not deviate significantly from the model obtained by the direct application of the unmodified Chauvenet criterion. This criterion is preferred in this step to a sigma-clipping procedure, as the latter does not take into account the variable number of objects included in the different partitions.

Once the stellar locus has been located, it is used as a baseline to identify the outliers: each CCD is detrended by subtracting the corresponding linear model from the data. A second iterative application of Chauvenet criterion on the detrended CCD enables us to isolate the outliers, and hence to calculate the root mean square (rms) of the remaining sources. The objects that satisfy the following relation are selected as Hα-excess candidates, from either the CMD-based and/or the position-based partitions:

$$n_\sigma = \frac{y}{\sqrt{(\delta y)^2 + (m_{fit} \cdot \delta x)^2 + rms^2}} \geq 3. \tag{2.3.1}$$

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 $r-i$ colour, and m_{fit} is the slope of the best-fit line. Thus defined, n_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 actually assigned to each source: $CMD-significance$ ($\mathrm{n}_{\sigma_{\mathrm{CMD}}}$) and $POS-significance$ ($\mathrm{n}_{\sigma_{\mathrm{POS}}}$). Objects that satisfy relation 2.3.1, either from the CMD-based and/or from the position-based selection, will henceforth be referred to as "3σ

outliers". Fig. 2.5 provides a graphical depiction of the detrending (top panel) and outliers selection (bottom-left panel) processes relative to the CMD-partition 7331, as an example of a well-behaved partition.

I point out that Chauvenet criterion assumes an underlying Gaussian population, while a non-negligible amount of partitions seems to deviate from this (mainly due to unavoidable population-mixing). However, the application of Chauvenet criterion on non-Gaussian partitions provides a more robust outlier selection, since the standard deviation of these partitions is overestimated. As can be noticed from the bottom-right panel of Fig. 2.5, the detrended $r-\mathrm{H}\alpha$ distribution relative to the CMD-partition 7331 constitutes a good example of Gaussian underlying population. On the other hand, Fig. 2.6 presents an example of partition (CMD-partition 0, specifically) in which the underlying distribution deviates from a standard Gaussian behaviour. As a reference, Fig. 2.7 presents the CCDs (before and after the detrending process), and the histogram of the detrended $r-\mathrm{H}\alpha$ values relative to POS-partition 154. Two trends are identifiable from these latter CCDs: the top one represents the locus in which unreddened MS stars lie, while the bottom trend corresponds to the reddened RG track. This bifurcation, generally more prominent in POS-partitions than in CMD-partitions, emphasises that, by simply grouping together nearby objects in the sky, population-mixing effects are mostly unavoidable. The two trends that characterise the CCD of this partition reflect in the bimodality recognisable in the histogram in the bottom-right panel. However, this effect does not alter the number of outliers selected from partitions that present it, since the second Gaussian population is always redder than the main one.

The presented algorithm selects both positive and negative outliers; however, since the goal is to identify the Hα-*excess* candidates, the term "outliers" will henceforth refer to only the positive ones.

2.4 Selection outcome

Our selection identifies 28,496 $r-\mathrm{H}\alpha$ 3σ outliers (0.36% of the total dataset) above the previously identified stellar loci. More specifically, 25,030 outliers are selected from CMD-partitions and 8,550 from POS-partitions. In Fig. 2.8, the locations in the *Gaia* CMD and in the IPHAS CCD of these outliers are presented. It appears particularly noticable in the bottom panel that many of these candidates would have not stood out as outliers, if the chosen statistical analysis had been applied directly in the colour-colour domain. Mainly five

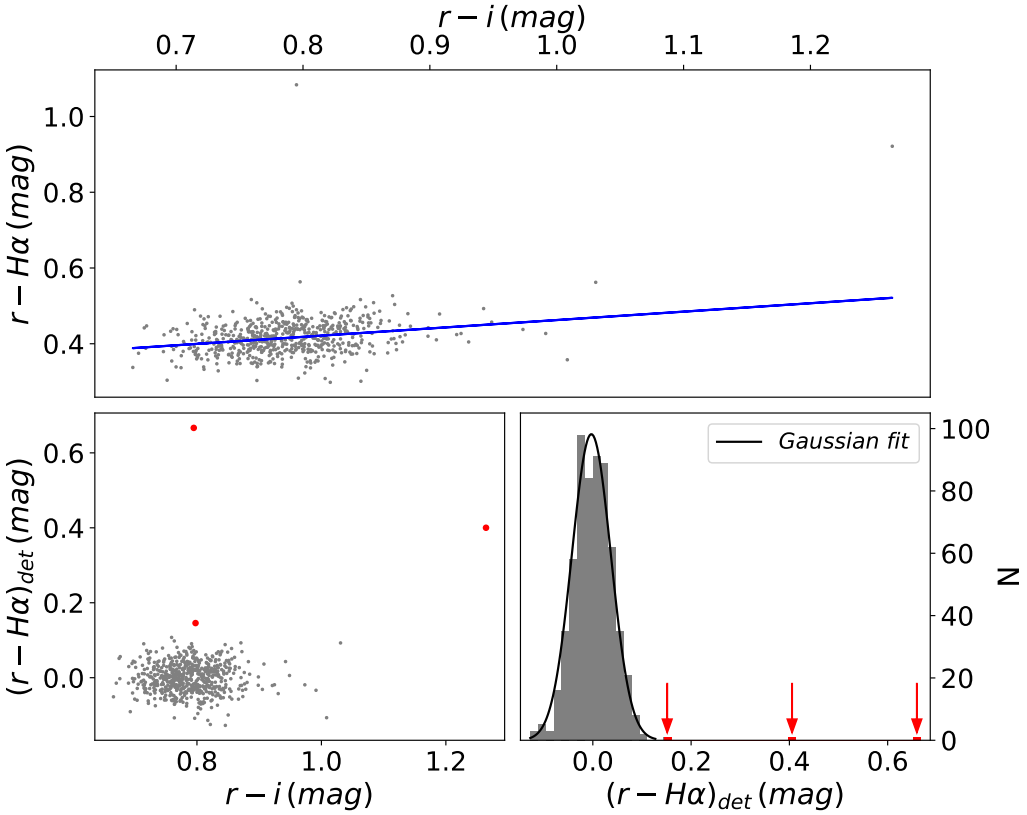


Figure 2.5: Graphical depiction of the detrending and outliers selection processes performed on CMD-partition 7331. The top panel shows the non-detrended IPHAS CCD for this partition, in which only one linear trend is clearly visible. The blue line depicts the best-fit linear model. It was subtracted from the data points to obtain the detrended r-Hα parameter (bottom-left panel). The red dots represent the positive outliers of the distribution. The bottom-right panel shows the detrended r-Hα distribution of this partition. The Gaussian behaviour of the underlying population is well described by the best-fit model (the black solid line). The red arrows point to the three outliers of the distribution. The position of the CMD-partition 7331 in *Gaia* CMD is shown in Fig. 2.4.

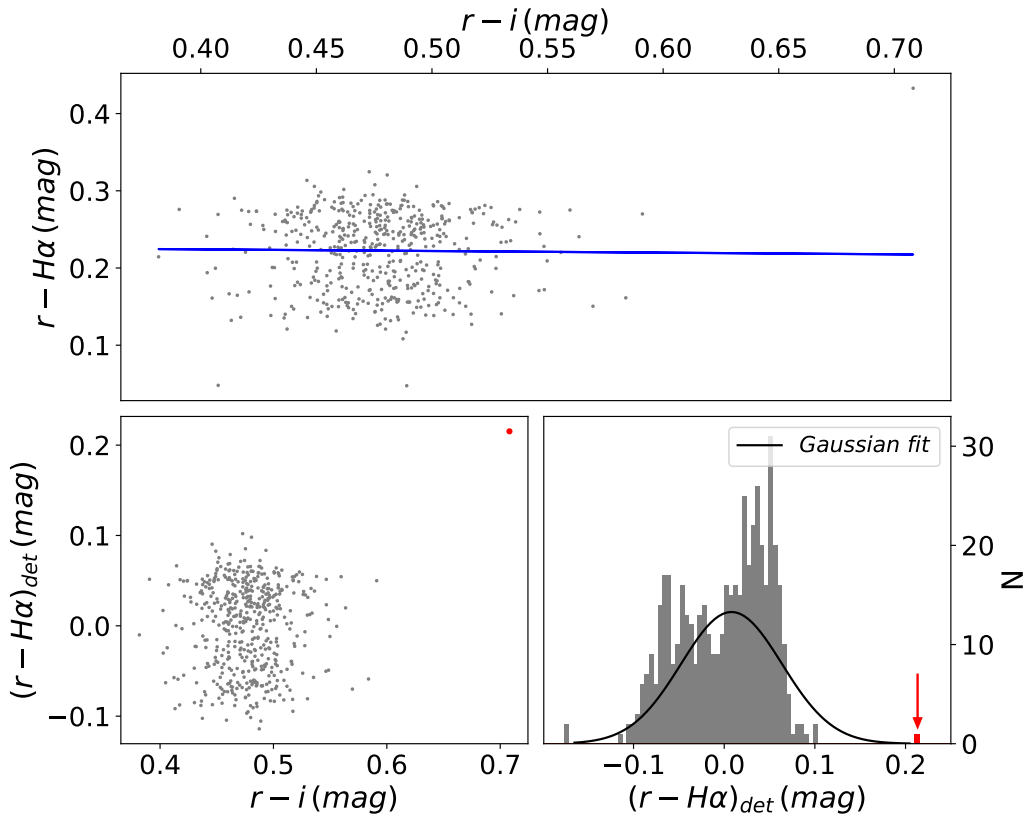


Figure 2.6: Graphical depiction of the detrending and outliers selection processes performed on CMD-partition 0. As it stands out clearly, the Gaussian model is not a good fit to the underlying population. The position of partition 0 in the *Gaia* CMD is shown in Fig. 2.4.

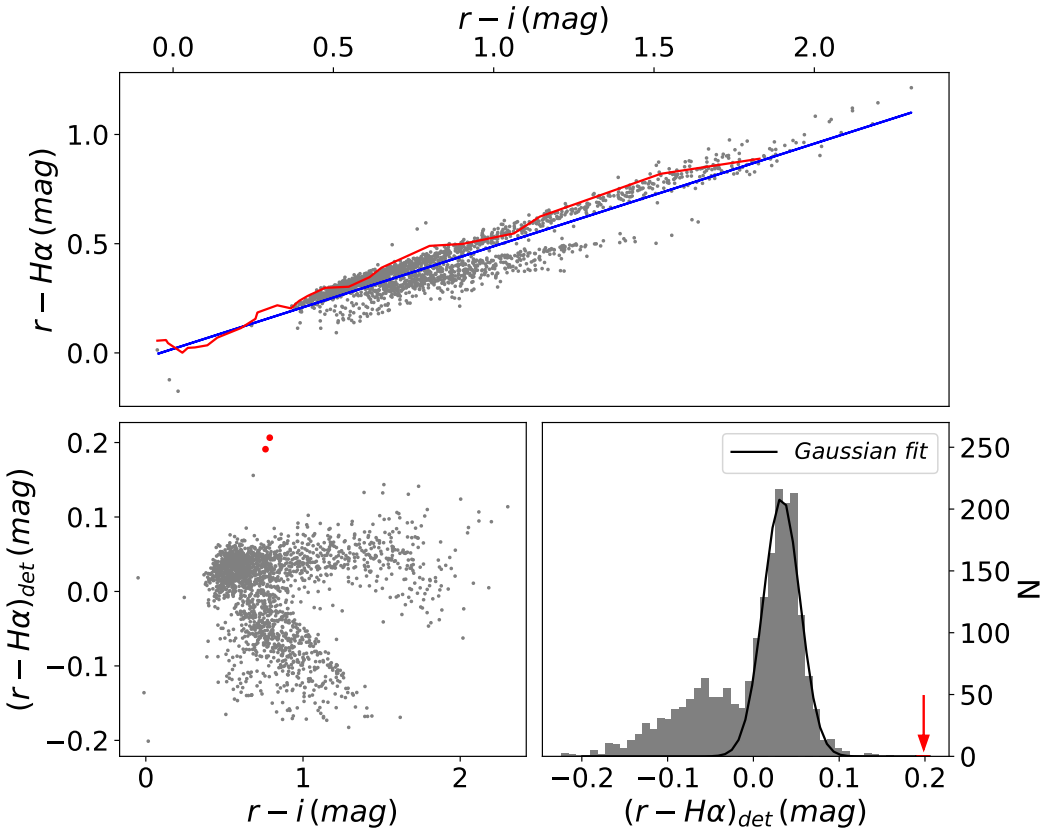


Figure 2.7: Graphical depiction of the detrending and outliers selection processes performed on POS-partition 154. The red line in the top panel represents the synthetic ZAMS track, for zero reddening (Drew et al., 2005).

regions of the CMD (identified with the letters from A to E) with a particularly high density of outliers can be highlighted from the top panel: the region where CBe stars sit (A), the YSOs region (C), the reddened RG track (D), the M-dwarfs area (E), and the region in the bluer side of the ZAMS (B), especially on the WD track. Region B is generally associated to CVs.

Fig. 2.9 displays the map of the fraction of outliers per CMD-partition. Above all, two regions with a relatively high fraction of outliers stand out. These regions are: the area commonly associated to YSOs, i.e. to the redder side of the ZAMS (the red line) and centred around $M_G = 7.5$ mag; and the region where CBe stars lie (i.e. the brightest area in the CMD, centred around $G_{BP} - G_{RP} = 1.2$ mag). Further information about the statistical composition of the Hα-excess candidates that occupy these areas of the CMD can be obtained through a cross-match[3] with SIMBAD database (Wenger et al., 2000). A cross-matching radius of 1 arcsecond is chosen, considering that the coordinates provided in SIMBAD are the most precise and accurate available coordinates (usually with uncertainties that are consistent with the chosen radius). 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 CBe stars (or candidates), 34 are emission-line stars, 24 are classified as "Star" and 10 are RGs.

In Fig. 2.10, the map of the fraction of outliers per POS-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. Moreover, the distribution of this ratio in the Galactic coordinate space is consistent with being homogeneous, with no significant trend in either direction. Nonetheless, some areas in the b vs. l diagram with a relatively high density of Hα-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 $l = 99.30\,^\circ$, $b = 03.74\,^\circ$; Kharchenko et al., 2013) and NGC2264 (centred at $l = 202.94\,^\circ$, $b = 02.30\,^\circ$; Dias et al., 2014, Kuhn et al., 2019, Barentsen et al., 2013), which are easily identifiable in Fig. 2.10. The latter star-forming region has been the subject of previous studies, such as the one presented by Barentsen et al. (2013). Through the application of the Bayesian inference

[3] All the cross-matches for this thesis have been produced with the software "TOPCAT".

[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.

method, they identified 115 accreting objects in NGC2264. POS-partition 1289 (highlighted by the yellow circle in Fig. 2.10), corresponds to the sky area in which NGC2264 is located, and is in fact the position-partition with the highest fraction of Hα-excess candidates: out of 2,826 objects, our algorithm selects 71 outliers (2.5%). Nine POS-partitions centred around this partition are shown in Fig. 2.11. The apparently empty areas in the sky are due to the quality cuts applied when compiling IPHAS DR2; because of these cuts, IPHAS DR2 provides photometric measurements for sources covering 92% 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 choosing different confidence levels during the selection process. One possibility consists in considering as outliers all the objects selected using Chauvenet criterion: this would be ideal if all partitions were to display Gaussian distributions in the r-Hα direction. Alternatively, all the objects that satisfy the $n_\sigma \geq 3$ constraint (Eq. 2.3.1) can be flagged as outliers; this constitutes a more relaxed threshold, when compared to Chauvenet 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 they are also not equally populated. The suggested threshold is the 5σ one ($n_\sigma \geq 5$), as a compromise to reduce false-positives fraction, while retaining a robust candidate selection. In fact, all candidates selected using a 5σ threshold would have been included using Chauvenet criterion as well. By applying the 5σ cut, 6,774 outliers (0.09% of the complete dataset, 23.8% if compared to the 3σ sample) are identified: 6,455 come from the CMD-partitions and 2,209 from the POS-partitions. For all the sources in the master-catalogue, two `flagCMD` 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 2.1. The full set of metrics computed during the catalogue generation is also published. In this full version, the necessary pieces of information are provided to trace each source back to the corresponding CMD-partition and POS-partition, as well as the detrending model information for each CCD. This additional information can aid the users of the catalogue to further tune

the selection of Hα-excess candidates to suit a specific task.

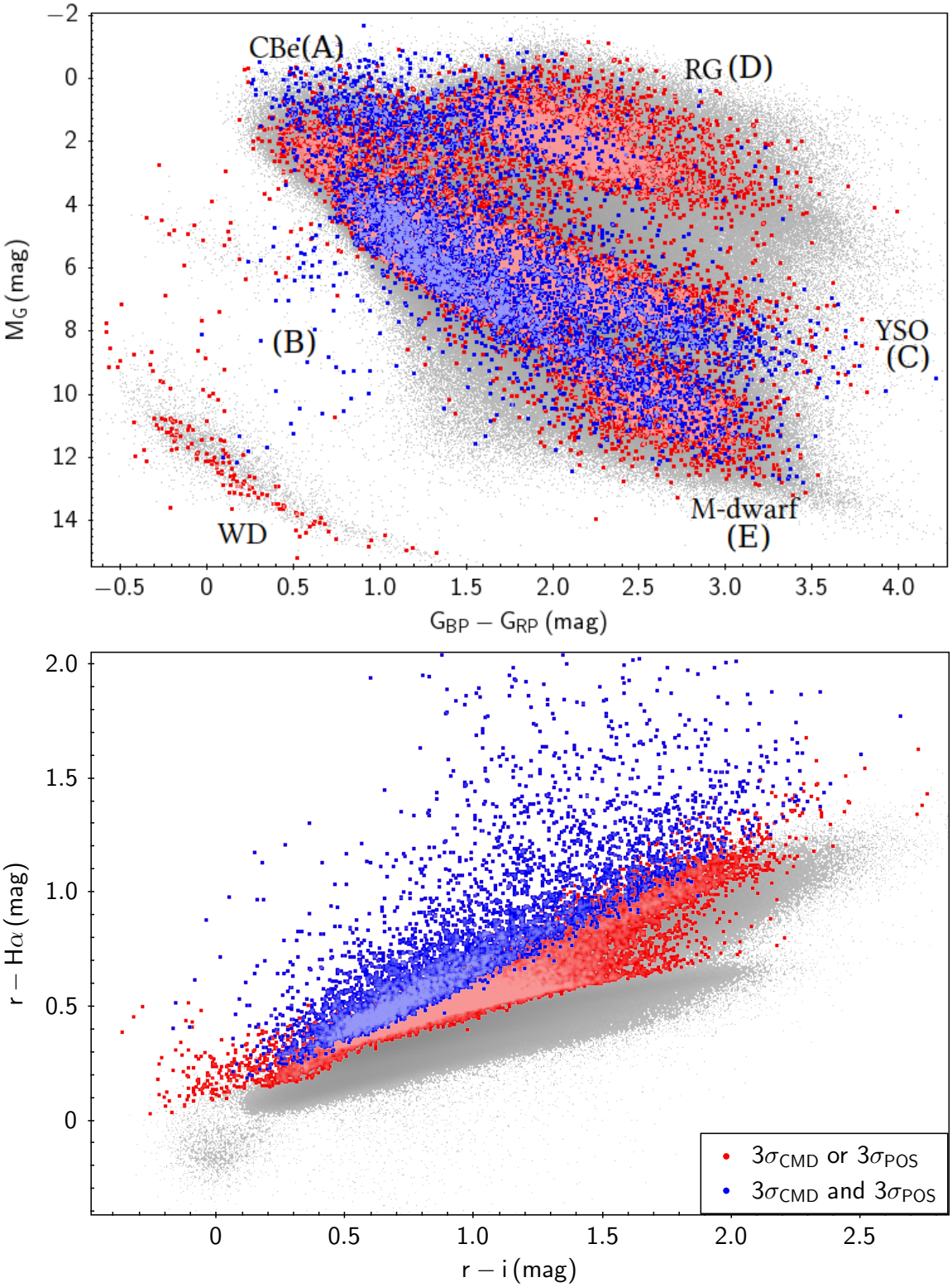


Figure 2.8: The top panel shows the position of the 3σ outliers in *Gaia* CMD, while their layout in IPHAS CCD is presented in the bottom panel. The red dots represent all the 3σ outliers identified by either our CMD-based or position-based selection, while the blue dots represent the subset of 5,084 outliers selected from both the partition types. The intensity of both these colours scales inversely with the density of objects.

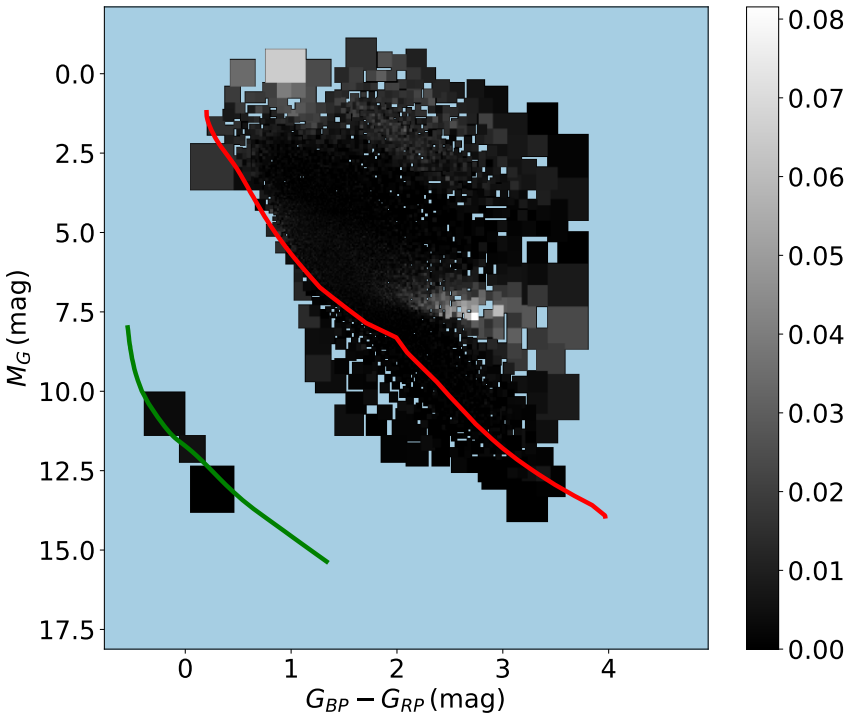


Figure 2.9: Fraction of outliers per CMD-partition.

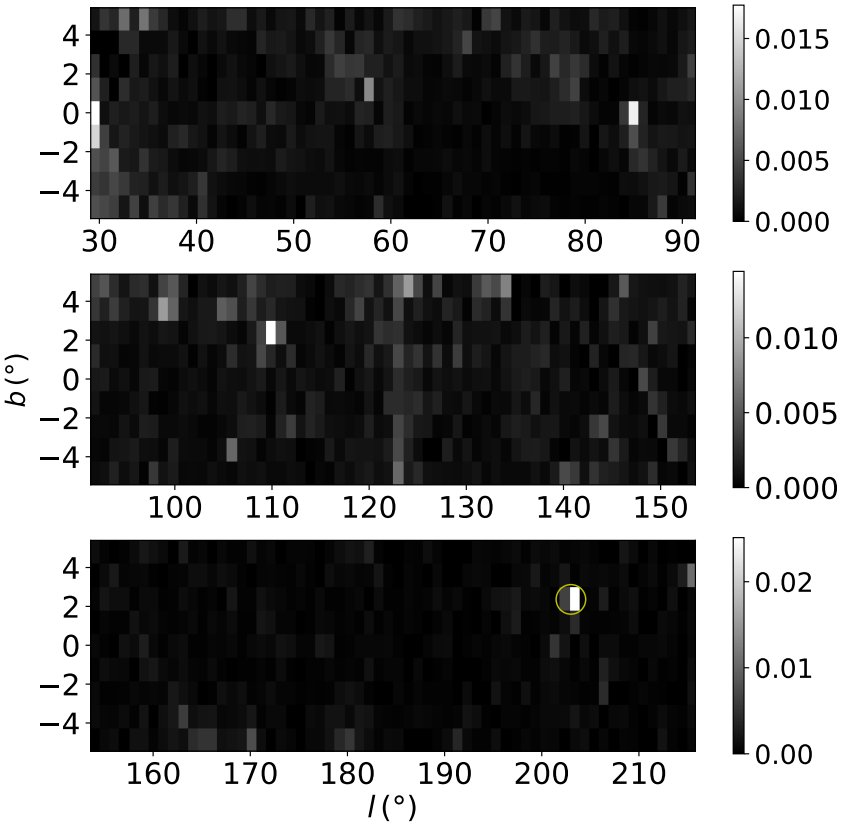


Figure 2.10: Fraction of outliers per POS-partition. POS-partition 1289 is highlighted by the yellow circle.

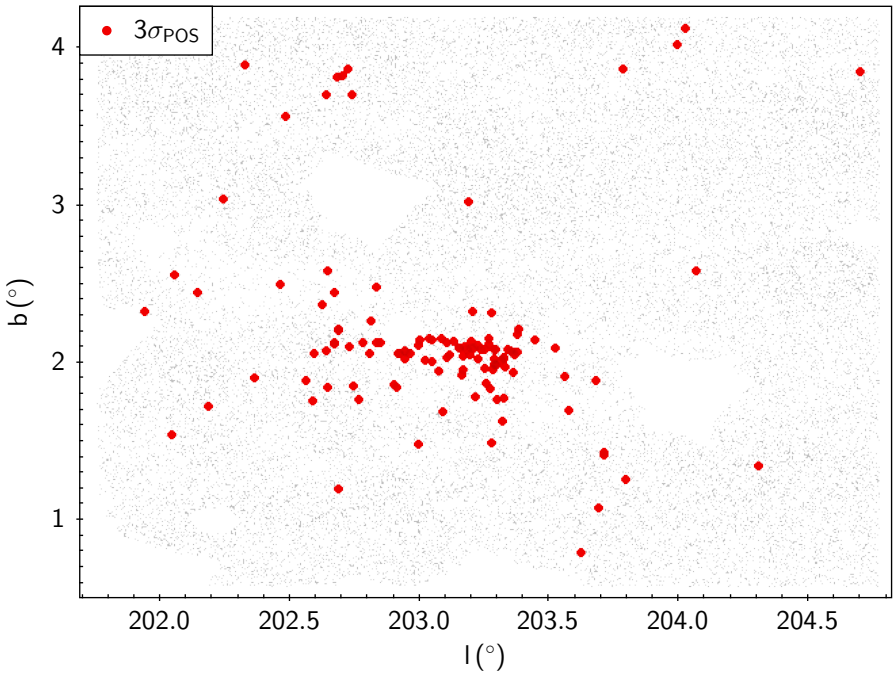


Figure 2.11: Graphical depiction of the nine POS-partitions centred around POS-partition 1289, in the Galactic coordinates space. This latter partition is the one with the highest fraction of Hα-excess candidates (the red dots).

SourceID (*Gaia DR2*)	**RA** (°)	**Dec** (°)	**Dist** (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 CMD**	**flag POS**
429950213735495552	0.01155	62.53021	4.91	13.217	0.001	12.919	0.002	12.963	0.001	13.418	12.938	-0.196	0	1
430049096757310848	0.04097	62.80630	0.25	17.815	0.017	17.868	0.026	17.663	0.018	17.612	17.808	10.816	0	1
432344808313438336	0.04268	66.34893	0.73	18.881	0.016	17.163	0.014	17.851	0.019	20.173	16.946	9.068	1	1
432167168466312448	0.05005	65.17739	0.81	18.114	0.009	16.901	0.012	17.378	0.013	19.016	16.783	8.374	0	1
429950179375766144	0.06964	62.53050	0.44	13.017	0.001	12.419	0.001	12.593	0.001	13.477	12.461	4.819	2	1
429952687636601728	0.09328	62.64748	0.33	13.639	0.001	13.100	0.002	13.276	0.002	14.408	13.114	6.252	0	1
429521606051123712	0.11978	61.69365	2.61	16.222	0.003	15.723	0.006	15.738	0.006	16.619	15.392	4.000	1	2
432168650239229696	0.12183	65.20698	0.50	13.080	0.001	12.362	0.001	12.617	0.001	13.615	12.402	4.591	2	0
429915231214176768	0.12447	62.17243	1.19	17.926	0.010	16.656	0.011	17.262	0.015	18.942	16.643	7.368	1	0
430048723107383808	0.12939	62.77123	1.20	15.160	0.004	14.507	0.003	14.679	0.003	15.498	14.192	4.520	2	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 2.1: The table shows the first ten rows of our *meta-catalogue*. 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 as suggested in Scaringi et al., 2018); the IPHAS DR2 photometric measurements with corresponding errors; the *Gaia* DR2 photometric measurements; the two labels `flagCMD` and `flagPOS`. These two latter parameters express how likely each source is to be an outlier, within its corresponding $r-$Hα distribution (either in the CMD-partitions and/or in the POS-partitions).

2.5 Outliers overview

In Fig. 2.12 the positions of our 3σ outliers in *Gaia* CMD (left column) and in IPHAS CCD (right column) are shown. The most evident differences between the selections applied on the CMD-partitions and on the POS-partitions are:

1. The CMD-based selection is less efficient, compared to the position-based one, in identifying outliers along the WD track. This effect stands out clearly from the comparisons of both the CCDs and the CMDs. This is due to the constraints set when partitioning *Gaia* CMD: the partitions in the least populated regions of this parameter space must 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 in the surrounding area of the CMD. This results in a limited number of outliers.

2. The CMD-based selection identifies more Hα-excess candidates in the region of the CMD where the CBe stars are, compared to the position-based algorithm.

3. Most M-dwarf Hα-excess candidates are identified mainly by the position-based selection. We believe that this is related to the fact that active M-dwarfs have an intrinsically more intense $r-$Hα IPHAS colour, compared to other MS stars. In contrast with the CMD-partitions, these objects are blended with other populations in the POS-partitions, and hence they stand out more clearly as outliers.

4. The CMD-based selection is more efficient in identifying YSOs of various types. This can be observed in the top CMD in Fig. 2.12 as the cluster of Hα-excess candidates on the redder side of the zero age Main Sequence (ZAMS) track, centred at around $M_G{\sim}7.5$ mag. These systems would be difficult to identify from POS-partitions, unless they display strong Hα emission.

A cross-match between our outliers sample and SIMBAD database provides an indicative statistical representation of the populations to which our Hα-excess candidates belong. As for the preliminary analysis presented in Section 2.4, a cross-matching radius of 1 arcsecond with SIMBAD database is chosen; out of the 1,825 matches, 866 are classified as YSOs[5] (or candidate YSOs), 376 sources are classified with the generic label "Star", 233 are emission-line stars, 113 are classified as CBe stars (or candidates), 13 as WDs (plus 43 WD candidates), and

[5]This classification also includes T-Tauri stars and Orion Variable stars.

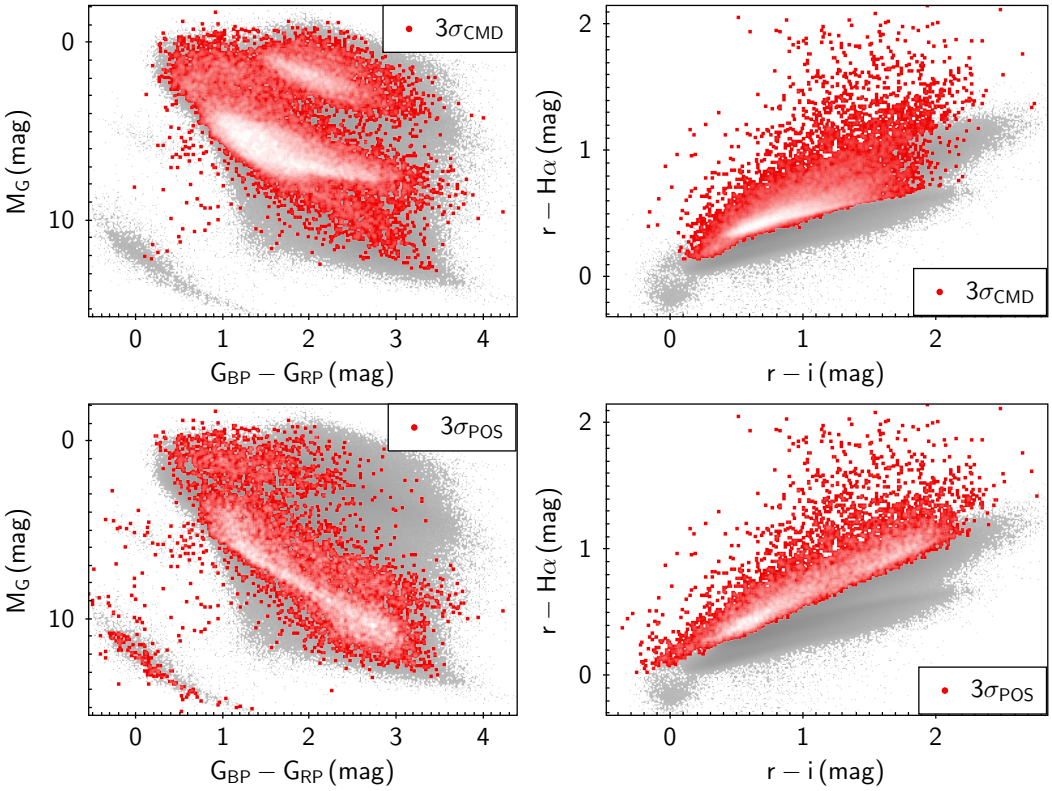


Figure 2.12: The top row presents the 3σ outliers (the red dots) identified from CMD-partitions, while the ones obtained from POS-partitions are shown in the bottom row.

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 of the 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 column in Fig. 2.13 shows the r magnitude distributions of our 5σ outliers. The bin size for these histograms is 0.2 mag. Both distributions are clearly bimodal; the peak around 13.1 mag for the CMD-outliers, as well as the one around $r = 13.6$ mag for the POS-outliers, is to be partially imputed to an observational bias. The secondary mode for the $5\sigma_{\rm CMD}$ 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 $5\sigma_{\rm POS}$ outliers is 18.2 mag (i.e. more than 1.5 magnitude fainter than the mode of the whole data set), confirming the fact that, generally speaking, our position-based selection is more efficient in identifying fainter outliers than the CMD-based selection. The blue areas in the histograms represent the most populated bins around $r = 13$ mag (for the CMD-outliers) and $r = 13.5$ mag (for the POS-outliers), while the red areas indicate the three most populated

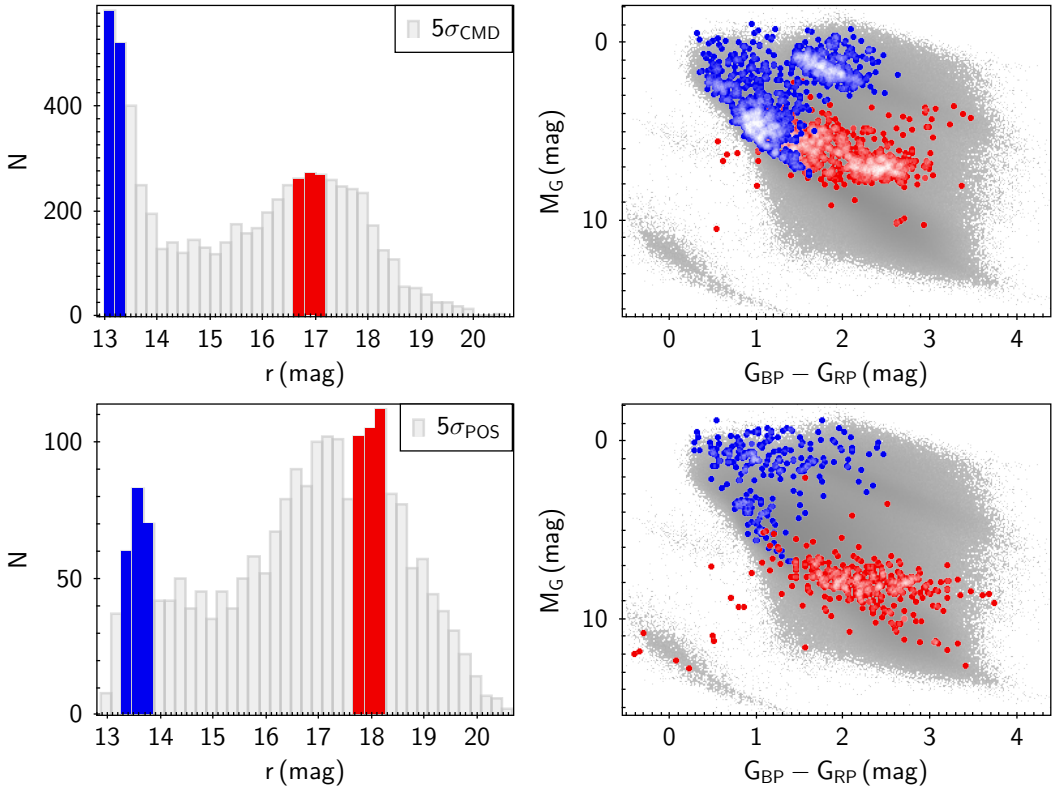


Figure 2.13: r magnitude distributions (left column) of the $5\sigma_{\mathrm{CMD}}$ outliers (top row) and of the $5\sigma_{\mathrm{POS}}$ outliers (bottom row), with a bin width of 0.2 mag. Both the distributions are bimodal; the blue areas in the histograms point to the bins around the brightest of the two modes, respectively, while the red areas in the histograms display the three most populated bins around the secondary modes. In the right column, the density in *Gaia* CMD of the objects that occupy the areas around the modes in the corresponding histogram are shown.

bins around the respective secondary modes. In the right column of the same figure, the CMD densities of the sources belonging to these coloured regions are shown. The brightest outliers in both distributions are located in the region of the CMD commonly associated to CBe stars. The CMD-outliers belonging to the red bins mainly cluster in the region of the CMD where the YSOs generally lie. Also the POS-outliers with an r magnitude close to the secondary mode mainly occupy the region of the CMD associated to 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 active M-dwarfs region.

2.6 Comparison with similar selections

A preliminary validation of our selection can be obtained by comparing our results with similar previous studies. In the following subsections, our results are compared with the catalogue developed by Witham et al., 2008 and with the IGAPS catalogue.

2.6.1 Comparison with Witham's catalogue

The cross-match between the list of emitters in Witham et al., 2008 (which counts 4,853 objects) and the *Gaia*/IPHAS catalogue (after the application of the quality cuts described in Sec. 2.1.2), yields 1,213 common sources (25.0% of Witham et al., 2008 outliers sample). This cross-match is performed using a radius of 1 arcsecond, since the IPHAS pre-calibration catalogue (analysed by Witham) has a comparable spatial resolution (Drew et al., 2005). Despite the proper motion correction applied during the production of the *Gaia*/IPHAS catalogue, sources with particularly large proper motions might be the cause of spurious matches, at this stage; however, the number of matches does not change significantly if this parameter is increased up to a generous 5 arcseconds (32 additional matches result from the use of this larger cross-matching radius). Although the remaining 75% of Witham's outliers can be found in *Gaia* DR2 archive, their astrometric/photometric measurements did not satisfy the f_C and f_{FP} quality constraints described in Scaringi et al. (2018).

Of the 1,213 common targets, 1,115 (91.9%) are also identified as outliers by our algorithm, with a significance (either $n_{\sigma_{CMD}}$ and/or $n_{\sigma_{POS}}$) equal to or greater than 3. By comparing the common sources with our CMD-based selection, 1,053 of common outliers (94.4%) are recovered, whilst the position-based selection finds 1054 (94.5%) of them (i.e. 992 common outliers are identified by both our selection criteria). A subset of 933 out of 1,115 common objects is characterised by $n_{\sigma_{CMD}} \geq 5$ and/or $n_{\sigma_{POS}} \geq 5$; 893 of them have a $n_{\sigma_{CMD}} \geq 5$, while 671 of them have $n_{\sigma_{POS}} \geq 5$ (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. 2.14, the positions in *Gaia* CMD of the results of the cross-match with Witham's catalogue are presented. More specifically, the top panel shows the matches between Witham's list and our 3σ CMD-outliers, while the bottom panel shows an analogous diagram for our POS-outliers. This comparison confirms that the CMD-based selection is more efficient than the position-based one in recovering bright objects, such as the ones that lie around the area of the CMD commonly associated to CBe stars. On the other hand, the cross-match with our position-based selection yields more matches in the active M-dwarf region of the CMD, and on the WD track.

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, it is worth pointing out that this apparent variability might be due to

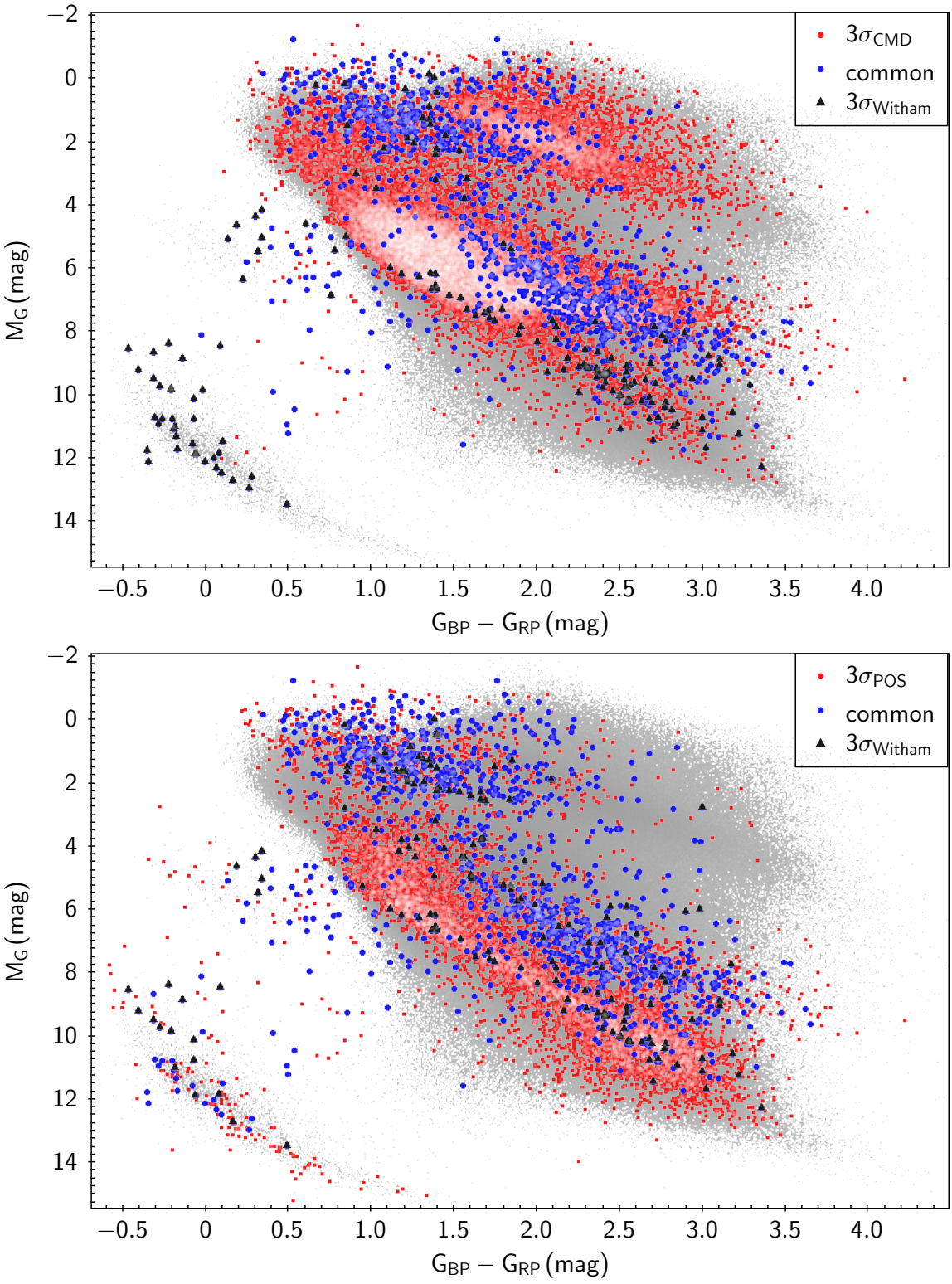


Figure 2.14: Top panel: position in the CMD of the matches between the outliers in Witham et al. (2008) and our 3σ CMD-outliers. Bottom panel: position in the CMD of the matches between the outliers in Witham et al. (2008) and our POS-outliers. The red dots represent the totality of our CMD/POS-outliers; the blue dots are the common outliers between Witham's list and our CMD/POS selection; the black dots are Witham's outliers not selected by our algorithm.

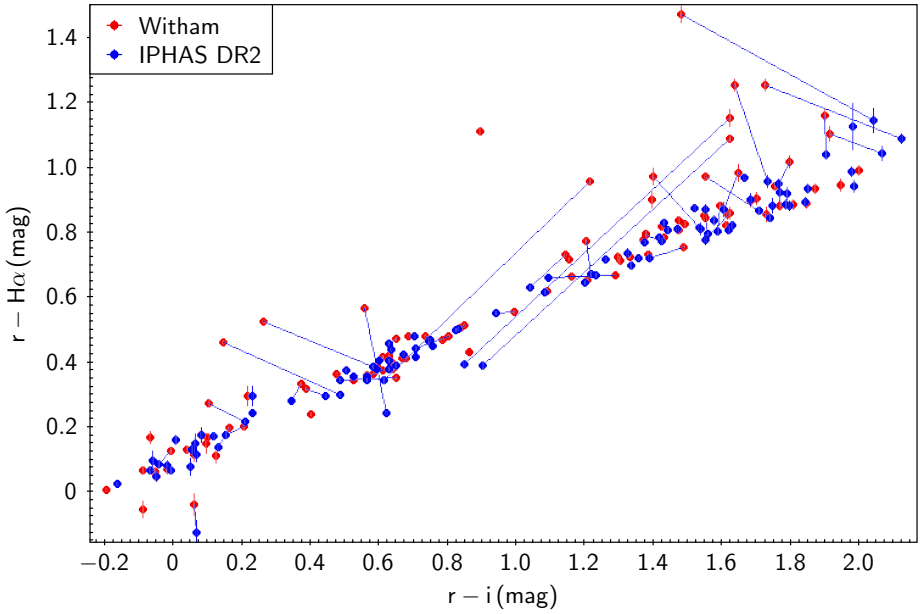


Figure 2.15: Positions in IPHAS CCD of 42 sources that were identified as outliers by Witham et al. (2008), but not by our algorithm. These are the objects the $r-$Hα intensity of which was higher in Witham's catalogue (the red dots) with respect to the analogous IPHAS DR2 intensity (the blue dots). For most of these targets, the error bars included in the plot are too small to be visible.

the different calibrations[6]. Of the 98 3σ objects missing in our list of outliers, 42 showed a stronger $r-$Hα emission at the epoch of Witham's study. This decrease in the $r-$Hα intensity might be the reason why these particular sources were identified as Hα emitting candidates by Witham et al. (2008), but not by us. Fig. 2.15 shows this apparent $r-$Hα intensity drop.

2.6.2 Comparison with IGAPS

The identification of emission-line candidates performed by Monguió et al. (2020) followed a strategy similar to that implemented for Witham et al. (2008). The main differences between these two works are related to the calibration of the data and to the morphology classes tested (Monguió et al., 2020 only excluded sources of "morphology class 0", i.e. the "noise-like sources", from being tested for Hα excess; see also Farnhill et al., 2016). Of the 53,234,833 objects in IGAPS that were tested for Hα excess, Monguió et al. (2020) produced a list of 20,860 3σ outliers (0.04% of the tested targets). The sub-sample of these candidates selected with significance greater than 5 contains 8,292 objects (0.02% of the tested sample).

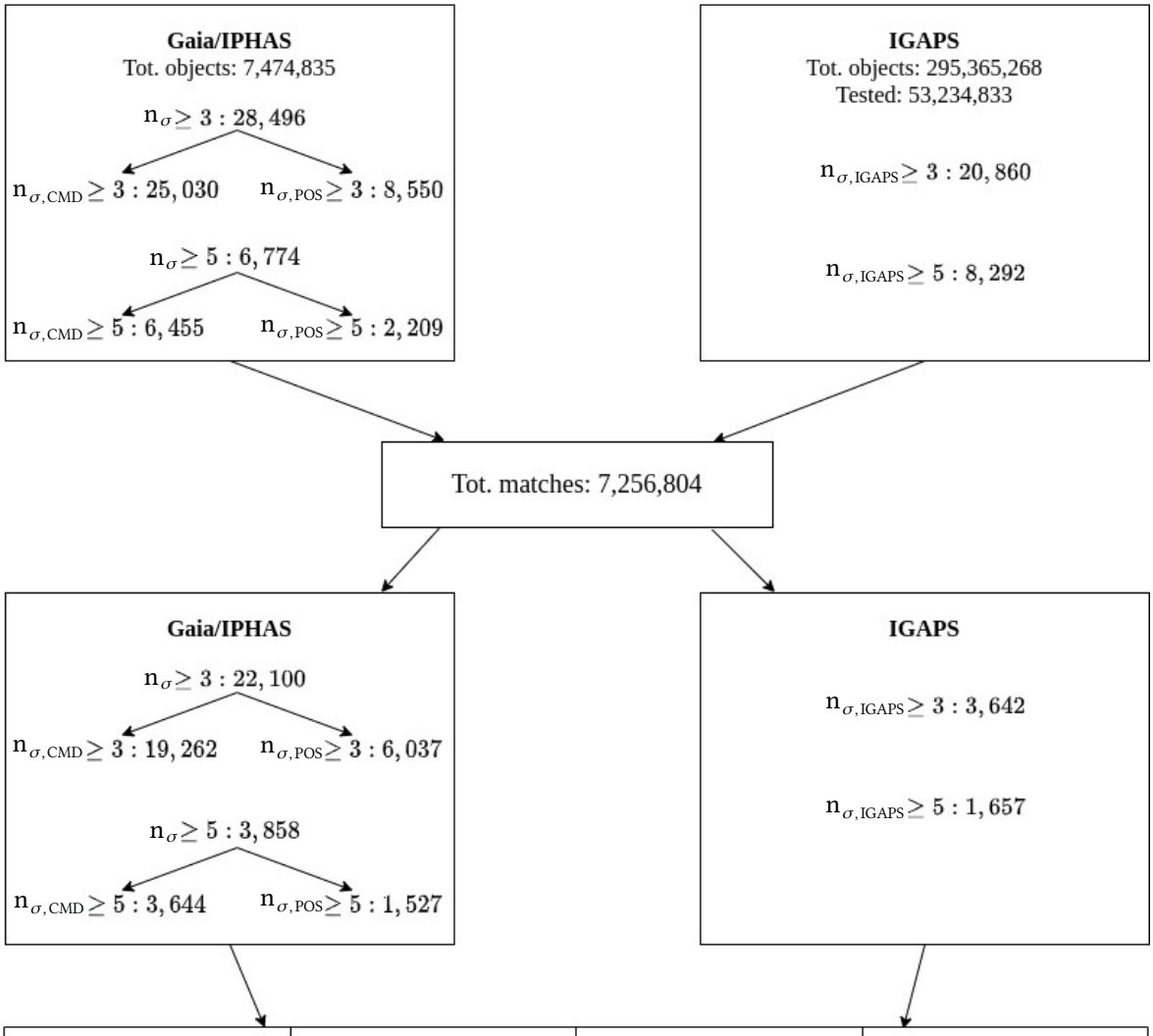
A cross-match between the *Gaia*/IPHAS catalogue (after the application of the quality

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

cuts described in Sec. 2.1.2) and the $\sim$ 53 million IGAPS tested sources yields 7,256,804 matches. As suggested in Monguió et al. (2020), the cross-matching radius used at this stage is 1 arcsecond. A test for false matches, performed by accepting duplicate matches, suggests a rate of spurious matches of 0.05%. This subset includes 3,642 IGAPS outliers, 1,657 of which with an associated significance greater than 5. It also includes a subset of 22,100 of our 3σ outliers: 19,262 of them derive from the CMD-based selection, and 6,037 from the position-based one. We point out that our CMD-based partitioning algorithm is performed on a different parameter space from 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. 2.16.

The positions in *Gaia* CMD and in IPHAS CCD of the 843 IGAPS outliers not identified by our algorithm are shown in the top row of Fig. 2.17 (the red and blue dots). Photometric measurements from IPHAS DR2 are used to produce the CCD. As can be noticed from the CMD, the vast majority of these objects can be associated to the M-dwarf region of the CMD. As previously mentioned, these objects are characterised by significantly different IPHAS colours, with respect to the other MS stars. This appears to fail our selection through the use of CMD-based partitions. In the bottom row of the same figure, our 3,417 POS-outliers that were not identified by Monguió et al. (2020) are placed in the CMD and in the CCD. These objects mainly lie in the CMD on the CBe stars region, on the ZAMS, or on the WD track. However, some red dots are placed between these two latter loci, making them good CV candidates.

The mismatches between the results obtained by Monguió et al. (2020) and by us are to be mainly attributed to two factors: the different definitions used to calculate n_σ, and the different calibrations applied to the photometric measurements in the input databases. In fact, when producing the *Gaia*/IPHAS catalogue, we based our work on IPHAS DR2 calibrated data, while IGAPS calibration relies on the more recent Pan-STARRS reference ladder (Magnier et al., 2013). This latter effect is visible in Fig. 2.18, where the $\delta\,(r{-}\mathrm{H}\alpha)$ vs. $\delta\,(r-i)$ 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 POS-outliers that Monguió et al. (2020) did not identify as Hα-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



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

Figure 2.16: The flow-chart describes the cross-matching process between the *Gaia*/IPHAS catalogue and IGAPS, as well as its detailed results.

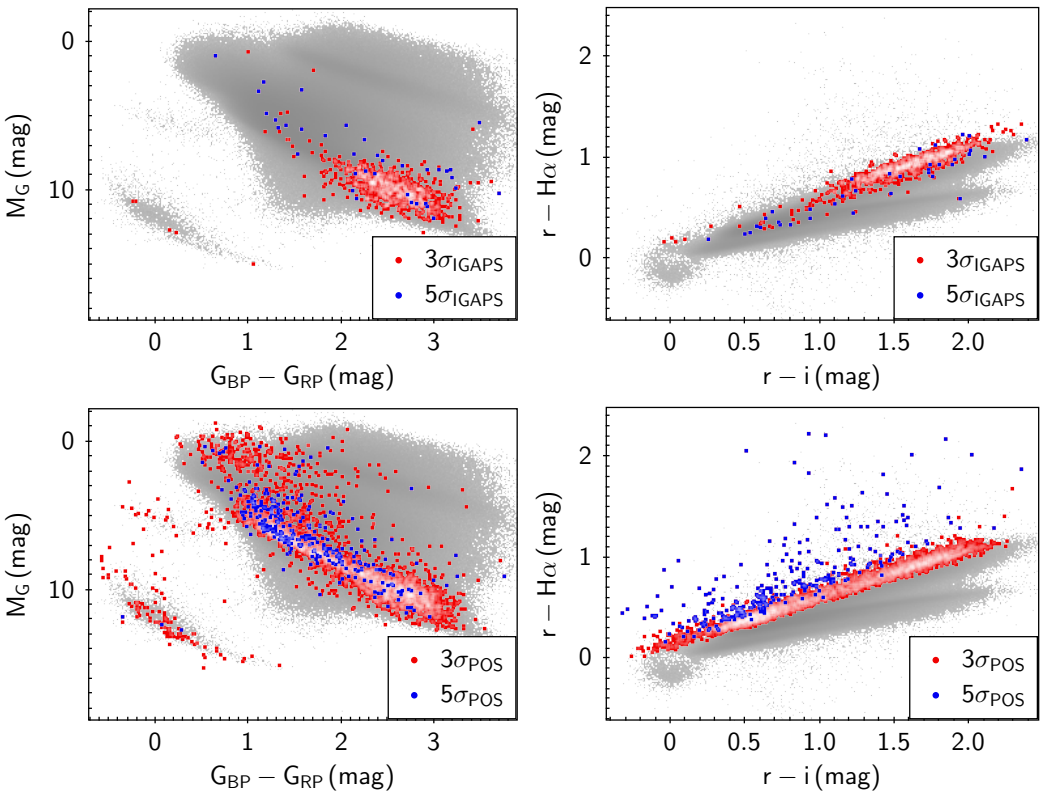


Figure 2.17: Top row: location in *Gaia* CMD (left panel) and in IPHAS CCD (right panel) of the 843 IGAPS $r-H\alpha$ outliers that are not selected by our algorithm. The red dots represent objects that Monguió et al. (2020) identified with a significance included between 3 and 5, while the objects with a higher significance are depicted with blue dots. Bottom row: position in CMD (left panel) and in the CCD (right panel) of our 3,417 POS-outliers that are not listed as Hα-excess candidates in IGAPS. The red dots represent the POS-outliers with a significance included between 3 and 5, while the blue dots represent outliers with a higher POS-significance.

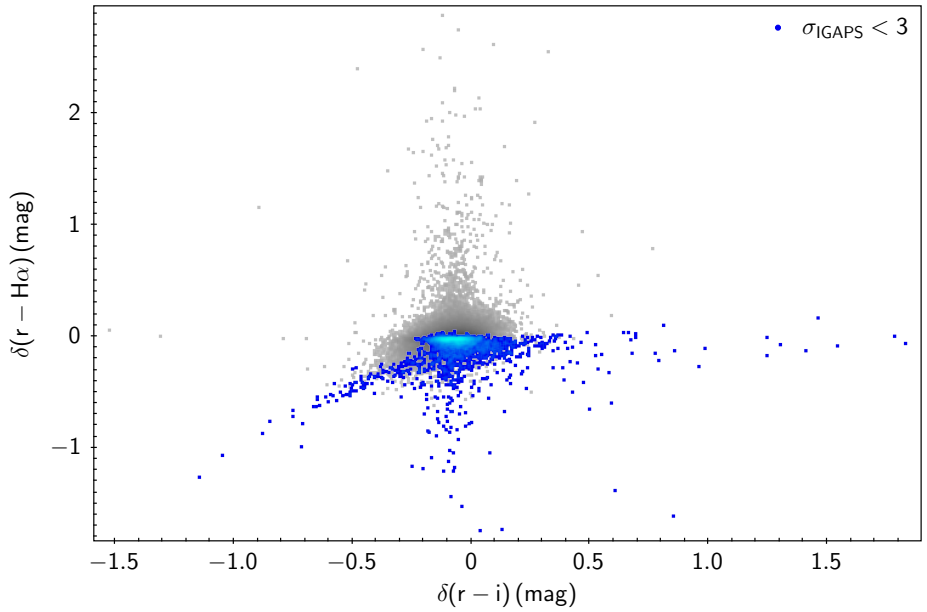


Figure 2.18: Position in the $\delta\,(r-\mathrm{H}\alpha)$ vs. $\delta\,(r-i)$ diagram of all the matches between IGAPS and the *Gaia*/IPHAS catalogue (the grey dots), and of our POS-outliers that Monguió et al. (2020) did not select as Hα-excess candidates (the overplotted blue dots).

-0.03 mag. If IGAPS and IPHAS DR2 data had the same calibration, the points in this diagram would cluster around the (0,0) coordinates. However, almost all our POS-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 position-based selection and IGAPS.

2.7 Spectra-based selection statistics

A more direct validation of our selection comes from visually inspecting the spectra of the photometrically identified Hα-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 Hα-excess sources selected by our algorithm may not display clear Hα emission, and vice-versa.

2.7.1 Purity and completeness

A cross-match between LAMOST DR5 archive and our Hα-excess candidates list (with a conservative 0.5 arcsec cross-matching radius) yields 1,873 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 1,873 objects, 916 (48.9%) are confirmed as reliable Hα-excess candidates, while 939 (50.1%) seem to show Hα 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* candidates. 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 POS-outliers, with respect to the CMD-outliers. Fig. 2.19 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 panels. Both these ratios are significantly higher than the mean in correspondence with the wavelength of interest, 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 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 $n_{\sigma_{CMD}} \geq 5$ and/or $n_{\sigma_{POS}} \geq 5$, 306 (49.7%) seem to be solid Hα-excess candidates. Out of the 603 5σ CMD-outliers for which LAMOST spectra are available, 294 (48.8%) show Hα-line emission. On the other hand, 128 spectra out of 157 (81.5%) seem to validate our 5σ position-based selection. The ratio of spectroscopically confirmed outliers improves significantly, if a more rigid cut on IPHAS magnitudes is applied (and hence if the effects due to saturation are reduced). 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 1,141 (67.7%) confirm our 3σ outlier selection (either CMD-based and/or position-based). Constraining the spectral analysis to our 5σ outliers, 231 LAMOST spectra out of 282 (81.9%) 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:

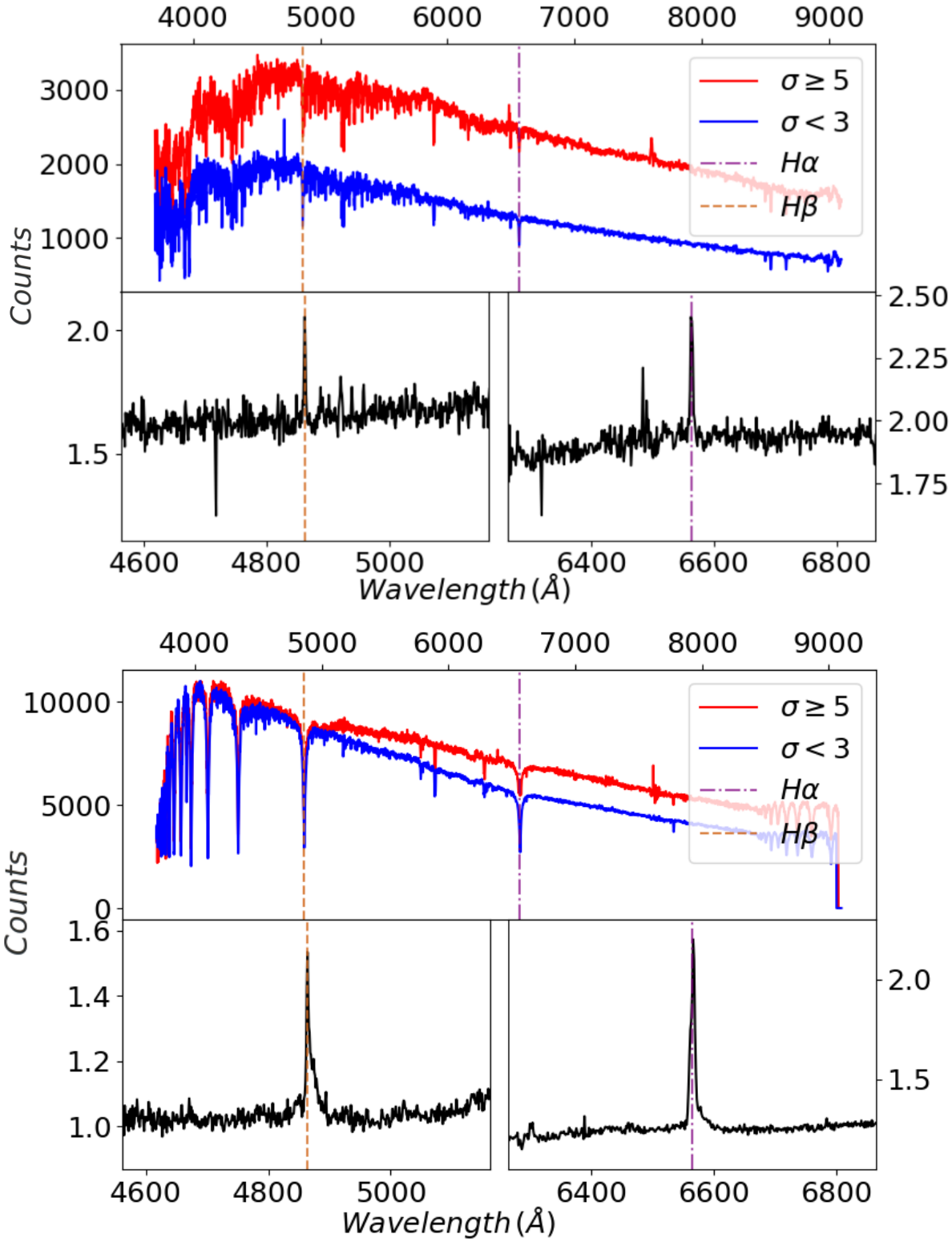


Figure 2.19: Example of two 5σ *excess* sources that show Hα absorption (the red lines in the top panels). The blue lines represent the fluxes of two objects in the same partitions, with a significance lower than 3. The bottom panels display the ratios between the fluxes in the top panels, centered around the Hβ wavelength (bottom-left panels) and around the Hα wavelength (bottom-right panels).

- the first method consists of the evaluation of:

$$C = \frac{N_{n_\sigma \geq 3} \cdot P}{N_{n_\sigma \geq 3} \cdot P + N_{n_\sigma < 3} \cdot fn} \tag{2.7.1}$$

  Here, $N_{n_\sigma \geq / < 3}$ represents 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%), and *fn* is the *false-negative* fraction (5.6%). This latter parameter derives from the visual inspection of 1,000 spectra belonging to randomly selected sources with n_σ lower than 3, and corresponds to the fraction of these spectra that show Hα emission. This method yields a completeness of around 3%. The positions of the 56 false-negative objects in the *Gaia* CMD and in the IPHAS CCD are shown in Fig. 2.20 (top panel and bottom panel, respectively). Most of these sources lie in the M-dwarfs region of the CMD;

- the second method consists of the visual inspection of 2,000 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%.

Such apparently low completeness values can partially be 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 problem are ruled out by the fact that none of our false-negative objects are included in IGAPS list of outliers. For reference, the same calculations applied on IGAPS catalogue yield a completeness percentage below 1%.

In Fig. 2.21, four spectra associated to our outliers are shown: two of these spectra belong to sources with an associated n_σ included between 3 and 5, while the other two belong to objects with a higher significance. For each of these significance-based outliers sub-samples, one example of confirmed Hα-excess, and one example of clear absorption in the Hα band are shown. These four sources are located in the *Gaia* CMD and in the IPHAS CCD in the top panel and in the bottom panel in Fig. 2.22, respectively.

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

Out of the 98 Witham's outliers that are not identified by our selection, 10 have an associated LAMOST DR5 spectrum. Four of these spectra show a clear Hα 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%) show a significant Hα emission line.

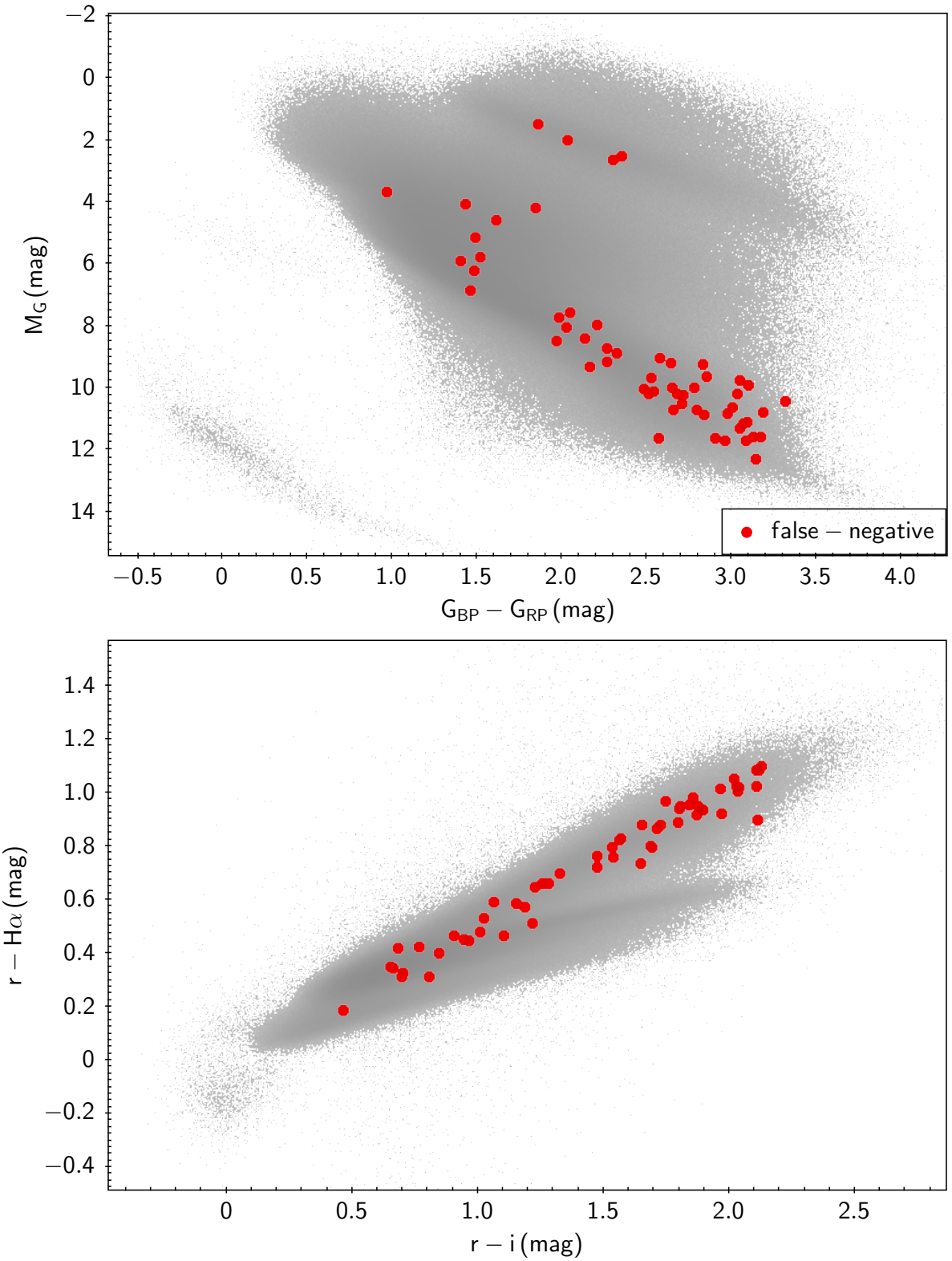


Figure 2.20: Position in the *Gaia* CMD (top panel) and in the IPHAS CCD (bottom panel) of the 56 "false-negative" objects discussed in the text.

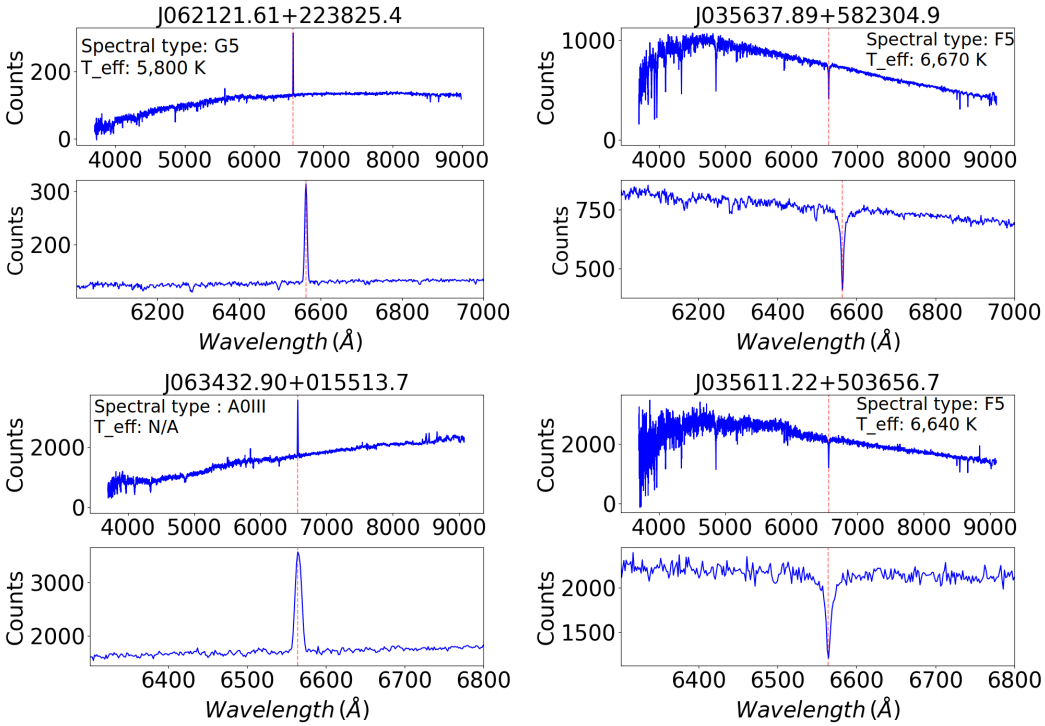


Figure 2.21: LAMOST spectra of four objects that are selected as Hα-excess sources by our algorithm. For each of these spectra, a zoom to the region around the Hα line is shown. The top two spectra belong to the sources *Gaia* DR2 3377220715714066304 (LAMOST DR5 ID: J062121.61+223825.4) and *Gaia* DR2 470024698144186112 (LAMOST DR5 ID: J035637.89+582304.9) respectively, which have an associated significance (either CMD-based and/or position-based) included between 3 and 5. The bottom two spectra refer to the objects *Gaia* DR2 3120920947508407808 (LAMOST DR5 ID: J063432.90+015513.7) and *Gaia* DR2 250435321081819392 (LAMOST DR5 ID: J035611.22+503656.7), which are characterised by a higher significance. The red dashed line indicates the Hα wavelength.

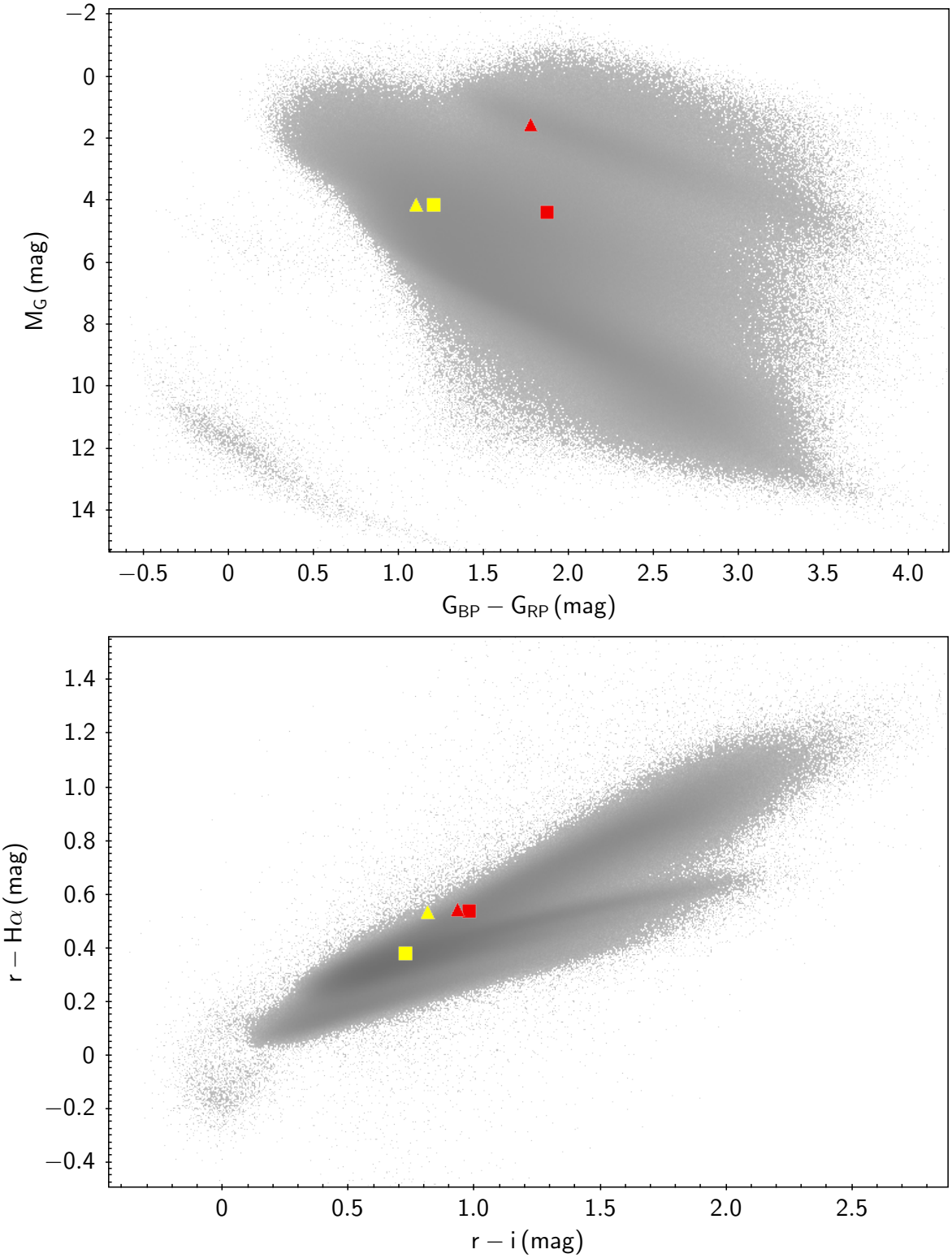


Figure 2.22: Positions in the *Gaia* CMD (top panel) and in the IPHAS CCD (bottom panel) of the objects in Fig. 2.21. The squares represent the objects with a significance (either $n_{\sigma_{\rm CMD}}$ and/or $n_{\sigma_{\rm POS}}$) included between 3 and 5, while the triangles depict the sources with a higher significance. The colour-code is: yellow for the objects whose spectra show absorption in the Hα band, and red for the objects that show Hα emission.

Regarding IGAPS 3σ outliers, 543 of them are in LAMOST DR5 archive, and 491 of these spectra (90.4%) show Hα emission. Of the 843 IGAPS 3σ outliers that our algorithm does not identify, 21 have an associated LAMOST DR5 spectrum. The absolute majority of these spectra (18/21) shows Hα emission. On the other hand, out of the 3,417 3σ position-based outliers not included in IGAPS outliers list, 149 have an associated LAMOST spectrum. By visually inspecting these spectra, 57 of them (38.3%) belong to clear Hα-excess sources. If constraining the subset to our 5σ POS-outliers, 6 out of 9 available spectra present clear Hα emission.

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

Accretion onto compact objects is often related to X-ray emission, with an intensity that increases with the compactness of the accreting source. Table 2.2 provides the results of the cross-matches 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 Chandra Source Catalogue (CSC, Evans et al., 2010).

Among the 972 matches with faint-*ROSAT*[7], 33 are identified as Hα-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[8]", and the remaining two as CVs. LAMOST spectra are accessible for 3/33 targets, and they all present a clear Hα emission line. Our algorithm assigns a significance (either CMD-based and/or position-based) higher than 5 to 12 of these 32 objects, including the two known CVs.

Out of the 69 matches with bright-*ROSAT*, 10 are 3σ outliers, and all of them find a classification in SIMBAD. Of them, three are classified as "X-ray emitting source", three as CVs plus two as dwarf novae, one as WD, and one as T-Tauri star. LAMOST spectra are available for three members of this group of sources, as well, and they all show the Hα 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

[7] No duplicate matches with either faint or bright *ROSAT* catalogues are found.

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

radius used for the cross-matches against both *ROSAT* surveys is 15 arcsec, included in the 13-25 arcsec positional accuracy of the X-ray surveys (Voges et al., 1999).

A total of 2,553 matches are found by cross-matching our catalogue and the CSC, 376 of which are identified as outliers by our algorithm. A radius of 3 arcsec is used for the cross-match; this value is overestimated with respect the suggested positional accuracy suggested in Evans et al. (2010). However, the false matches introduced by this relatively large cross-matching radius are negligible: repeating the matching process while accepting duplicate matches suggests a rate of 0.3% of spurious matches. Out of the 376 Hα-excess candidates in the CSC, 325 are classified in SIMBAD. Among them, 243 are YSOs (or candidates), 72 are Emission-line stars, two are CVs, and eight are classified as "Stars". Four of these 325 objects find a LAMOST spectrum, and the Hα emission line is visible in all of them. 181 of these 325 targets have a significance higher than 5, and the vast majority of them (153/181) are classified in SIMBAD as YSOs. In Fig. 2.23, these three groups of 325, 32 and 10 sources are located in the *Gaia* CMD (top panel) and in the IPHAS CCD (bottom panel). In agreement with their SIMBAD classification, most of the matches with CSC cluster around the area of the CMD 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".

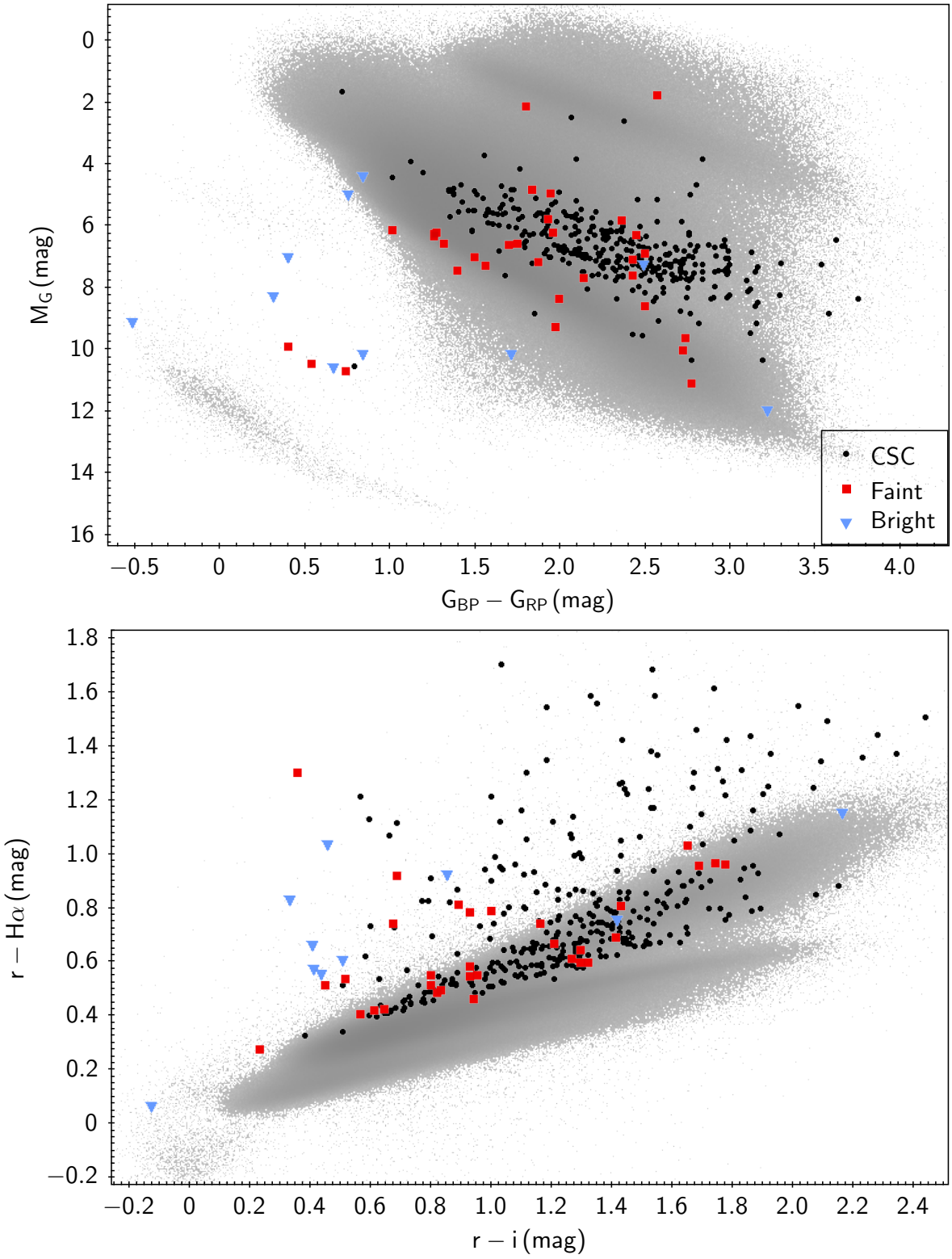


Figure 2.23: The top panel presents the positions in the *Gaia* CMD of our $r-\mathrm{H}\alpha$ 3σ outliers, that are included in the surveys bright-*ROSAT* (Voges et al., 1999, the light blue triangles), faint-*ROSAT* (Voges et al., 2000, the red squares), and CSC (Evans et al., 2010, the black dots), for which a SIMBAD classification is available. The bottom panel shows the positions of the same objects in the IPHAS CCD.

	Total	$flagCMD = 0$ $flagPOS = 0$	$flagCMD = 1$	$flagCMD = 2$	$flagPOS = 1$	$flagPOS = 2$
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	2,553	2,177 (85.3%)	157 (6.1%)	192 (7.5%)	64 (2.5%)	56 (2.2%)

Table 2.2: The table shows the results of different cross-matches between our meta-catalogue and the three X-ray surveys faint-*ROSAT* (Voges et al., 2000), bright-*ROSAT* (Voges et al., 1999) and CSC (Evans et al., 2010). The objects in our dataset are grouped before the cross-matches, with reference to the corresponding `flagCMD` and `flagPOS` specifications. These entries refer to the significances (either CMD-based or position-based) associated to each object in the meta-catalogue.

Chapter 3

Spectroscopic follow-up

In Section 2.7, I presented a preliminary validation of the photometry-based identification of Hα-excess candidates, based on LAMOST spectroscopy of randomly selected targets. However, it is necessary to point out that LAMOST observation strategy privileges relatively blue, bright objects in the Galaxy (Carlin et al., 2012). Therefore, an unconstrained cross-match with these data does not necessarily result in a homogeneous representation of the objects in the *Gaia* CMD.

In the following Sections, I describe a pilot program that enables me to validate and test the CMD-based selection technique, while simultaneously identifying and classifying new Hα-excess sources. To these ends, a set of spectroscopic follow-up observations of bright, robust, Hα-excess candidates is presented. The targets examined here are chosen in an attempt to homogeneously sample the *Gaia* CMD. In addition, I present a study on how the distribution of the spectroscopically confirmed Hα-emitters across the *Gaia* CMD, supplemented by UVEX and *WISE* parameter spaces, can be used to efficiently separate the different populations. This is essential to ease the task of target selection from large spectroscopic surveys.

3.1 Observations and reduction

3.1.1 Sample selection

The spectroscopic data analysed here were obtained at the Gemini (North and South) telescopes, via a poor-weather queue[1]. The *Gaia*/IPHAS catalogue was also cross-matched to

[1]PI: Scaringi; Program IDs: GN-2019B-Q-401, GS-2019B-Q-403.

other spectroscopic surveys, but the overlaps between its footprint and those of other surveys are generally very small. As an example, SDSS DR14 contains spectra of only 378 sources when matched against the IPHAS, and none of these have an Hα excess larger than 3σ. Other spectroscopic surveys searched, such as RAVE (Steinmetz et al., 2006), SEGUE (Yanny et al., 2009), and GALAH (DR2; Buder et al. 2018) contain no matches.

In order to reach a compromise between a minimum required signal-to-noise ratio of 10 (even for the faintest objects), and the necessary observing time, only objects with $r < 17.5$ mag were included in this follow-up. Only targets with a significance of being Hα-excess sources higher than or equal to 5 were considered.

As mentioned in Section 2.4, the Hα-excess candidates in the *Gaia*/IPHAS catalogue can be tentatively divided into five groups, corresponding to the five visually identified regions of the *Gaia* CMD with the highest density of outliers (labelled as region A-E, in Fig 2.8). The strategy to attempt to obtain a homogeneous sample across these tentative clusters has been to request follow-up observations of up to 2,230 targets, sampled as evenly as possible across the *Gaia* CMD. The location of these targets in the *Gaia* CMD is shown in Fig. 3.1. Out of these 2,230 spectra, 114 were obtained (67 from Gemini North, 47 from Gemini South). These are displayed in Fig. 3.2 as the red and blue dots in the *Gaia* CMD and in the IPHAS CCD (top and bottom panels, respectively). The red candidates were identified by both position-based and CMD-based outliers selection, while the blue ones were detected only by the CMD-based selection. None of these outliers resulted only from the position-based selection. The grey dots in the background represent all the identified 3σ outliers. As it can be seen in the top panel, the minimum-brightness constraint excludes from our spectral follow-up all the Hα-excess candidates that crowd the lowest stripe of the *Gaia* CMD. Among them we can list, for instance, all the objects that lie on (or around) the WD track. The bottom panel in the same figure brings out the value of the CMD-based selection technique. In fact, a non-negligible amount of the Hα-excess candidates (mainly blue dots) blends with the main stellar locus, in the IPHAS parameter space; without the prior population discrimination, these objects would have been hidden.

3.1.2 Gemini spectra and data reduction

The spectra analysed in here were acquired with the Gemini Multi-Object Spectrograph (GMOS; Allington-Smith et al. 2002; Hook et al. 2004), with individual exposure times of 450 s. Gemini observatory consists of two 8.1 m telescopes, based in Mauna Kea (Hawaii;

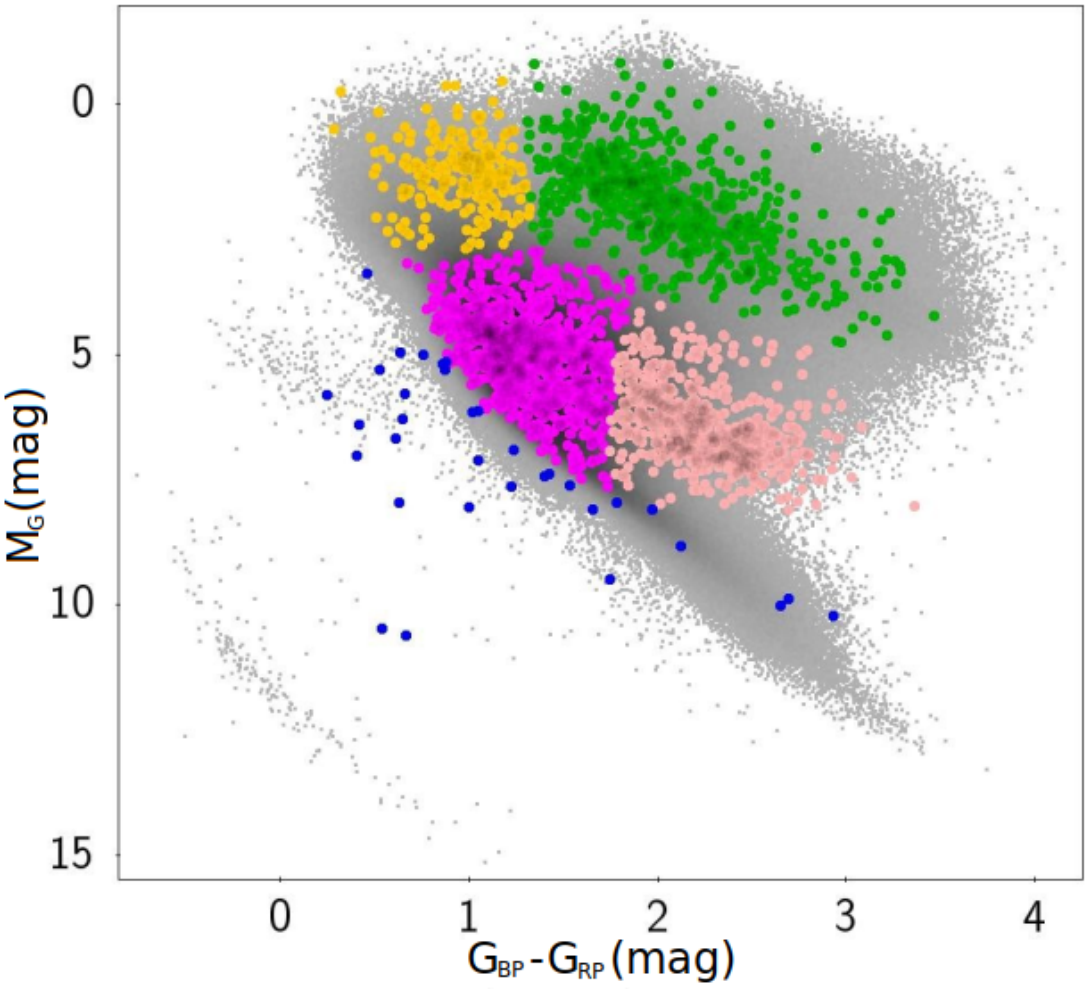


Figure 3.1: Location in the *Gaia* CMD of the targets included in the request for spectroscopic follow-up observation, divided into five cluster. These are visually defined to homogeneously sample the 5σ, bright outliers across the *Gaia* CMD.

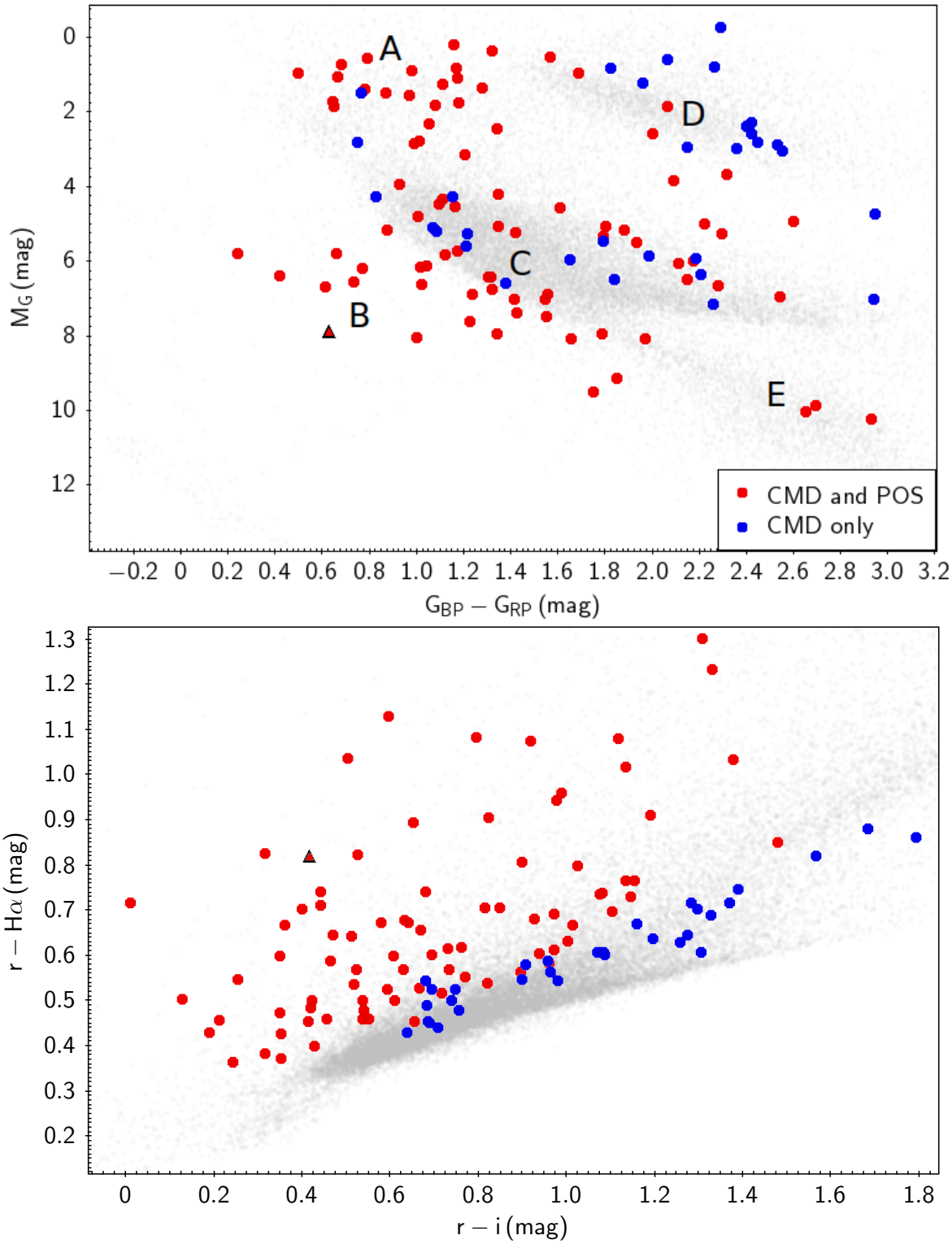


Figure 3.2: Locations in the *Gaia* CMD and in the IPHAS CCD of the 114 spectroscopically inspected targets (top and bottom panels, respectively). While red colour refers to the Hα-excess candidates selected from both partition types, the targets depicted in blue were identified from CMD-partitions only. No objects were identified from POS-partitions only. The grey dots plotted in the background as a reference represent all the 3σ outliers. The black-red triangle refers to the object *Gaia* DR2 2590186886644933l2, mentioned in the text.

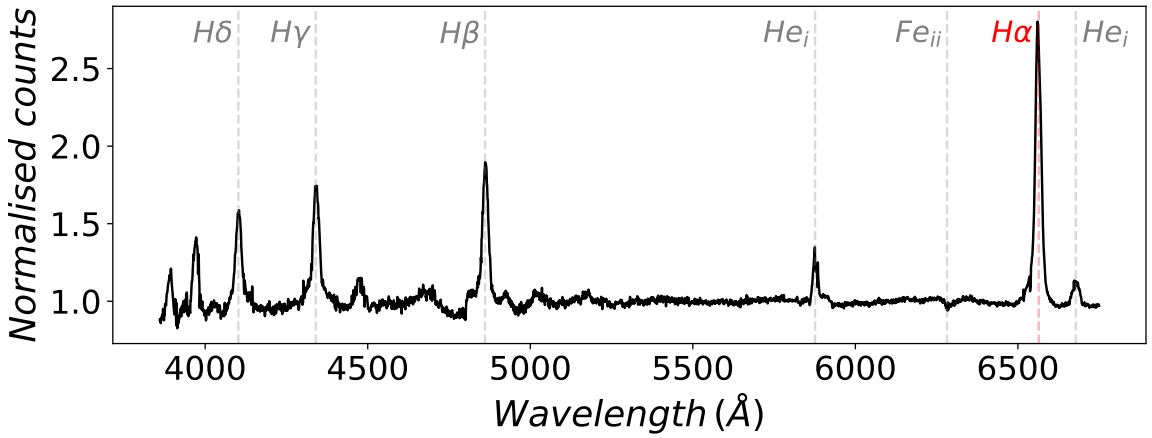


Figure 3.3: Example of reduced Gemini spectrum, belonging to the object *Gaia* DR2 259018688664493312. The most prominent emission/absorption lines are highlighted by dashed vertical lines.

Gemini North) and in Cerro Pachon (Chile; Gemini South), respectively. The observations were carried out between August 2019 and the end of January 2020. The use of the R400 grating yielded a spectral resolution at the blaze wavelength (764 nm) of R$\sim$ 1,900. The obtained wavelength resolution is 1 Å.

The setup of the detector caused systematic gaps in the data. To eliminate this effect, two spectra with different central wavelengths (more specifically, 525 and 535 nm, respectively) were consecutively collected for each target[2]. These two spectra were then averaged, in order to associate one single spectrum to each source in our list. Gemini raw data were reduced with the Image Reduction and Analysis Facility Software System (IRAF; Tody 1986) and the SAOImage DS9 tool (DS9; Smithsonian Astrophysical Observatory 2000). The process involved the subtraction of cosmic rays, as well as bias and flat-field corrections. CuAr arcs were used during wavelength calibration step. Flux calibration could not be performed, because the poor-weather program did not include the observation of any flux standard objects. To compensate for this, a relative flux calibration was carried out in two steps. First, the continuum was assessed by fitting a spline to the uncalibrated counts. The spectral lines were masked-out while this process was being done. The uncalibrated flux was then divided by the spline. Fig. 3.3 displays an example of a reduced spectrum. It belongs to the object *Gaia* DR2 259018688664493312, the position of which is highlighted with a black-red triangle in Fig. 3.2.

[2]For a total observation time of 900 s per target.

3.2 Classification

While testing the CMD-based selection technique, this spectral analysis also aims to break the degeneracy that emerges from tentatively characterising the sources solely on the basis of their location in the *Gaia* CMD. The spectroscopic classification is made more solid with the support of previous designations found in SIMBAD database (when available). Out of 114 targets, 32 are already labelled in SIMBAD. Our classifications confirm the ones found in said database for almost half of the matches (15/32). In the remaining 17 cases, SIMBAD does not provide unambiguous classifications: while 12 of these targets are classified as "emission line stars", 4 are labelled as "variable stars", and one simply as "Star".

The most common spectral features that distinguish the stellar populations mentioned in Section 3.1.2 are the following. Given the distribution of our targets in the *Gaia* CMD, MS stars other than CBe stars and active M-dwarfs are also listed here:

- **MS stars (except for CBe stars and active M-dwarfs)**: their spectra are characterised by absorption in the main Balmer and He lines (the latter ones are shown mainly in the hottest objects of this category). Among them, γ-Doradus are typically A-F spectral types stars, characterised by line-profile and radial velocity variability (predominantly produced by g-mode oscillations; Kaye et al. 1999);

- **coronally-active M-dwarfs**: molecular (TiO and CaH) absorption is one of the most prominent spectral features of (mainly early type) M-dwarfs (Rajpurohit et al., 2012, 2018) due to the low atmosphere temperature. This population usually shows relatively narrow Hα (and sometimes Ca) emission lines due to coronal activity;

- **CBe and Ae stars**: the spectra of both these object types are characterised by a blue continuum, with narrow Hα emission. Broad absorption lines might be found in correspondence to the other Balmer lines. Furthermore, Fe II and O I emission lines might be observed in the spectra of Be stars, while Ae stars often show emission in some forbidden lines. One of the peculiarities of these classes of objects is the high variability of the strength of the emission lines (Porter & Rivinius, 2003);

- **RGs**: these objects are characterised by a red continuum emission. Narrow absorption is commonly observed in the Hα, He and Fe II lines;

- **YSOs**: due to the diverse nature of the Pre-Main-Sequence objects, their spectra can show very different features. Nonetheless, they usually present an IR excess, as well as

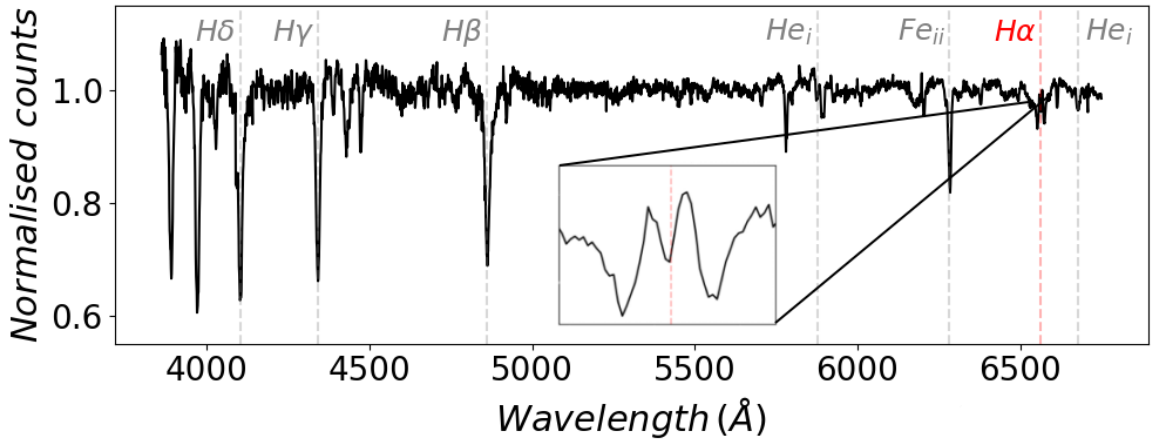


Figure 3.4: Spectrum of the object *Gaia* DR2 3455444577118544768. The shape in correspondence of the Hα wavelength does not allow the measurement of its FWHM.

the Hα line in emission. Additionally, TiO and VO molecular absorption bands can be often spotted from T-Tauri stars and from Orion variables (Itoh et al., 2010). Sources in this latter group might also present Paschen emission lines in their spectra (Hillenbrand et al., 2013);

- **CVs**: generally characterised by a blue continuum; broad Balmer - and often He - emission lines originate from the hot accretion disk (Sheets et al. 2007; Han et al. 2020).

Target classification can be further supported by the analysis of the full width at half maximum (FWHM) of the Hα line. Henceforth, the acronym "FWHM" will refer to the full width at half maximum of the Hα line, unless otherwise specified. The width of a given line can provide an indication for the presence of an accretion disk (in compact accretors): due to Doppler broadening, disk-fed accretors are characterised by relatively broad spectral lines. Out of the inspected spectra, 69/114 belong to confirmed Hα emitters, and the FWHM is measured for 68 of them. The peculiar shape of the Hα line of one of our sources does not allow the measurement of its FWHM, as testified by Fig. 3.4. The distribution of our targets in the *Gaia* CMD, colour-coded with respect to their FWHM, is displayed in Fig. 3.5. The Hα emitters (the coloured dots), are evenly spread on the CMD. The objects with the largest FWHM can be found between the ZAMS and the WD tracks. Said region of the *Gaia* CMD is commonly associated to accreting WDs; both these types of sources are characterised by the presence of a disk surrounding the accreting object (Smak 1992; Szkody 1992, Beltrán 2020; Zhao et al. 2020). The remaining 45 objects, characterised by absorption in the Hα line, are depicted by the turquoise triangles. These objects are discussed later in the text.

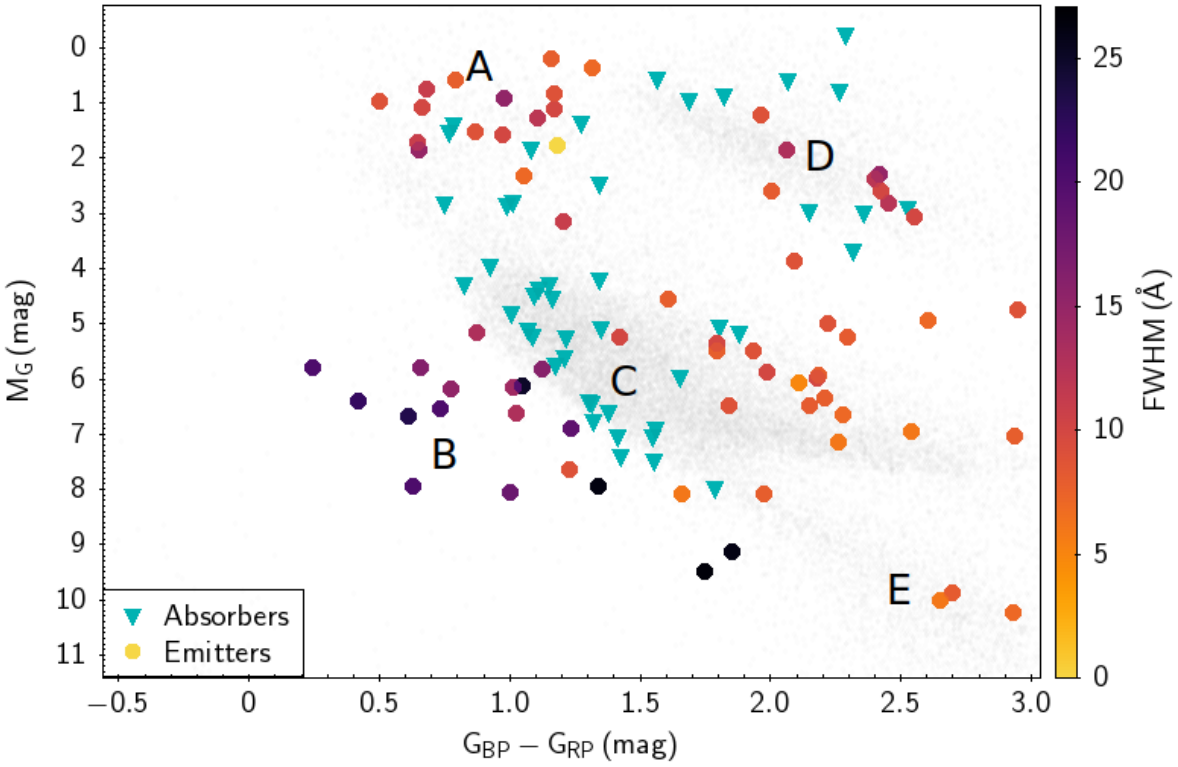


Figure 3.5: Positions of our targets in *Gaia* CMD, colour-coded according to the FWHM of the Hα line. The objects that show Hα absorption are depicted by the turquoise triangles. The grey dots represent all the 3σ outliers.

As an outcome of our spectral classification, the examined dataset is found to be a compound of: 35 MS stars (except for CBe stars and M-dwarfs); 30 YSOs (while Marton et al. 2019 recently produced a catalogue of more than 1.1 millions of YSO candidates, selected from *Gaia* DR2 and *WISE* surveys with machine learning techniques); 18 CBe stars (whereas the all-sky Be Star Spectra (BeSS) database described in Neiner et al. 2011 counts more than 2,000 Be stars); 17 CVs (to be added to the 1,429 CVs listed by Ritter & Kolb 2003); 10 RGs; 4 active M-dwarfs (Lépine & Gaidos 2011 found ∼9,000 bright active M-dwarfs from the SUPERBLINK survey. However, the authors estimate that their selection is only ∼75% complete).

Fig. 3.6 shows six exemplary spectra, one for each identified population. However, all the 114 analysed spectra are presented in Appendix A, in Figs. A.1-A.12. Their classification is included in Table A.1. The taxonomy of the spectroscopically analysed sample is shown in Fig. 3.7, where the various populations are represented in the *Gaia* CMD with different colours and shapes. As can be noticed, targets labelled likewise generally tend to cluster in well defined regions of the *Gaia* CMD - the only noticable exception to this statement is provided by YSOs which, for their manifold nature, are more widely spread in this parameter

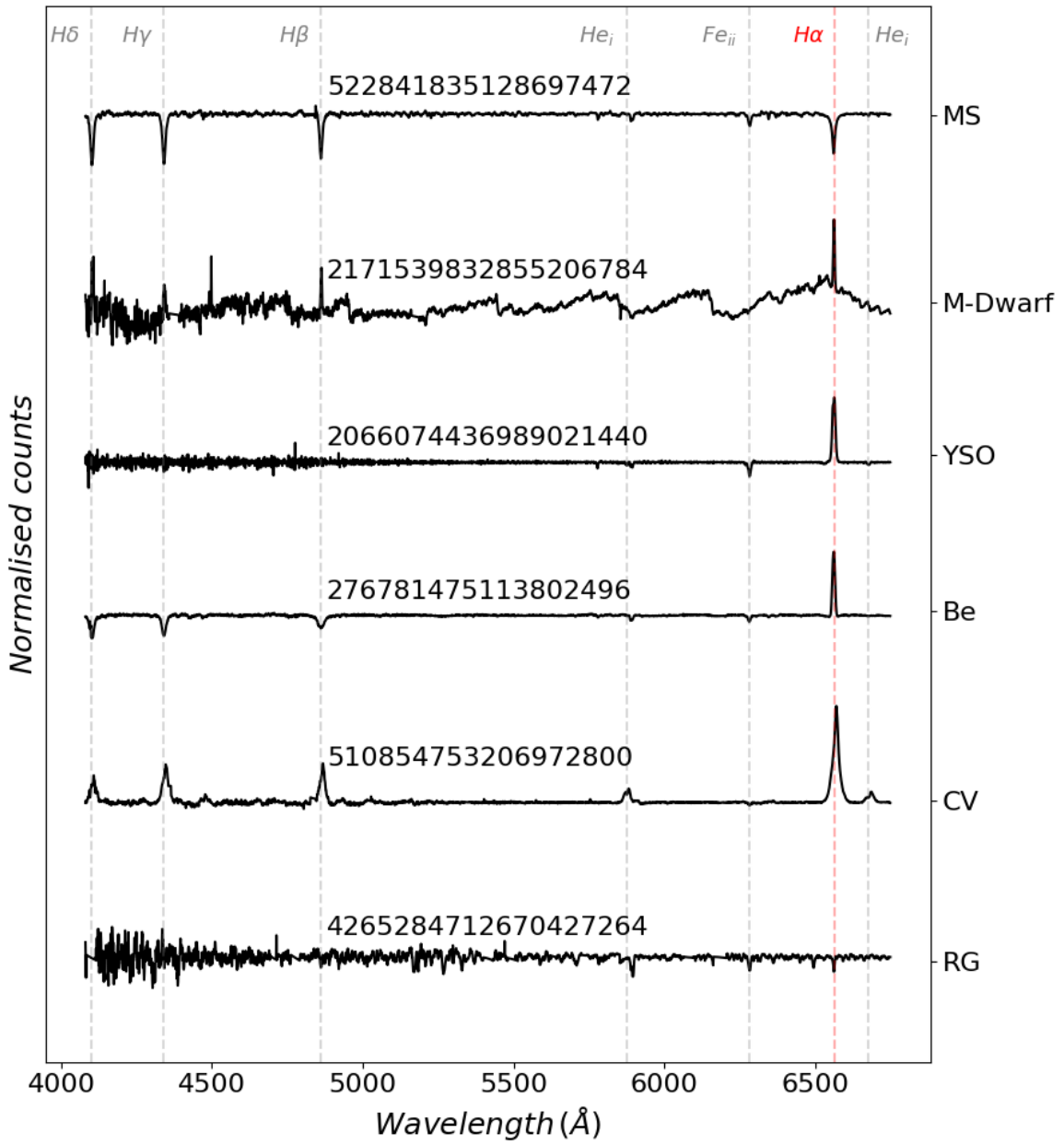


Figure 3.6: The spectra shown here (flux-normalised for visualisation purposes) exemplify the six categories to which the objects in our sample are assigned. The labels on the right vertical axis associate each spectrum to its corresponding population. The respective targets are identified by their (*Gaia* DR2) Source ID. Some Balmer, He and Fe lines are highlighted with dotted lines.

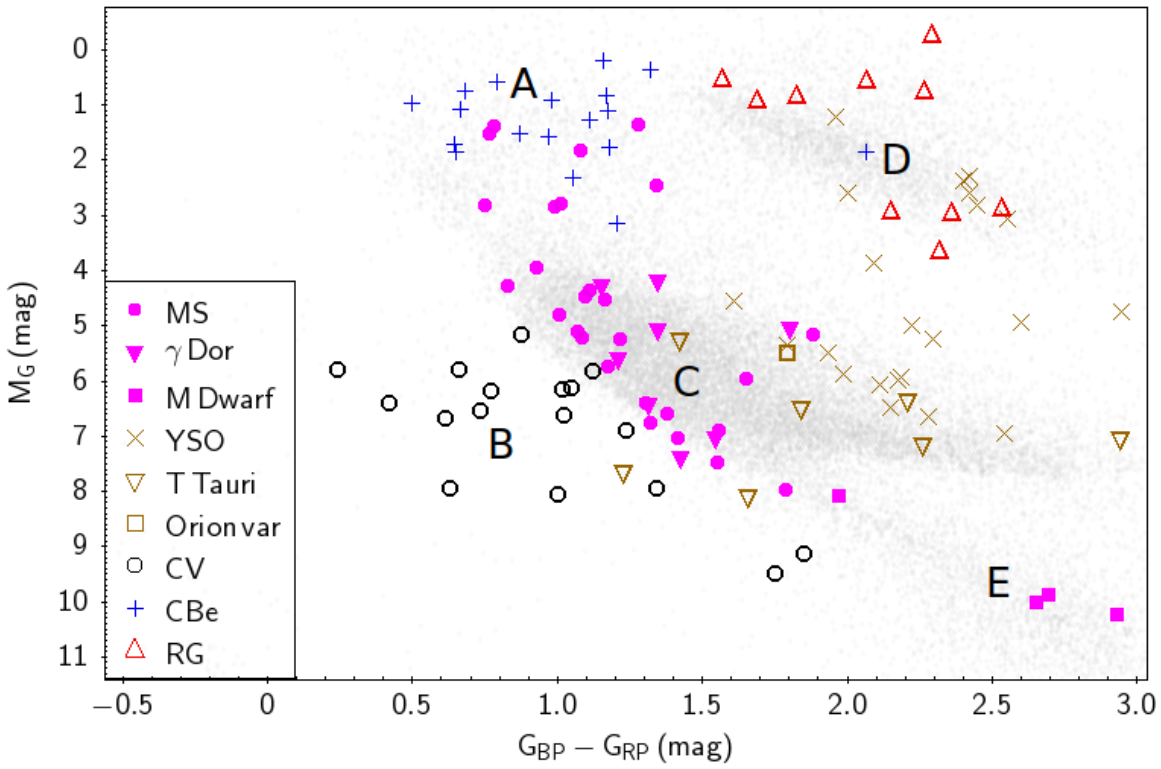


Figure 3.7: Location of the 114 inspected sources in the *Gaia* CMD, colour-coded with respect to their spectral classification. The grey dots represent the totality of the selected Hα-excess candidates.

space. However, the overlaps of these regions clearly demonstrate that the location in the *Gaia* CMD does not provide enough information to unambiguously tie a source to a stellar population. For instance: a particularly dominant companion would make it an extremely difficult task to discern a CV from a MS star; moreover, the bright and red tail of the YSO distribution seems to blend with the RGs; furthermore, distinguishing line emitting CBe stars from common, bright, MS stars can be very tricky.

3.3 Testing the CMD-based selection

The distribution in the *Gaia* CMD (top panel) and in the IPHAS CCD (bottom panel) of the sample analysed here is shown in Fig. 3.8. The colour-code is the same applied in Fig. 3.2. The dots represent the spectroscopically confirmed Hα emitters, while the triangles display the absorbers. Out of the 69 emitters, 55 were selected by both the position-based and CMD-based techniques. The remaining 14 emitters were obtained only through the CMD-based method (no outliers in the sample were identified by the position-based selection only), demonstrating the improvements introduced by this selection in terms of completeness. This

is further shown by the fact that the CMD-significance is higher than the POS-significance for 45/55 of the common emitters. All the 14 CMD(only)-emitters are spectroscopically classified as YSOs. This confirms that the CMD-based method is more efficient than the position-based one in identifying emission-line sources within this population (as stated in Section 2.5).

Out of the 45 absorbers, 28 (62%) were selected by both position-based and CMD-based selections. The remaining 17 absorbers were identified only by the CMD-based method. The 45 absorbers include 35 MS stars (other than CBe stars and M-dwarfs) and all the 10 RGs. The variability timescale typical of these stellar populations is significantly longer than the ~ 10 years that separate IPHAS and LAMOST observations. For this reason, the photometric identification of these absorbers as Hα-excess candidates cannot be justified solely on the basis of temporal variability. Out of the 10 RGs, 7 were selected only by the CMD-based method. To test the reliability of this selection criterion in region D of the CMD, a group of outliers that lie in that area is visually isolated. Considering that the vast majority of them were identified only by the CMD-based method, the purity yielded by this selection criterion in region D of the *Gaia* CMD seems to be significantly lower than the one from the position-based selection. However, half of the YSOs identified only by the CMD-based method also lie in that area. The position-based selection alone would have missed these outliers. A similar test is performed for the MS stars population: the majority of the visually selected outliers along the ZAMS were identified only by the CMD-based technique. However, 25/35 of the MS stars considered in this Chapter were identified by both selections. Therefore, the position-based one does not seem to provide advantages in terms of purity, along the ZAMS. On the other hand, one YSO that lies in that region was selected only by the CMD-based technique.

Table 3.1 summarises how many members of each population were identified as Hα-excess candidates by the CMD-based selection and how many by the position-based one. The number of spectroscopically identified Hα absorbers in each population is provided within the parentheses.

3.4 Photometry-based pre-selection

The upcoming availability of large spectroscopic surveys makes the exploration of different pre-selection methods a crucial task. The last paragraph of Section 3.2 shows how the distribution in the *Gaia* CMD can be used as an indication to classify Hα-excess sources. However, the physical and chemical properties of diverse source types can affect wider different

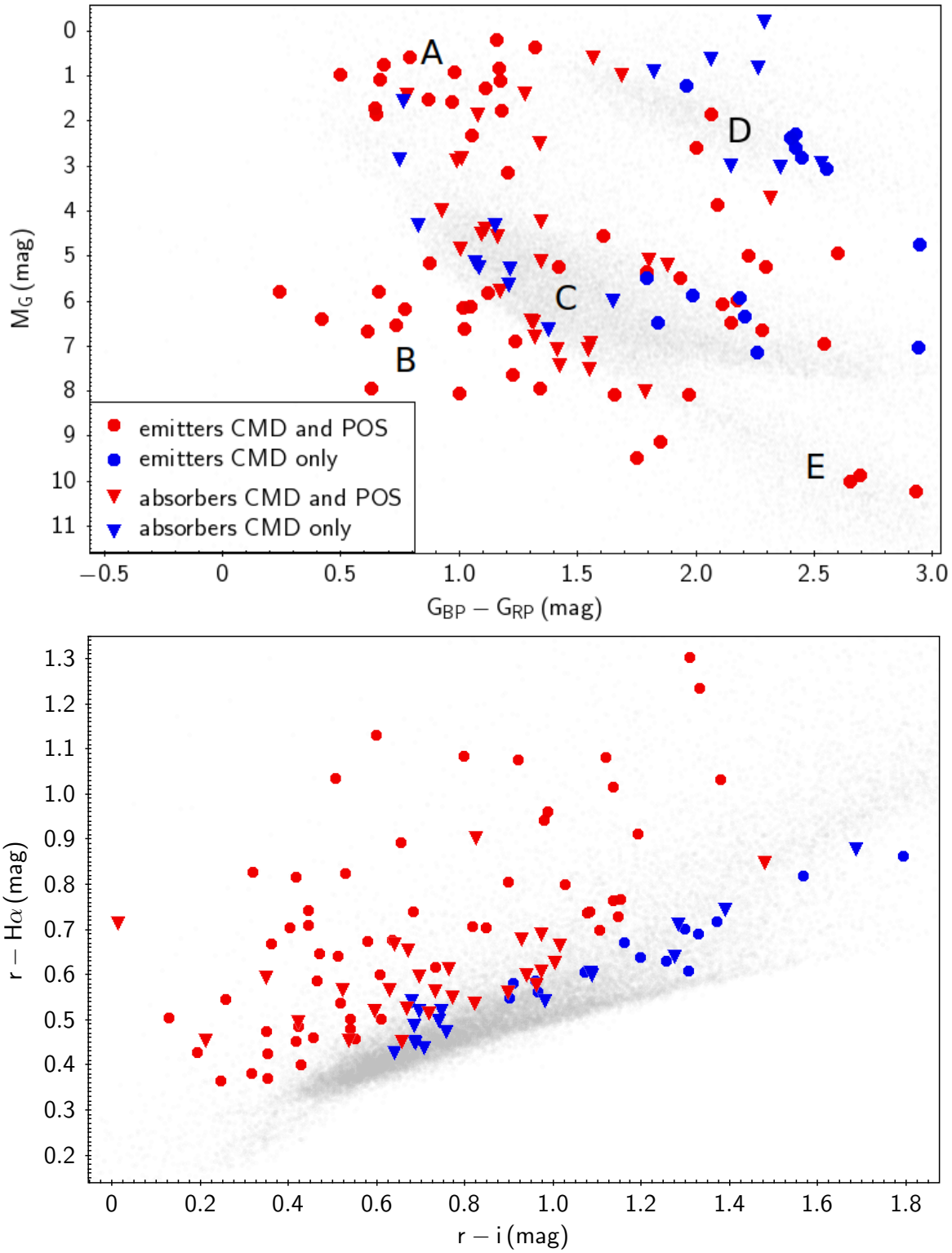


Figure 3.8: Locations in the *Gaia* CMD and in the IPHAS CCD of the 114 spectroscopically inspected sources (top and bottom panels, respectively). The colour-code is the same of Fig. 3.2. While the dots display the Hα line in emission, the triangles show it in absorption.

Population	CMD-based selection (abs)	position-based selection (abs)
YSO	30 (0)	16 (0)
CBe	18 (0)	18 (0)
CV	17 (0)	17 (0)
RG	10 (10)	3 (3)
Active M-dwarf	4 (0)	4 (0)
Other MS star	35 (35)	25 (25)
Total	114 (45)	83 (28)

Table 3.1: Number of members from each population detected by the CMD/position-based selections, respectively. The amount of absorbers from each population is included inside the parentheses.

pass-bands out of the *Gaia* ones. Therefore, supplementary information from bluer/redder pass-bands can better inform on the exact class. In this Subsection, the optical photometry from *Gaia* is complemented with UV and IR intensities, leveraged from UVEX and *WISE* surveys, respectively.

Out of the 114 spectroscopically inspected Hα-excess candidates, 78 have an available measurement in UVEX *U, g* and *r* bands, whilst 73 have associated *W1, W2*, and *W3 WISE* measurements[3]. The colours in UVEX bands are obtained through a cross-match with IGAPS catalogue with a 1 arcsec radius, as shown in Section 2.6.2. The same radius is used for the cross-match against *WISE*, as suggested in Cutri et al. (2012) and Scaringi et al. (2013b). A cross-match between these two groups (obtained by comparing *Gaia* Source IDs) yields 57 matches, i.e. half of our dataset. The non-detection of a source carries valuable information for classification purposes, per se. In fact, objects that appear to be too faint (or too bright), in relation to the limiting magnitudes of the survey, are not detected. Among the 36 objects that are not included in UVEX survey, 27 are MS stars[4] (one of them being a CBe star), and eight are RGs. These two populations are usually faint in UVEX bands, as confirmed by the fact that the MS stars and RGs included in UVEX are among the faintest objects in our sample, in the *U* band. The faintest out of the 114 inspected targets sets a lower limit of $U_{\rm MS/RG}>21.3$ mag on the apparent magnitude of the remaining MS stars and RGs. Similarly, 23 MS stars (including three Be stars and one M-dwarf) and 13 CVs have no detections in *WISE* bands. The sources in these categories with also a *WISE* counterpart are among the faintest in all *WISE* bands of interest. The $W1_{\rm MS/CV}>14.8$ mag, $W2_{\rm MS/CV}>14.8$ mag, and

[3]It is to be pointed out that the large sizes of *WISE* pixels might produce some spurious cross-matches, especially in crowded fields such as the one being studied here.

[4]However, it is worth noticing that all the four M-Dwarfs are included in UVEX catalogue.

$W3_{\mathrm{MS/CV}}$>12.7 mag constraints for the remaining MS stars and CVs are set by the faintest objects with *WISE* counterparts.

Fig. 3.9 shows the locations in the M_U vs. g-r CMD[5] (top panel) and in the *W2*-*W3* vs. *W1*-*W2* CCD (bottom panel) of all the objects with valid measurements in the corresponding bands. Limiting the analysis to the Hα-emitting populations, a visual inspection of the three parameter spaces combined shows that CBe stars and YSOs seem to be better separated in UVEX CMD than in the *Gaia* CMD. CVs are clearly bluer compared to the ZAMS in UVEX CMD than in the *Gaia* CMD while M-dwarfs appear to be equally well identifiable in UVEX CMD or the *Gaia* CMD. Future studies may be able to further leverage multi-band photometry to quantify the class separations of Hα-emitting sources.

[5]The zero-extinction assumption is also applied to obtain the absolute magnitude in the UVEX U band (M_U).

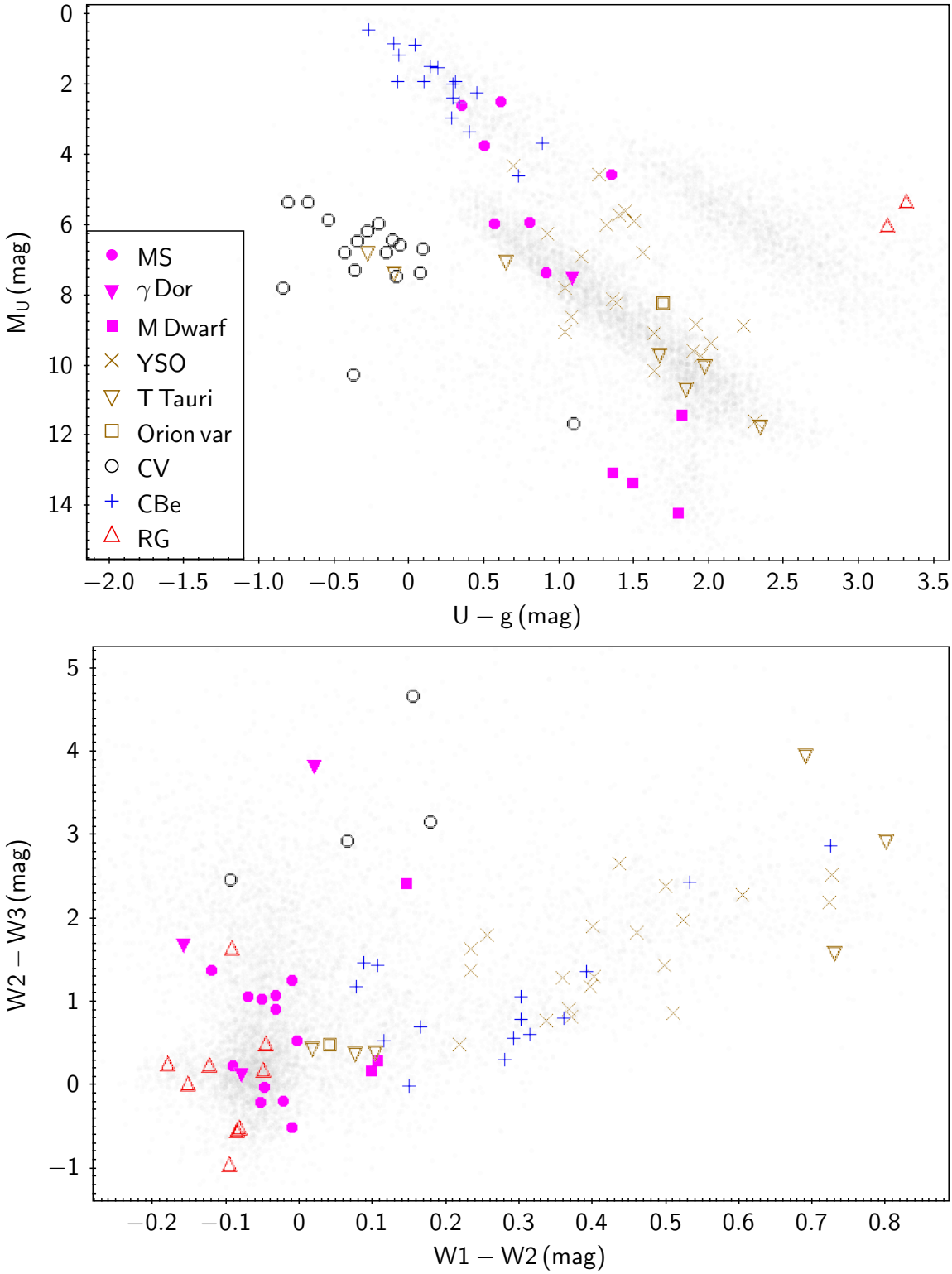


Figure 3.9: Location in the UVEX CMD and *WISE* CCD of the sub-samples with available photometric measurements in the corresponding bands. The colour (and shape)-code applied here is the same of Fig. 3.7. The grey dots in the background depict all the 3σ Hα-excess candidates, included in UVEX or in *WISE* surveys.

Chapter 4

Identifying accreting WD candidates

The previous Chapters describe how the novel population-based selection of Hα-excess candidates was performed and further validated. However, the relevance of this technique is to be found in its several potential applications.

Section 1.1 illustrates the value of examining the Hα emission to identify new members of the varied listed populations. Section 3.2 shows how the distribution in the *Gaia* CMD can be used as indication to preliminarily classify Hα-excess candidates, while in Section 3.4 said classification is improved with the inspection of UVEX and *WISE* parameter spaces.

In this Chapter, I bring two examples of how the information about the excess in the Hα flux can be complemented with further additional data, to obtain a deeper insight on the nature of the candidates. More specifically, I focus on how the identification of accreting WD candidates in the *Gaia*/IPHAS catalogue can be supported by a) inspecting the X-ray range of the electromagnetic spectrum, and by b) evaluating the variability amplitude of the optical intensity of our candidates. I conclude the Chapter by exploring the temporal variability of the newly classified CV HBHA 4204-09. This is performed through the inspection of the light curve observed by *TESS*.

4.1 X-ray emission

Accreting WDs usually occupy a well defined region in the CMD (between the MS and the WD tracks), and most of them are associated to both Hα and X-ray emission. The information about X-ray emission from our Hα-excess candidates are leveraged from the cross-matches with the three X-ray surveys presented in Section 2.8. Fig. 4.1 shows the positions in the

Gaia CMD (top panel) and in the UVEX CMD (bottom panel) of all the 3σ outliers in the *Gaia*/IPHAS catalogue that also present X-ray emission. Among the ones that are not yet classified in SIMBAD, four (the black triangles) are found in the region of the *Gaia* CMD between the WD and MS tracks, and clearly detach from the latter. Such location makes them robust accreting WDs. In the UVEX CMD, these objects lie very close to the WD track, showing that the flux from the WD is dominant with respect to that of the companion star, in UVEX bands.

However, the significance of the addition of UVEX photometric measurements stands out more clearly when more ambiguous objects, such as the black squares, are considered. In the *Gaia* CMD, these objects do not significantly detach from the MS track, and could be easily mistaken for odd YSOs (as shown in Fig. 3.7). On the other hand, the same targets appear bluer in the UVEX CMD, and are more clearly separated from the MS track. Therefore, their tentative identification as CVs is more robust[1].

4.2 Variability

Another feature that characterises accreting WDs is their relatively large variability amplitude, with a corresponding root mean square of ∼20-30% (or greater, in some case). Deason et al. (2017) and Abrahams et al. (2020) developed a method that enables the calculation of a parameter (ϵ) that quantifies the fractional rms relative to the flux of a source. With the use of *Gaia* metrics, this parameter is given by:

$$\epsilon = \sqrt{N} \cdot \frac{\delta f_G}{f_G}. \tag{4.2.1}$$

Here, f_G is the mean *Gaia* G-band flux obtained with N observations, while δf_G represents the corresponding standard error. The sources in the *Gaia*/IPHAS catalogue are binned with respect to their G-band magnitude (with a bin width of 0.2 mag); the ones with an ϵ larger than five standard deviations above $\epsilon_{\mathrm{mean,i}}$ (i.e. the average ϵ value for the i-th bin) are selected as variables. With this method I identify 22,199 significantly variable sources, 2,243 of them are also Hα-excess sources[2]. These represent 10.1% of the variable sources.

Since most of the Hα-emitting populations are also highly variable, a quantification of their variability can provide a further robust constraint to identify Hα-excess sources. This can

[1] Moreover, the spectrum of the brightest of these objects was observed with Gemini during the spectroscopic validation of our selection, and it confirms the "CV" classification.

[2] Different binnings produce results that do not significantly deviate from these values

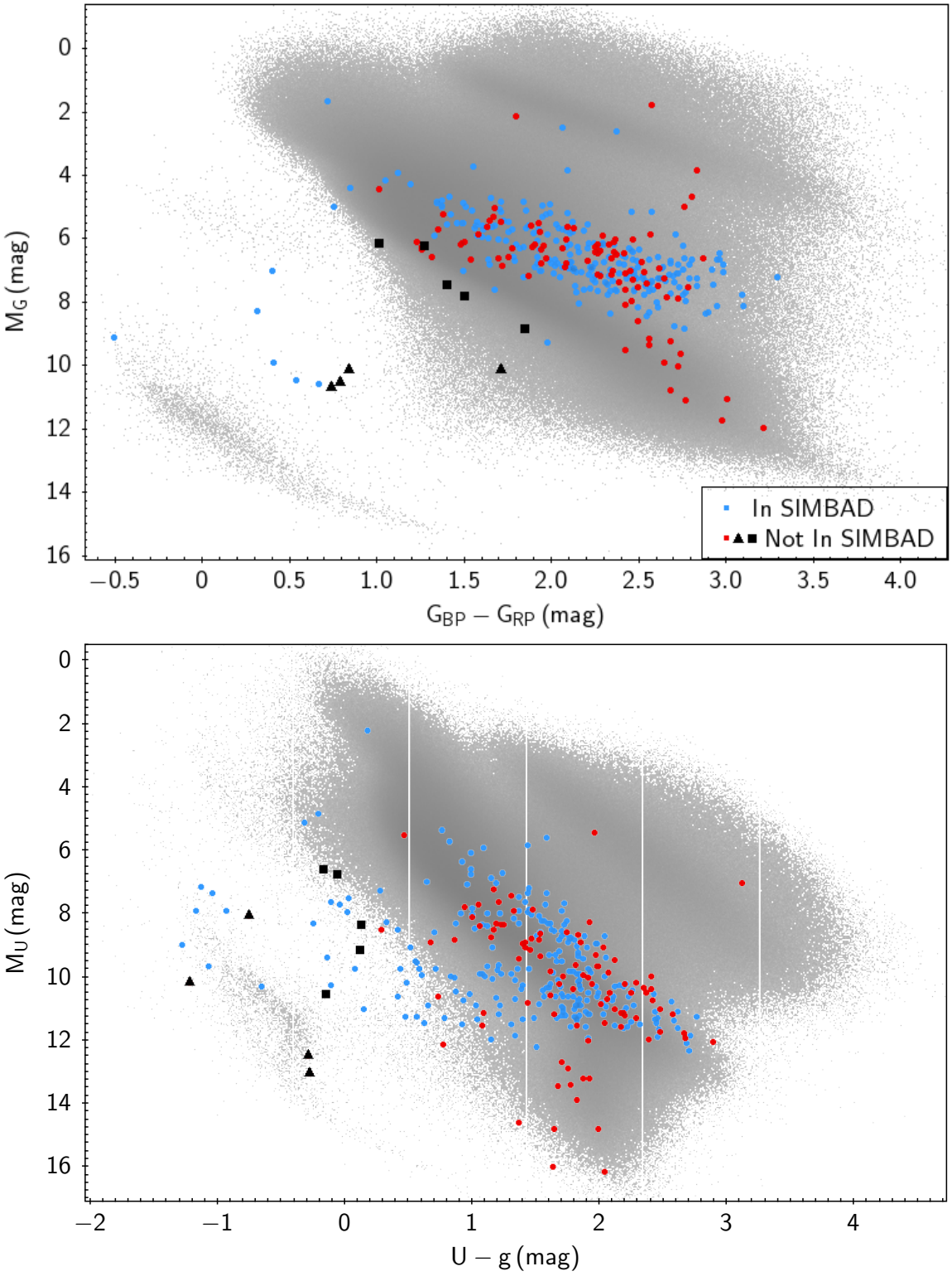


Figure 4.1: Position in the *Gaia* CMD (top panel) and in the UVEX CMD (bottom panel) of the Hα-excess candidates that show X-ray emission. The few vertical lines in the UVEX CMD are simply a visual effect connected to the sensitivity of the instrument. The light-blue dots represent the objects that are already classified in SIMBAD, while the red dots, and the black triangles and squares are still unclassified (or simply classified as "Star" or "X-Ray emitting source").

be confirmed by considering that the identified Hα-excess candidates represent 0.36% of the full *Gaia*/IPHAS catalogue. A subset of variable sources from a group of 28,000 randomly drawn targets (similar to the amount of Hα-excess candidates), with no constraint on the Hα-excess significance, constitutes on average[3] the same percentage (0.36%) of the identified variable sources. This ratio is ~ 28 times lower than the percentage of variable sources that also show Hα-excess.

The top panel of Fig. 4.2 shows the distribution of the objects in the *Gaia*/IPHAS catalogue in the ϵ vs. G-band magnitude diagram: the blue dots represent all the variable objects, whilst the red ones are the variable Hα-excess sources. The bottom panel in the same figure shows the location of the variable Hα-excess sources in the *Gaia* CMD. While most of them cluster in the region associated to YSOs, many targets are found in the region of the CMD where stars of B or A types lie, and between the MS and the WDs tracks. These latter group constitutes a promising selection for follow-up spectroscopic observations, in search for new CVs.

A combination of X-ray emission and variability makes the identification of accreting WD candidates more solid. Fig. 4.3 presents the position in the *Gaia* CMD (top panel) and in the UVEX CMD (bottom panel) of the Hα-excess candidates that show X-ray emission and variability. The nine black objects in Fig. 4.1 are all included among them. The fact that these objects are: a) Hα-excess candidates, mostly with a significance greater than five (only one of them has a significance included between three and five); b) significantly variable; c) detected in the X-ray; d) located in a region of the UVEX CMD usually associated to CVs, make them extremely robust CV candidates. The blue dots are included only in the top panel, since they represent the variable Hα-excess candidates that show X-ray emission, but were not detected by UVEX survey. Their location in *Gaia* CMD, complemented with the excess in the Hα flux and the variability information, makes these objects good YSO candidates.

All these pieces of information can be combined to provide a criterion to reliably identify new CV candidates. An example of a newly identified CV is provided by the target HBHA 4204-09 (or also *Gaia* DR2 2162336164614235648), the classification of which was confirmed by the spectral investigation presented in Chapter 3. In the next Sections, I first show the classification of this object based on its spectro-photometric properties; then I focus on its

[3]The process is repeated 10 times.

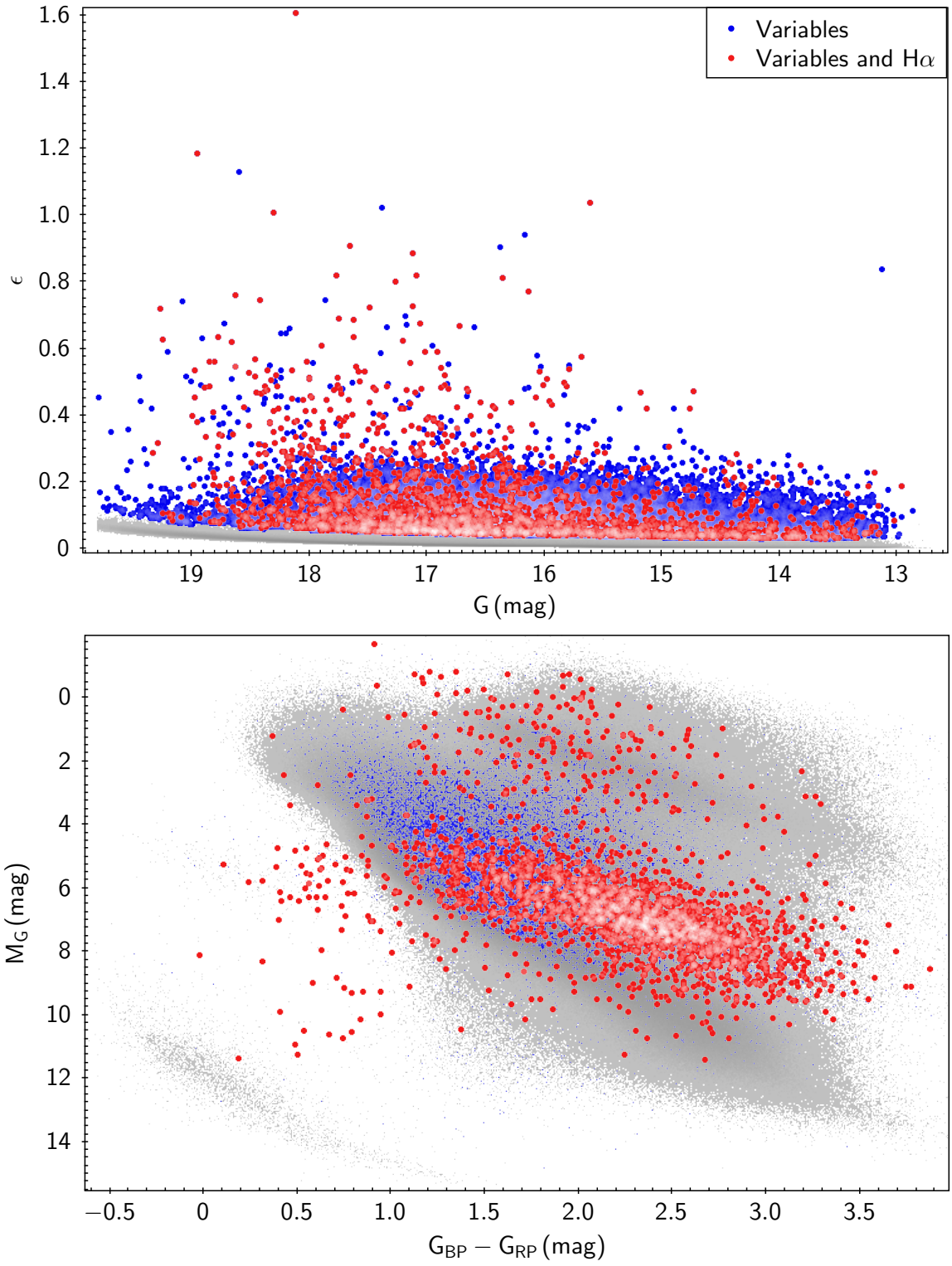


Figure 4.2: Graphical depiction of the location in the ϵ vs. G (mag) diagram (top panel) and in the *Gaia* CMD (bottom panel) of the variable objects in the *Gaia*/IPHAS catalogue (the blue dots) and of the variable Hα outliers (the red dots). The grey dots represent the non-variable objects.

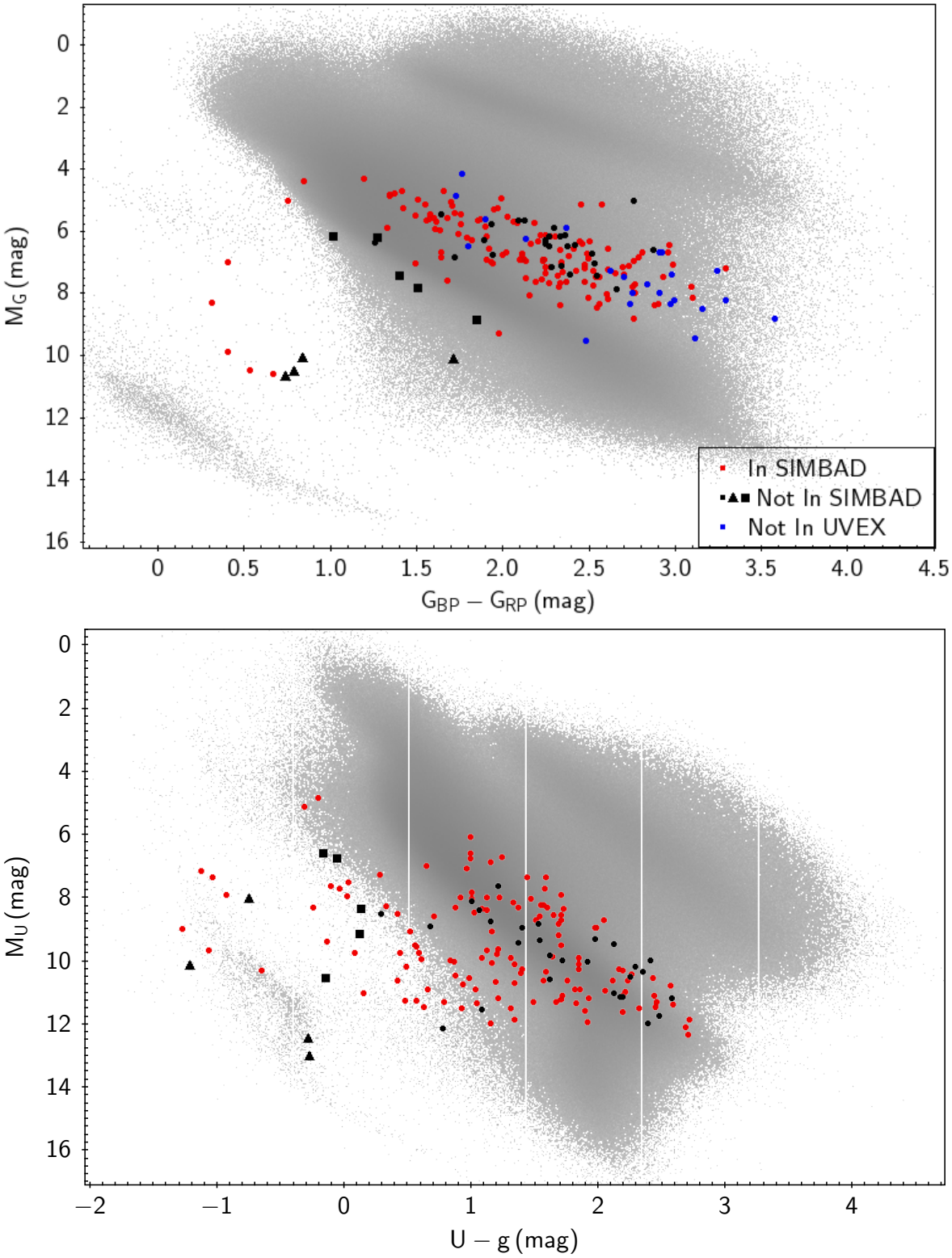


Figure 4.3: Location in the *Gaia* CMD (top panel) and in the UVEX CMD (bottom panel) of the variable Hα-excess candidates in the *Gaia*/IPHAS catalogue that also present X-ray emission. The red dots represent the subset that is already classified in SIMBAD, while the black objects are not. The blue dots in the top panel corresponds to the objects in this group that do not have a UVEX counterpart to *Gaia* photometry. The black triangles and squares are the same targets shown in Fig. 4.1.

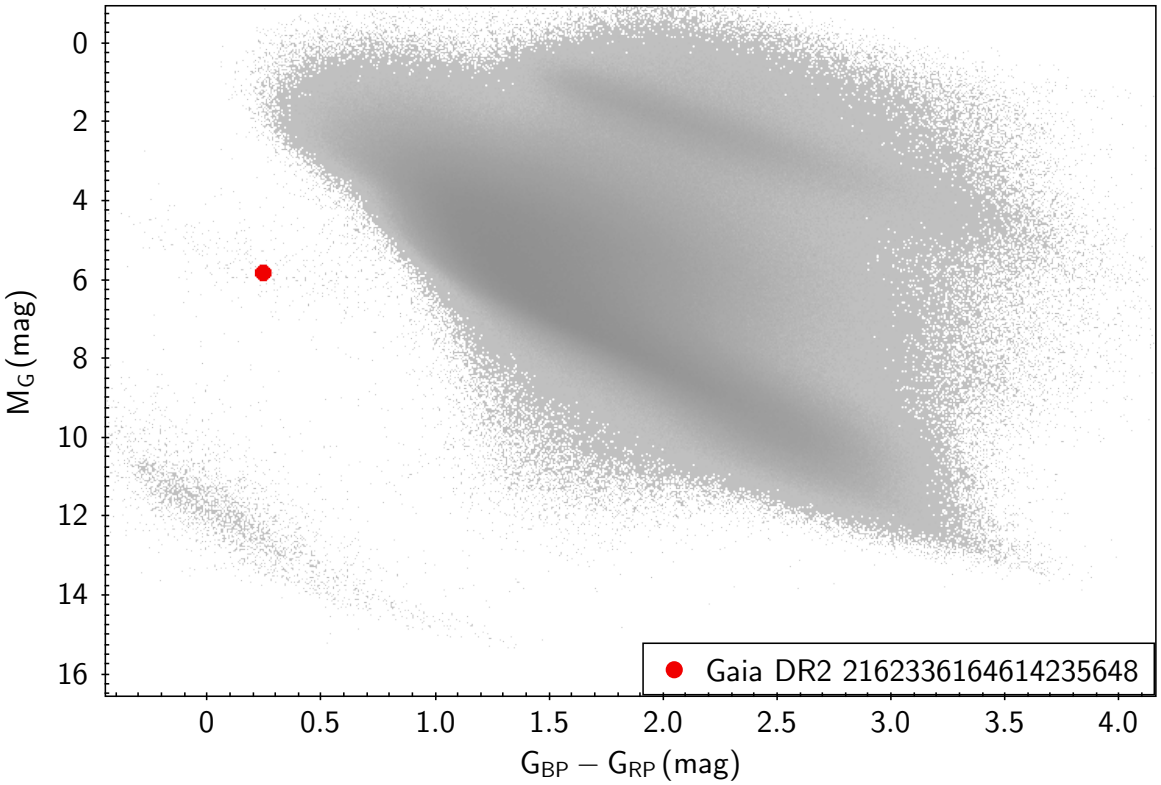


Figure 4.4: Location in the *Gaia* CMD of the source HBHA 4204-09.

variability features, by implementing a timing analysis of its *TESS* light curve.

4.3 HBHA 4204-09

4.3.1 Classification

HBHA 4204-09 was first identified as an Hα-excess candidate by Kohoutek & Wehmeyer (1997), and later confirmed as such by the selection of Witham et al. (2008). More recently, Geier et al. (2019) included it in their list of hot subdwarf stars, consistently with its location in the *Gaia* CMD (Fig. 4.4). The CMD-based selection identifies this target as an Hα-excess candidate with a significance higher than 10. In addition, HBHA 4204-09 is also included in the list of variable stars introduced in the previous Section.

The prominent emission in the Balmer lines (Hα - Hϵ), as well as in the He I and He II, clearly visible in its spectrum (top panel of Fig. 4.5) are typical of a CV. The observed central wavelength for the Hα peak (6569Å), longer than the rest wavelength (6564Å), suggests that

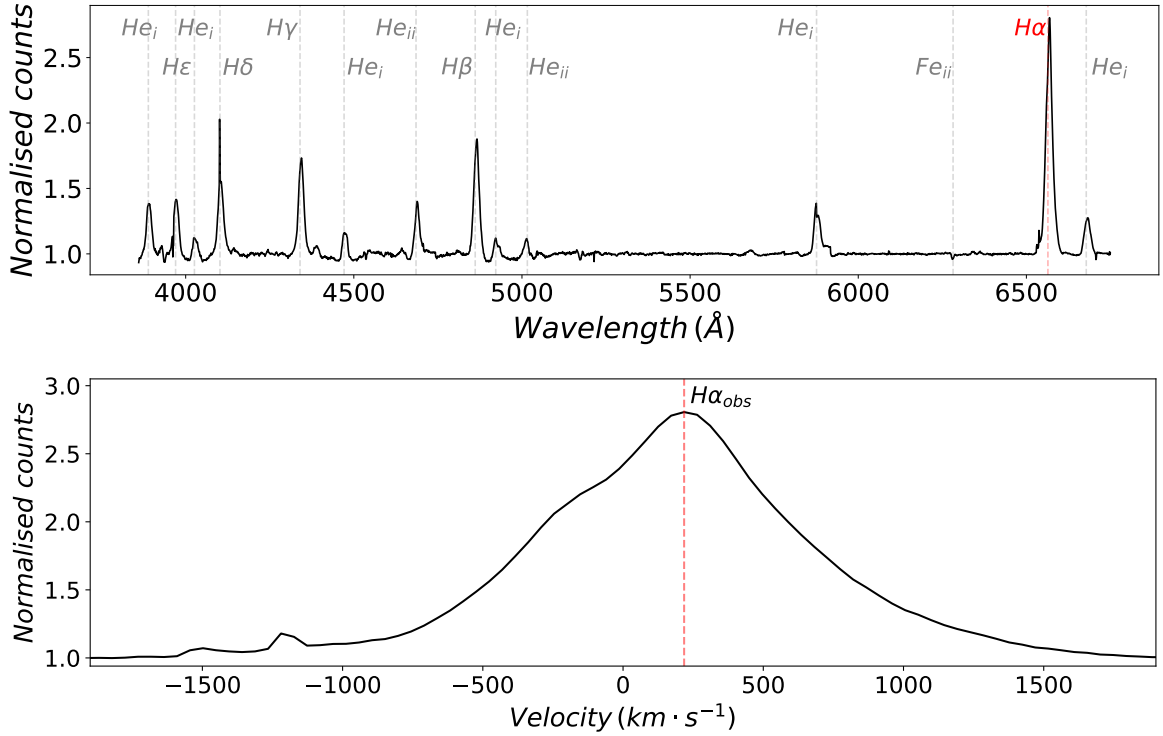


Figure 4.5: Top panel: spectrum of the target HBHA 4204-09, observed by Gemini telescope. The main spectral lines inspected to classify this source as a CV are highlighted with dashed grey lines (except for the red line, that marks the Hα wavelength), and labelled. Bottom panel: Doppler velocity profile of HBHA 4204-09, calculated on the basis of the displacement of the observed Hα line with respect to the rest wavelength.

the target is moving away from us with a Doppler radial velocity[4] of 218(46) km·s^{-1} (bottom panel of Fig. 4.5). On the other hand, the large FWHM of the Hα line ($\sim$ 21 Å) indicates the presence of an accretion disk, and corresponds to a velocity dispersion of 910(91) km·s^{-1}. The high relative intensity of the He II λ4686 line with respect to the Hβ line (HeII λ4686/Hβ>0.5) suggests a classification of this object as a nova-like variable, rather than as a dwarf nova (Warner, 2003; Hou et al., 2020). However, a solid discrimination between these two classes based exclusively on their spectral features is not possible, and it is left for future analyses.

HBHA 4204-09 has not been detected by any X-ray survey. This can be due principally to two reasons: a) the X-ray telescopes have not observed in the direction of the coordinates of HBHA 4204-09 at all, or b) they have not observed that patch of sky with long enough exposure times to detect it. A possible way to address this issue is to assume an average nova-like X-ray luminosity from HBHA 4204-09. Here, I used the X-ray luminosity of the well known

[4]It is necessary to point out that the Heliocentric correction is not performed at this stage. However, the measured velocity is only partially associated to the motion of the Earth around the Sun, as the average velocity of revolution is $\sim$ 30 km·s^{-1}.

nova-like variable MV Lyrae as a reference: Balman et al. 2014 measured this value with the *Swift* X-Ray Telescope (*Swift* XRT; Burrows et al. 2003) to be $L_x \sim 6.6 \times 10^{32}\,\mathrm{erg{\cdot}s^{-1}}$ (measured during a period in which the source was in the high state), using an exposure time of 6,600 s. Considering that the distance of MV Lyrae and HBHA 4204-09 from us are comparable (respectively 0.497 kpc (Gaia Collaboration, 2020) and 0.475 kpc[5]), in order for the latter object to be detected, it should have been observed by such an instrument with a comparable exposure time. However, *Swift* never observed in the direction of HBHA 4204-09. Similarly, *ROSAT* observed the patch of sky corresponding to HBHA 4204-09 for a total of 115 s; even considering that the effective area of the X-ray detector on *ROSAT* is approximately three times larger than that on *Swift*, a detection of HBHA 4204-09 by *ROSAT* is unlikely, with an observation shorter than 2.200 s. Therefore, further follow-up observations in the X-ray are needed to detect this source in this band.

4.3.2 *TESS* light curve

TESS observed HBHA 4204-09 during sectors 15 and 16, for a total of ~ 51 days (from BJD 2458711 to BJD 2458762), with a cadence of 120 s. The uncalibrated light curve, found in the Mikulski Archive for Space Telescope (MAST[6]), is shown in Fig 4.6. Sector 15 is displayed in the top panel, whilst the bottom panel shows the light curve observed during Sector 16. The calibration is performed by splitting the data from each Sector in two batches of roughly equal duration, and by comparing them to the flux from the All-Sky Automated Survey for SuperNovae (ASAS-SN; Shappee et al. 2014, Kochanek et al. 2017). Splitting the Sectors in two parts is necessary to account for the data gaps in the light curve. These correspond to the intervals during which *TESS* temporarily stops observing, and points towards the Earth to downlink the collected data. The calibration process is shown in the four panels of Fig. 4.7, while the full calibrated light curve is presented in the top panel of Fig. 4.8. The increase of $\sim 50\%$ in the optical flux, between BJD 2458742 and (roughly) BJD 2458751, may be caused by a temporary increase in the mass-transfer rate. The physical reasons behind it are still uncertain. In the bottom-left panel of Fig. 4.8 the light curve is zoomed-in around the first ~ 8 days of observation, in order to better highlight the eclipses of this binary system. Moreover, the bottom-right panel presents a further reduced portion of the light curve (zoomed-in around almost two days of observation) is displayed, in order to better

[5] See Chapter 3.
[6] http://archive.stsci.edu/

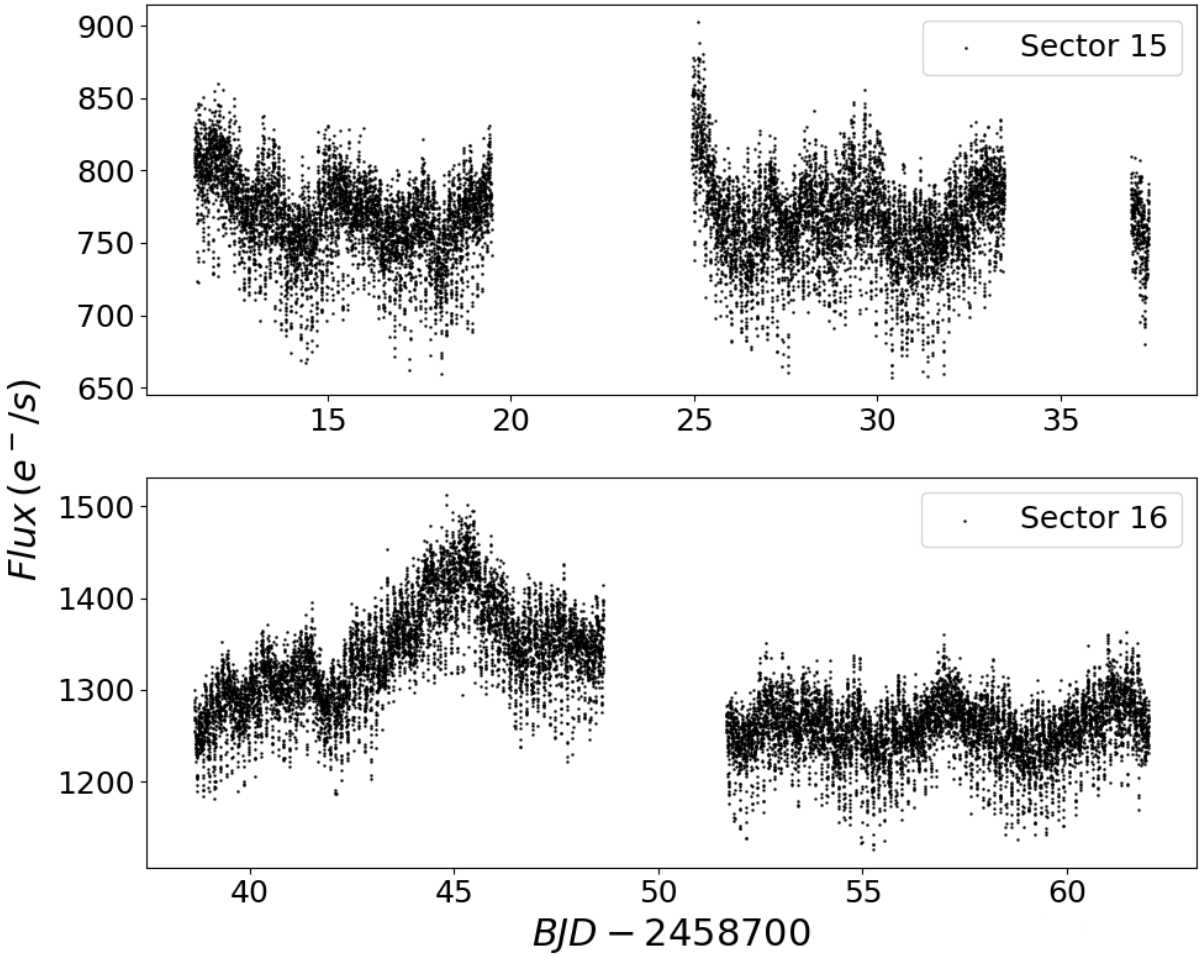


Figure 4.6: Uncalibrated light curve of HBHA 4204-09, observed by *TESS* during Sector 15 (top panel) and Sector 16 (bottom panel), with a cadence of 120 s.

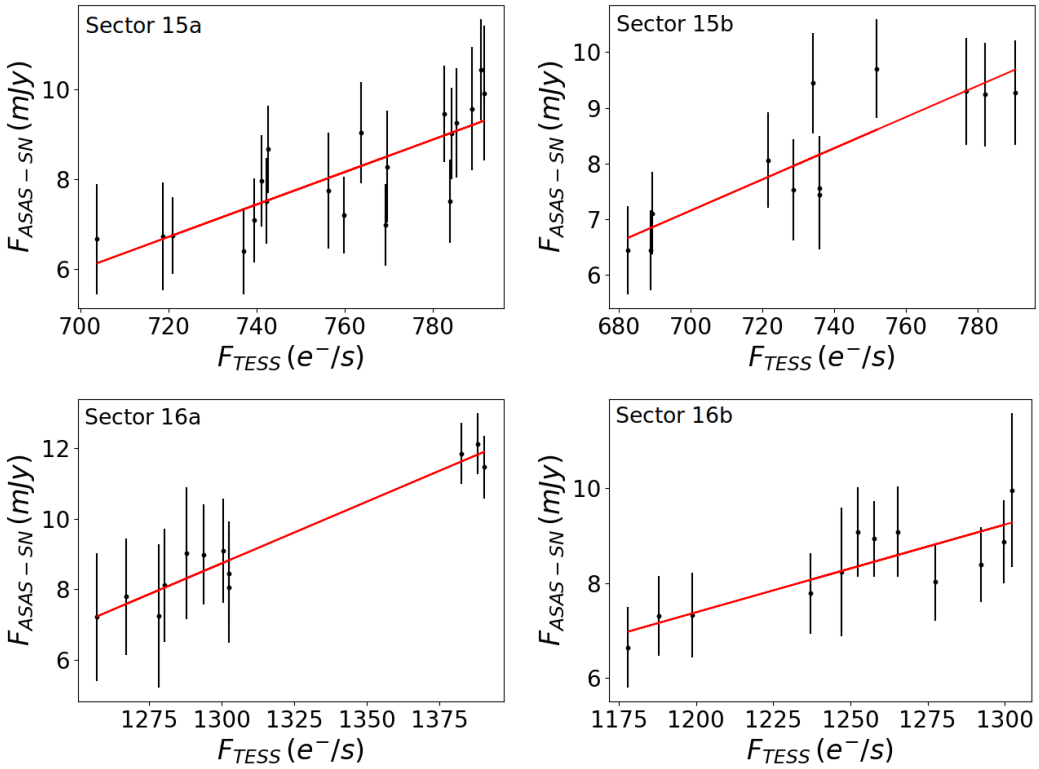


Figure 4.7: Calibration of *TESS* light curve, divided in four batches of equal duration, against ASAS-SN data. The top row describes the calibration process of *TESS* light curve observed during Sector 15, while the calibration of the light curve from Sector 16 is shown in the bottom row.

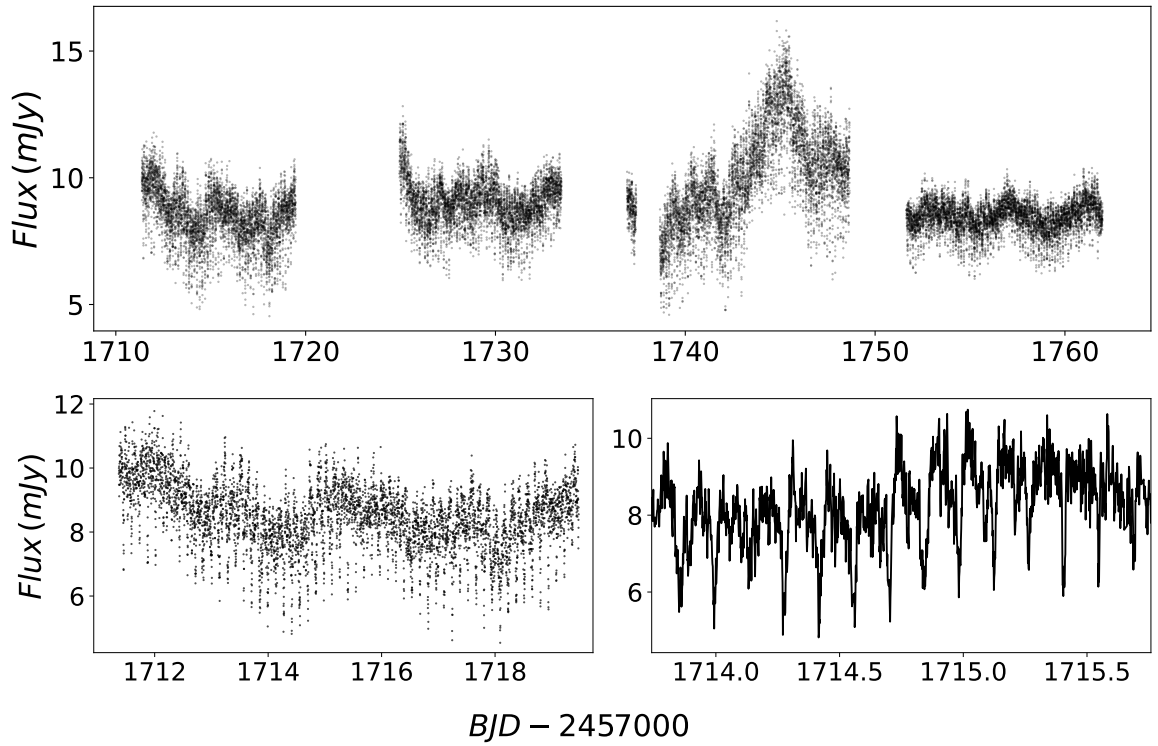


Figure 4.8: Top panel: light curve of HBHA 4204-09, observed by *TESS* during sectors 15 and 16, and calibrated against ASAS-SN data. Bottom panels: zoom in of the first ~ 8 days of observation (left panel), and of two randomly selected days of observation.

define the shapes of the profiles of the eclipses.

The periodicities of this object are identified through the analysis of its Lomb-Scargle periodogram (Lomb 1976; Scargle 1982), shown in Fig. 4.9. In the top panel, the power density spectrum is presented in a linear space, while in the bottom panel, it is displayed in the log-log space. The most prominent signal comes from the orbital frequency (Ω), found at $\Omega = 7.078\,\mathrm{c/d}$ (cycles/day), and corresponding to an orbital period of $\mathrm{P_{orb}}$=0.1412896 days ($\sim 3.39\,\mathrm{hr}$). Just above the period gap, this is a typical value for nova-like variables (usually, these objects are characterised by orbital periods included between three and four hours). On the other hand, the fraction of dwarf novae in that specific period range drops significantly (Knigge et al., 2011). Complementing the spectral information, this orbital period seems to provide a stronger indication towards a nova-like classification for HBHA 4204-09. Harmonics of Ω up to the 7th order, deriving from the highly non-sinusoidal signal, are visible in the power spectrum. The periodogram of the full light curve, zoomed-in around the orbital frequency is shown in the top row of Fig. 4.10. For reference, the central and bottom panels show the zoomed-in periodograms of Sector 15 and Sector 16, respectively.

Another signal, associated to a *negative* (or *nodal*) *superhump* (Wood et al. 2000, 2009) is

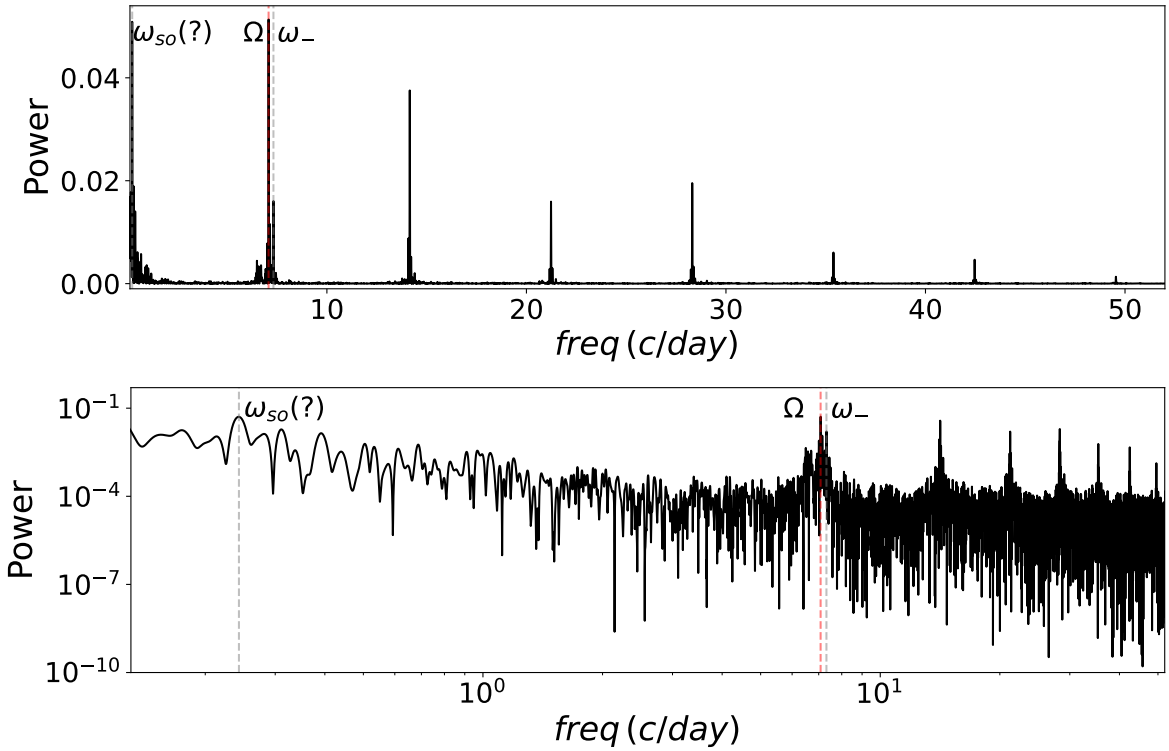


Figure 4.9: Lomb-Scargle periodogram of the full light curve of HBHA 4204-09 in linear space (top panel) and in log-log space (bottom panel). The signals associated to the orbital frequency (Ω), the negative superhump (ω_-) frequency and (possibly) the superorbital frequency ($\omega_{so}(?)$) are labelled in the two panels.

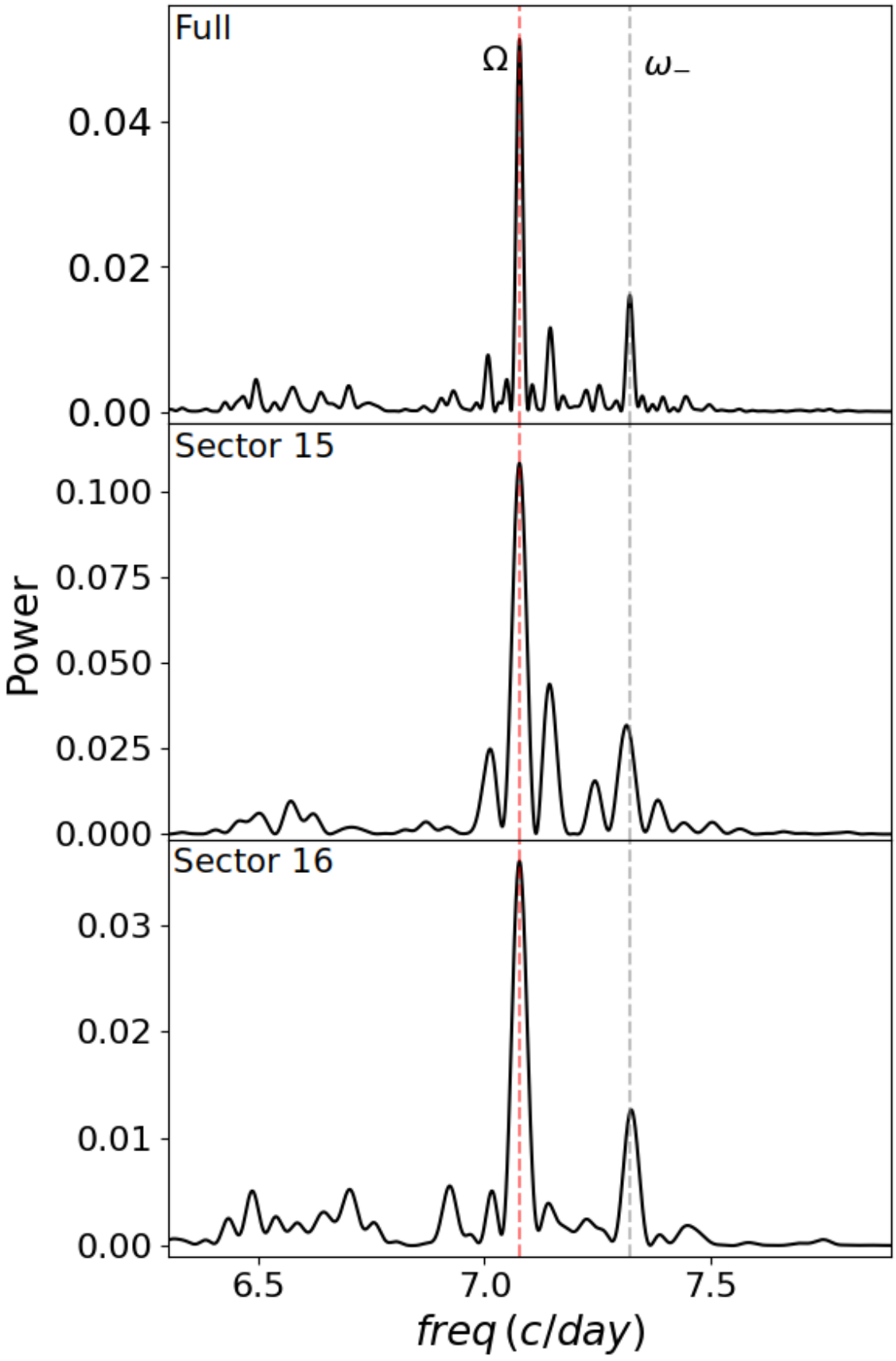


Figure 4.10: Periodogram of the full light curve of HBHA 4204-09 (top panel), zoomed-in around the orbital frequency. The middle and bottom panels display the periodograms, zoomed-in around the orbital frequency, for Sector 15 and 16, respectively.

detected at an immediate faster frequency: $\omega_- = 7.321\,\mathrm{c/d}$. Negative superhump frequencies are usually a few percent greater than the orbital frequency of the system, and are associated to the presence of a retrogradely precessing tilted accretion disk. Negative superhumps are often accompanied by a *superorbital signal*, the frequency (ω_{so}) of which is equal to the difference between the negative superhump frequency and the orbital one: $\omega_{so} = \omega_- - \Omega = 0.243\,\mathrm{c/d}$. This signal gives the duration of one precession of the tilted disk. The periodogram shown in the top panel of Fig. 4.9 shows a peak in correspondence of the expected superorbital frequency. However, the low significance of said peak does not provide a clear evidence to discern whether this signal is to be associated to a coherent superorbital frequency, or to red noise.

Positive superhump signals are often detected in CVs, too, with frequencies ω_+ a few percent smaller than the orbital one. Positive superhumps are described as the beat frequencies between the orbital period and the prograde precession of an elongated (to the 3:1 resonance value) accretion disk. This elongation derives, in turn, from tidal instabilities in the disk. The necessary conditions for the production of negative and positive superhumps signals are independent. Therefore, they can be simultaneously satisfied, and a CV can present both signals together. An example of CV that shows both negative and positive superhumps is AQ Men (Iłkiewicz et al., 2021). A set of signals in the frequency range 6.4-6.8 c/d stands out in the periodogram of HBHA 4204-09, in both linear and log-log space. These signals persist (as all the other ones shown in Fig. 4.9) even if the Lomb-Scargle periodograms of the two Sectors are analysed singularly (middle and bottom rows of Fig. 4.10). One of these frequencies might be associated to a positive superhump, but further analyses are necessary to assert this with more confidence.

A more precise timing analysis is enabled by the presence of (grazing?) eclipses in the light curve of the target being studied. Fig. 4.11 shows the flux of HBHA 4204-09 from Sector 15 (top panel), folded on the orbital period. Data from Sector 16 are characterised by a larger noise, due to the increase in the luminosity of the system (as can be seen from the bottom panel in the same figure). The folded light curves are binned with 100 equally spaced bins to produce the average shape outlined by the red lines. The first step to calculate the uncertainty that affects the orbital period, as well as that on the epoch of the first observed eclipse minimum, consists in fitting a spline to each orbital cycle, and registering the phase of eclipse minimum (Φ_{min}). The distribution of Φ_{min} as a function of the cycle number (n) is detrended via the subtraction of the best-fit linear model that describes the data. A graphical depiction of this process in shown in the top panel of Fig. 4.12. Some groups of

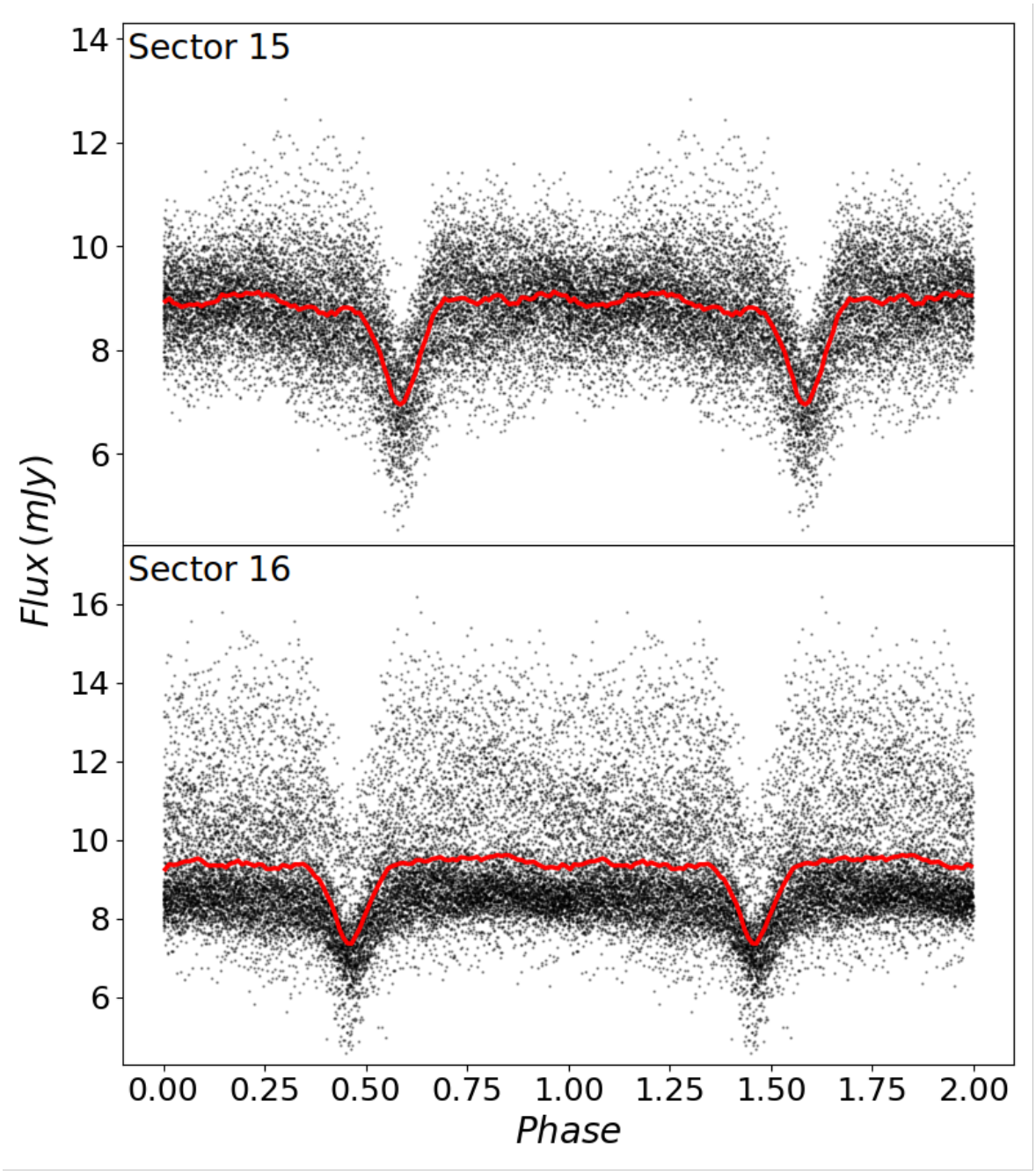


Figure 4.11: The light curve of HBHA 4204-09, limited to Sector 15, is folded on the orbital period in the top panel. On the other hand, the the light curve relative to Sector 16 is folded on the orbital period in the bottom panel. The red lines in the two panels results from binning the phase space in 100 equally spaced bins, and calculating the average flux in each bin.

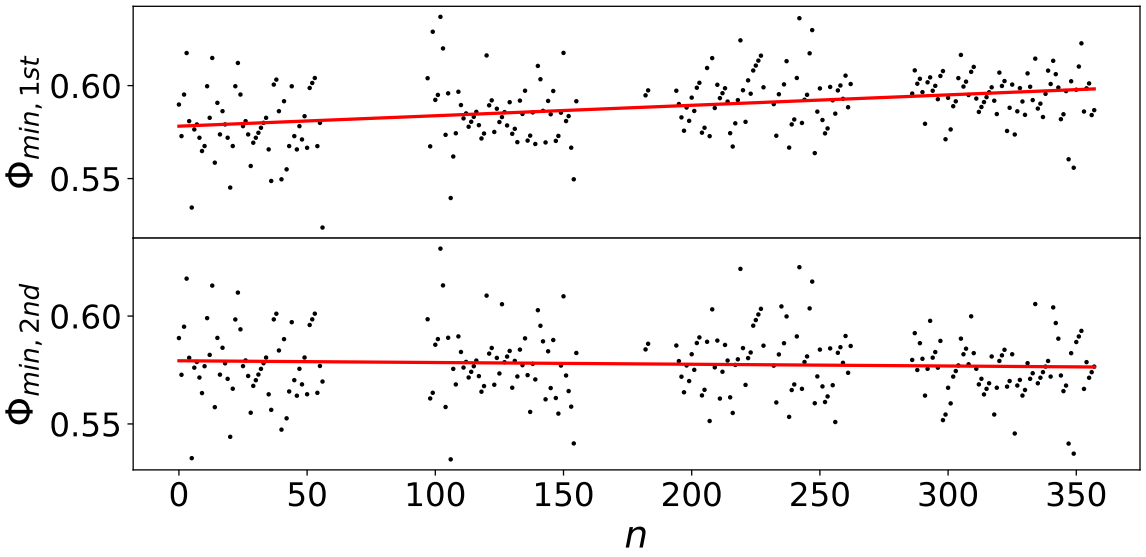


Figure 4.12: Graphical depiction of the process for: a) the refinement of the measured orbital period of HBHA 4204-09, and b) for the calculation of the uncertainties on the epoch of first minimum, and on P_{orb}. The top panel shows the phases of minimum flux, measured with the use of P_{orb} obtained from the Lomb-Scargle periodogram, as a function of the cycle number. The best fit model, used to correct this value of P_{orb}, is depicted by the red line. The bottom panel shows the new phases of minimum flux, obtained with the use of the refined P_{orb}, as a function of the cycle number.

adjacent points are aligned. This trend, particularly noticeable for example for $28 \leq n \leq 34$, is presumably due to the interference of the superorbital frequency. However, further analyses to confirm this hypothesis are required. The slope (5.62×10^{-5}) of the linear model (depicted by the red line) is used to refine the orbital period (new P_{orb}=0.1412975 days) of the system, while error on said slope corresponds to the uncertainty on P_{orb}. After that the phases Φ_{min} are recalculated with the corrected orbital period (shown in the bottom panel of Fig. 4.12), their standard deviation is calculated, to provide the uncertainty on the first epoch of eclipse minimum. The slope of the new best-fit model (-8.07×10^{-6}) is affected by a measurement error that sets the uncertainty on the refined value of P_{orb}. This process, described more in detail in Scaringi et al. (2013a), provides the following eclipse-minimum ephemeris:

$$BJD_{\rm min} = 2458711.4514(22) + 0.1412975(12) \cdot n. \tag{4.3.1}$$

Chapter 5

Future work and conclusions

In this thesis, I discussed the benefits introduced by a novel CMD-based technique, as a complement to classic position-based selections of Hα-excess candidates. The inclusion of information regarding their parent stellar population, provided by the availability of the accurate *Gaia* astrometry, enabled the production of a more complete representation of the Hα-excess sources in the northern Galactic plane.

In the following Section, I provide an example on how our photometric population discrimination can be further improved with the introduction of additional parameter spaces. This is also put in the context of the forthcoming availability of large spectroscopic surveys. In the final Section, I provide a full summary of the content of the previous Chapters of this thesis.

5.1 Hα-excess candidates in XGAPS catalogue

With the arrival of multi-wavelength narrow-band surveys such as J-PAS, the potential of the CMD-based Hα-excess populations identification could theoretically be limited only by the available computational power. Here, I illustrate a preliminary study about how our population-based selection can be improved with the introduction of additional colours, by implementing it on the recently available XGAPS catalogue (Scaringi et al. (2022), submitted).

5.1.1 The XGAPS catalogue

The XGAPS (Cross-match of Galactic Plane Surveys) catalogue is the result of a sub-arcsecond cross-match of the *Gaia* DR3 catalogue against the IGAPS catalogue, first, and

then against the United Kingdom Infrared Deep Sky Survey (UKIDSS, Lawrence et al. 2007) catalogue[1]. Besides the *Gaia* astrometric measurements, XGAPS provides photometric data in the (*Gaia*) G, G_{BP} and G_{RP}, (UKIDSS) J, H, K, (UVEX) U, g, (IPHAS) r, i, and Hα bands, for almost 34 millions objects in the Galactic plane. UVEX and IPHAS photometries are leveraged from the IGAPS catalogue.

The sources in the XGAPS catalogue are included in the $20^\circ < l < 220^\circ$ and $-6^\circ < b < 6^\circ$ coordinate ranges. The constraint on the parallax measurements (`parallax_over_error > 3`) ensures that objects within a 1.5 kpc distance radius are well represented. Moreover, in order to ensure the quality of *Gaia* photometry, an additional signal-to-noise constraint for the G-band flux (`phot_g_mean_flux_over_error > 3`) is applied.

Before performing the various cross-matches, *Gaia* DR3 coordinates are rewound to the observation epochs of the different surveys, in order to reduce the number of spurious mismatches. The procedure followed here is analogous to that applied for the production of the *Gaia*/IPHAS catalogue, described in Section 2.1.1. However, since the epochs of observation of IPHAS and UVEX differ substantially (UVEX started operating roughly three years after IPHAS), these two components of the IGAPS catalogue are considered separately, at this stage. While 33,987,180 objects[2] in *Gaia* DR3 are identified to have a counterpart in IGAPS (either in IPHAS or in UVEX), 32,138,484 of them have detections in both IPHAS and UVEX. The sources in this list are then further cross-matched with UKIDSS, upon the same coordinates rewinding procedure. This process yields 21,240,381 non-duplicate matches.

The top panel of Fig. 5.1 shows all the $\sim$ 34 million objects with IGAPS measurements, located in the *Gaia* CMD. Some spurious astrometric solutions from *Gaia* are not unexpected (Scaringi et al., 2018; Gaia Collaboration et al., 2021), especially when *Gaia* observes particularly crowded regions of the sky (such as the Galactic plane). An evident example of this effect can be noticed in the region of the *Gaia* CMD between the MS and the WD tracks, where a considerably lower density of objects is usually expected. To exclude the objects affected by spurious astrometric solutions, the targets in the XGAPS catalogue are classified with the application of a Random Forest machine learning technique (Breiman, 2001), that involves 26 astrometry-based predictors, and 1001 decision trees. To train the classifiers, the $\sim$ 20 million sources with `parallax_over_error > 5` are considered to have a good astrometric solution. On the other hand, the bad astrometry set includes the mirror-sample

[1] However, for the purposes of this Section, I only focus on the photometries from *Gaia* and IGAPS.
[2] This number results from the removal of duplicate matches

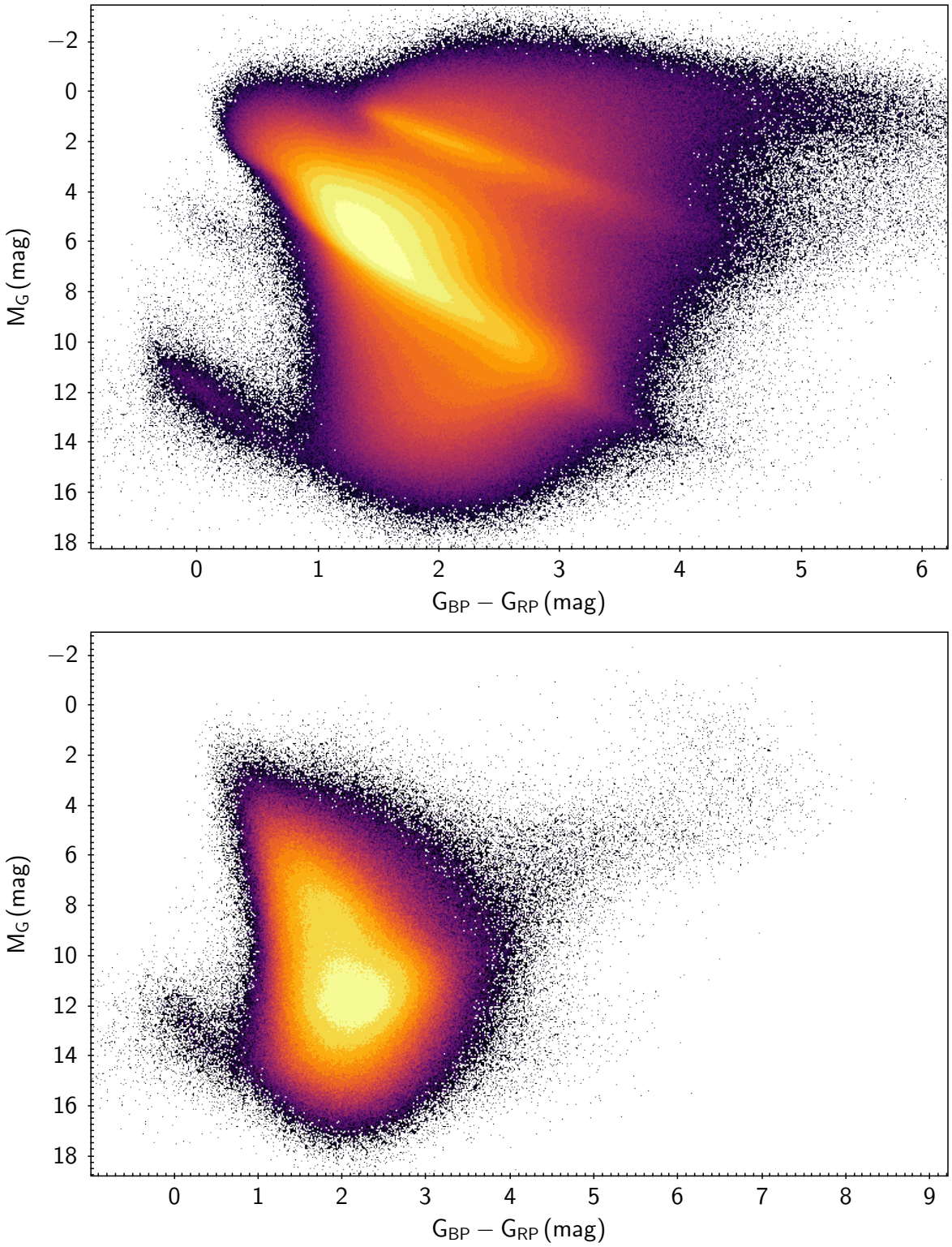


Figure 5.1: Top panel: location in the *Gaia* CMD of all the sources resulting from the cross-match between the *Gaia* DR3 and IGAPS catalogues. Bottom panel: subset of targets identified with a spurious astrometric solution by more than half of the Random Forest classifiers.

of $\sim 250,000$ objects with `parallax_over_error < -5`. The testing set is built by randomly selecting $\sim 50,000$ sources (20%) from the bad-astrometry sample, and the same number of objects from the good-astrometry set. The training set is constituted by the objects with good and bad astrometric solutions, with the exclusion of the testing sets. After the classifier algorithm is trained, it is applied to all the ~ 34 million sources in XGAPS, and a flag (`flagRF`) is assigned to to each of them. If an object is classified as genuine by more than half of the decision trees, the `flagRF` is equal to 1, otherwise it is equal to 0. The objects with `flagRF=0` are located in the *Gaia* CMD in the bottom panel of Fig. 5.1. While the vast majority of these targets cluster between the MS and WD tracks, many of them are also found on the WD track and among the objects affected by the highest extinction.

5.1.2 Hα-excess selection

The selection of Hα-excess candidates is performed on the subset of the XGAPS catalogue with valid measurements in the G, G_{BP}, G_{RP}, U, g, r, i and Hα bands, and with `flagRF=1`. The addition of UVEX colours in the partitioning procedure enables a more robust population separation, with respect to that obtained by grouping together nearby objects in the *Gaia* CMD, standalone. However, this comes at the expenses of an increased computational effort. The distributions in the *Gaia* and UVEX CMDs, and in the IPHAS CCD of the 22,455,168 non-duplicated sources that satisfy these criteria are presented in Fig. 5.2 (top to bottom panels). While the grey dots represent all the objects included in this analysis, the black dots in the top panel depict the excluded sources. As for the *Gaia*/IPHAS catalogue, the absolute M_G and M_U magnitudes are obtained with the use of *Gaia* parallaxes[3], with the assumption of zero extinction, and are therefore overestimated.

The selection of Hα-excess candidates is performed by following a similar process to that described in Chapter 2. However, the current analysis focuses on the CMD-based selection, only. The main improvement introduced here is in the partitioning process, that is performed simultaneously in the UVEX and *Gaia* CMDs. In fact, to constitute a partition, the targets have to simultaneously occupy similar locations in both the inspected parameter spaces. The full partitioning process yields 5,470 partitions. The least populated partition includes 528 objects, while the densest one contains 66,618 sources. The detrending and identification of outliers procedures are the same as those illustrated in Section 2.4, including the definition

[3]However, *Gaia* DR3 parallaxes were not available when the *Gaia*/IPHAS catalogue catalogue was generated, and the parallaxes from the DR2 were thus used.

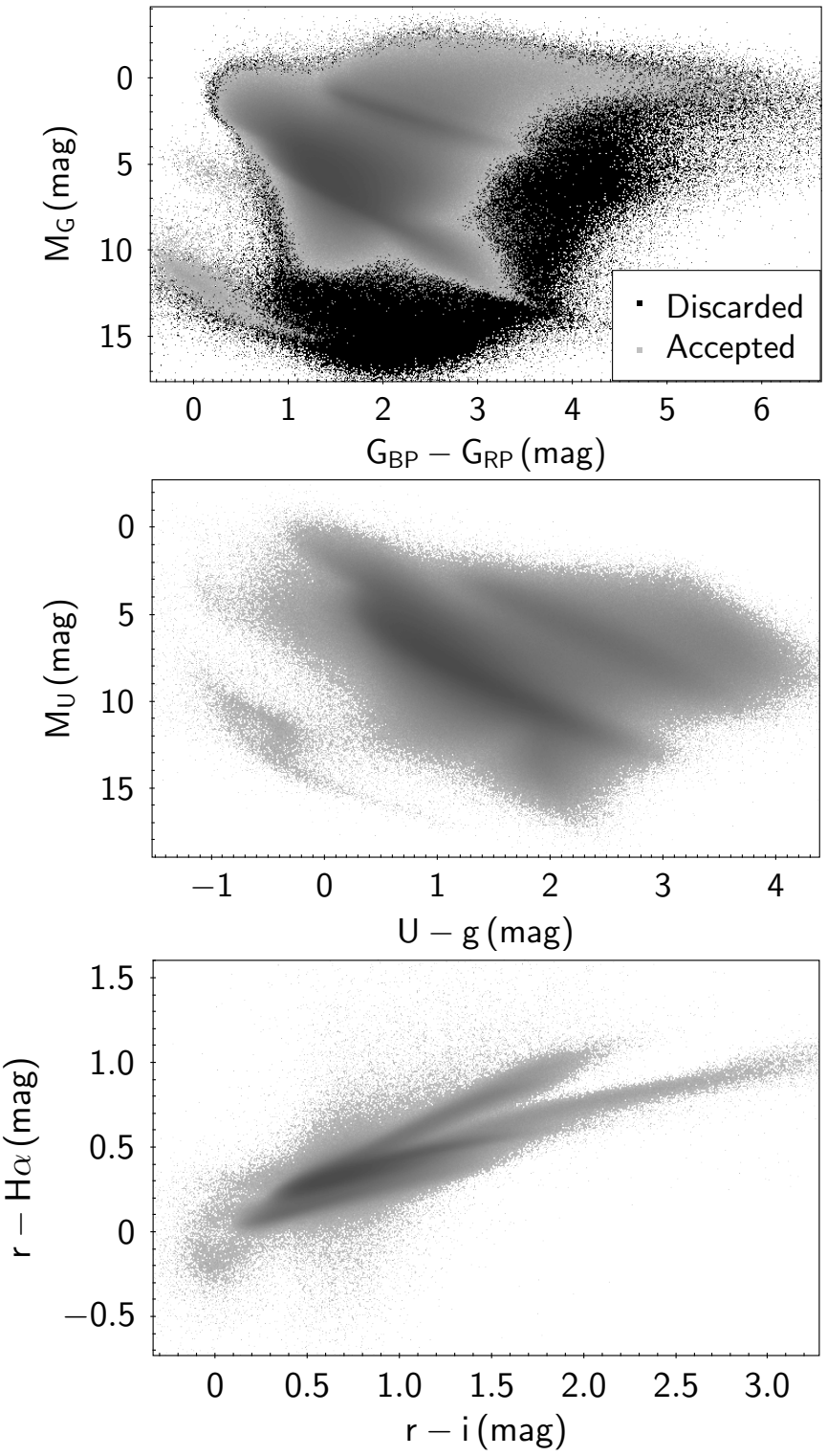


Figure 5.2: Locations in the *Gaia* CMD (top panel), in the UVEX CMD (central panel) and in the IPHAS CCD (bottom panel) of all the non-duplicated sources in the XGAPS that present valid measurements in all the *Gaia* and IGAPS bands of interest, and with a good astrometric solution (the grey dots). The black dots represent the objects in the XGAPS catalogue that do not satisfy all these criteria.

of the "significance".

The whole selection process yields 96,218 3σ outliers, and 24,550 5σ outliers. The percentages of Hα-excess candidates identified in XGAPS are similar to those selected from the *Gaia*/IPHAS catalogue, for both the significance thresholds: 0.43% against the 0.36% for the 3σ outliers, and 0.11% against the 0.09% for the 5σ ones. These are represented in the *Gaia* CMD (top panel), in the UVEX CMD (central panel), and in the IPHAS CCD (bottom panel) of Fig. 5.3 as the red and blue dots, respectively. Their locations in the two considered CMDs provide a strong indication about their classification. Consistently with what discussed in Sections 2.5 and 3.3, the highest density of outliers selected by this algorithm is in the regions of the CMDs associated to YSOs and RGs, but many accreting WD candidates can be visually identified between the MS and the WD tracks.

The full potential of this improved outliers identification is to be found in the context of the pre-selection of promising Hα-excess candidates from large spectroscopic surveys. While all these outliers can be potentially observed by the forthcoming WEAVE survey, 8,105 3σ outliers and 2,365 5σ outliers lie in the $-5^\circ < \mathrm{Dec} < 6^\circ$ coordinate range, that partially overlap with 4MOST field of view. These latter are distributed in the *Gaia* CMD and in the UVEX CMD as shown in the two panels (top and bottom, respectively) of Fig. 5.4. These new pieces of information constitute a valuable support to have a more comprehensive view on the varied Hα-emitting populations that form our Galaxy.

5.2 Summary

In Chapter 2, I outlined the application of the novel CMD-based Hα-excess selection on the *Gaia*/IPHAS catalogue, the generation of which is presented in Scaringi et al. (2018). I inspected 7,474,835 targets for Hα-excess, 28,496 out of which were identified as outliers. After describing each step of the selection algorithm, I showed the improvements that result from integrating classic position-based identifications with this novel technique. To provide a preliminary validation of this method, I compared the obtained list of Hα-excess candidates with those resulting from the studies presented by Witham et al. (2008) and Monguió et al. (2020). Moreover, I calculated the statistics of the selection of Hα-excess candidates (*purity* and *completeness*) with the inspection of randomly selected spectra from LAMOST DR5. On the one hand, the purity of the 3σ selection is 48.9%. This parameter increases to 81.9%, if only 5σ candidates are considered, and if objects close to the saturation limits of IPHAS

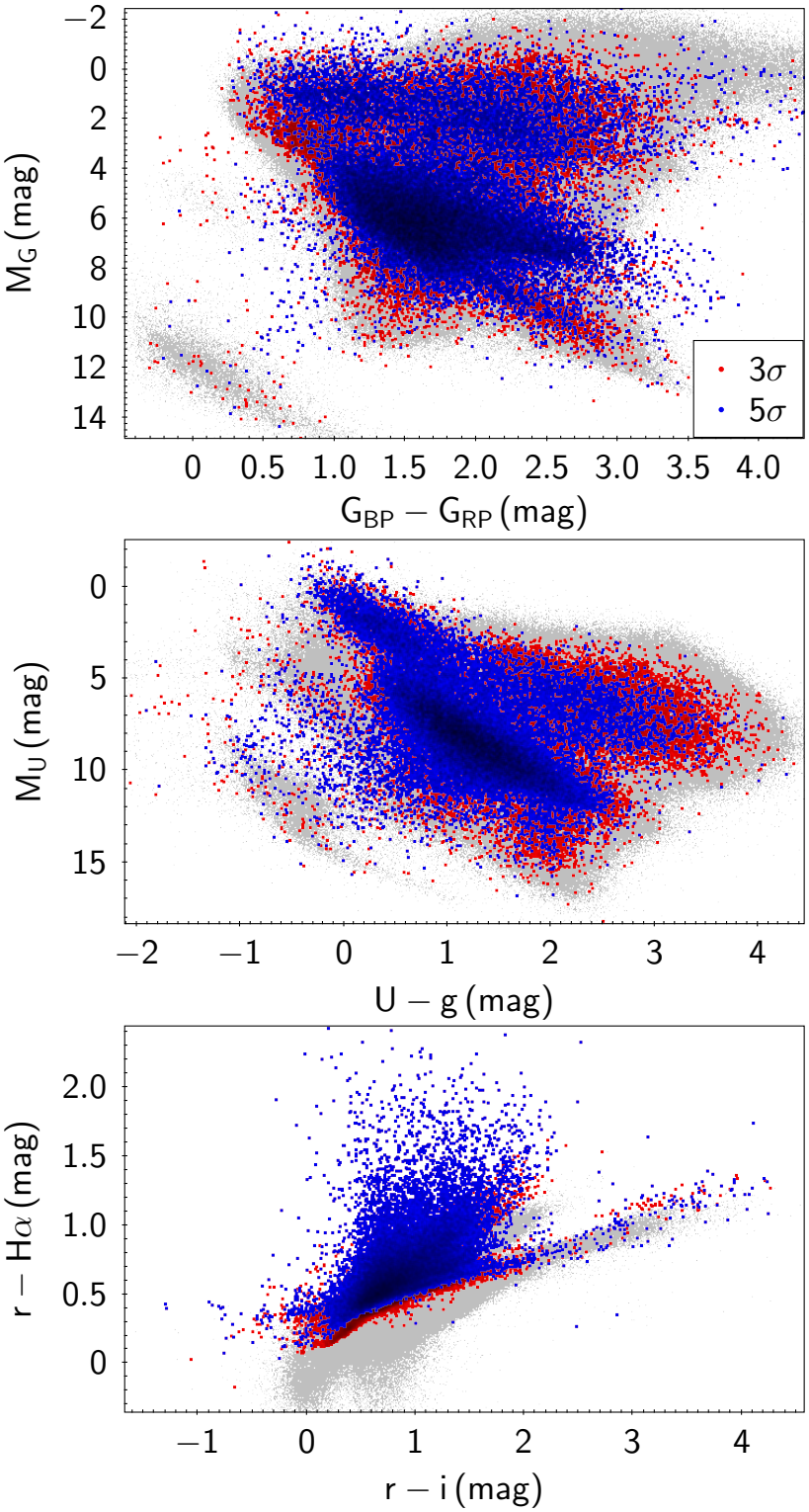


Figure 5.3: Locations in the *Gaia* CMD (top panel), UVEX CMD (central panel) and IPHAS CCD (bottom panel) of the Hα-excess candidates identified in the XGAPS catalogue. The red dots are associated to a significance higher than 3, while the blue ones have a significance higher than 5.

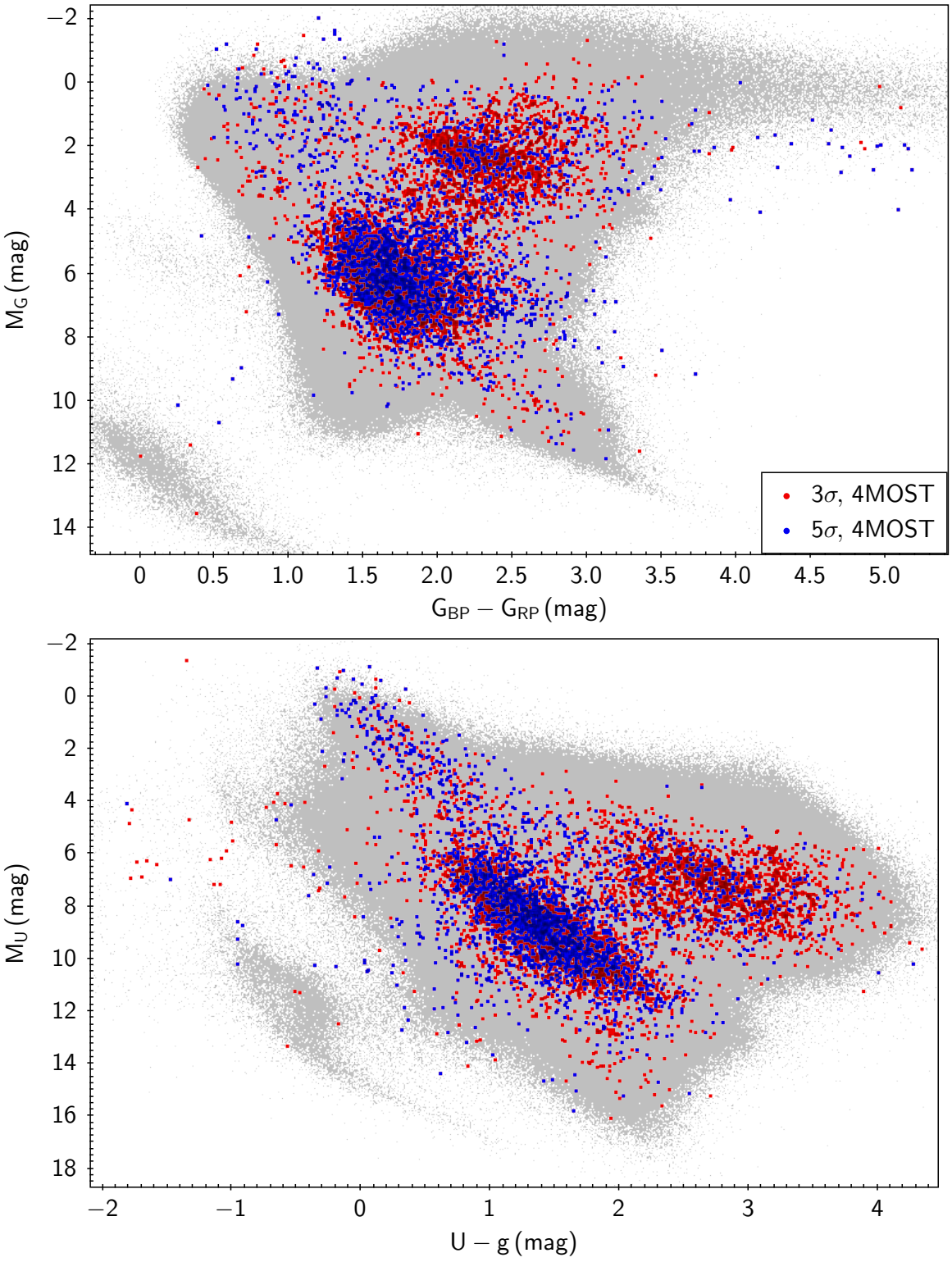


Figure 5.4: Location in the *Gaia* CMD (top panel) and in the UVEX CMD (bottom panel) of the Hα-excess sources in the XGAPS catalogue, included in the -5° < Dec < 6° coordinates range. This partially overlaps with 4MOST field of view. The grey dots represent all the objects in the XGAPS that were inspected for Hα-excess. The red dots are the outliers selected with a significance greater than 3, while the blue ones are associated to a significance larger than 5.

($13 \leq r < 13.5$ mag, $12 \leq i < 12.5$ mag and $12.5 \leq$ H$\alpha <$ 13 mag) are excluded. On the other hand, the combination of position-based and CMD-based selections yields a completeness included between 3% and 5%. As a reference, the same calculation performed on the IGAPS catalogue yields a completeness of 1%.

Chapter 3 presents a further spectroscopic validation of the CMD-based selection, based on a set of follow-up spectroscopic observations from Gemini facility. The 114 analysed spectra were drawn from a wider group of objects, that was specifically designed to provide a homogeneous representation of the brightest ($r < 17.5$ mag) and most promising ($n_\sigma \geq 5$) Hα-excess candidates in the *Gaia*/IPHAS catalogue. The spectroscopic classification of these sources enabled the quantification of the efficiency of the selection, on a population-by-population basis. In addition, this classification was used as a reference to illustrate how the location in the *Gaia* CMD can provide a reliable classification for Hα-excess candidates, especially if augmented with photometry in other bands (UVEX and *WISE* bands, in this case). This spectroscopic analysis seems to suggest that the most significant improvement introduced by the CMD-based selection is in the detection of Hα-excess candidates among the YSOs (14 more than the position-based selection). However, the CMD-based technique also enabled the identification and classification of previously unknown line-emitters belonging to other stellar populations, such as 11 CVs and 10 CBe stars. On the other hand, the number of spurious detections included in the population-based selection could be reduced with the introduction of Machine Learning clustering algorithms, for example. However, this task is left for future works.

In Chapter 4, information regarding X-ray emission and high variability levels are leveraged to identify new accreting WDs among the Hα-excess candidates. An example of these sources is the highly variable Hα-excess candidate HBHA 4204-09, that is also part of the spectroscopic analysis described in Chapter 3. While previously identified as a hot subdwarf star (Geier et al., 2019), the spectro-photometric analysis performed in this work locates this object among the CVs. A timing analysis of its short cadence light curve (observed by *TESS*) yielded an orbital period of 3.39 hr. This value supported a further classification of this target as a nova-like variable.

Finally, in Section 5.1 I shortly outlined one of the possible ways to improve the population-based selection of Hα-excess candidates. This consists in the introduction of additional colours (obtained from the XGAPS catalogue) in the partitioning process, to provide a more robust population discrimination.

Appendix A

Gemini spectra and classification

Table A.1: Results of our spectroscopic classification. For each target, the following entries are provided: *Gaia* DR2 SourceID; the *Gaia* DR2 barycentric Right Ascension and Declination, at epoch 2015.5; the distance, obtained inverting *Gaia* DR2 parallax; the apparent magnitude in IPHAS r band; the maximum significance (between the two provided by FR21) of being an Hα-excess source; the FWHM of the Hα line ("NA", if said line is seen in absorption); our spectral classification.

SourceID (*Gaia DR2*)	**RA** (°)	**Dec** (°)	**Dist** (kpc)	r (mag)	n_σ	**FWHM** (Å)	**Class**
183806256466977920	79.54732	35.13356	0.91(5)	16.35	13.56	19(1)	CV
184238295817134208	82.51049	36.82756	0.62(2)	14.69	6.42	NA	MS
185881786526045440	74.66526	35.09533	1.2(1)	16.77	8.49	14.2(6)	CV
188400875040872832	76.02748	38.72527	2.8(3)	13.66	5.76	9.14(7)	CBe
202583509889708800	77.95731	44.28475	0.58(3)	16.62	12.50	NA	MS
207006089252554240	75.83720	47.08507	1.3(2)	17.22	9.14	18.3(2)	CV
250899658585804544	59.41636	51.61552	0.506(7)	13.60	9.38	NA	MS
251693368542860288	59.30778	51.99074	1.19(6)	15.61	11.32	10.15(8)	YSO
258089193320478208	67.22134	48.52566	1.9(1)	15.44	7.74	11.17(4)	CBe
259018688664493312	71.64038	48.96539	0.56(2)	16.56	15.95	20.30(1)	CV
276781475113802496	62.89925	55.87940	4.7(7)	14.82	7.7	10.15(7)	CBe
422750061837839232	0.93751	58.46850	1.0(1)	17.29	6.67	NA	MS
424396924098269312	13.42935	57.96660	4.4(4)	13.86	10.47	NA	RG
425129194536945024	10.71283	58.32971	0.97(4)	16.51	14.89	NA	MS
425995304767655040	13.37703	59.19563	3.6(3)	13.83	13.05	NA	RG
427292453608805504	13.32775	60.77675	0.82(1)	14.02	12.10	NA	MS
427316058750043520	13.55209	60.94371	3.5(3)	13.73	7.32	NA	RG
427473838665897984	13.38334	61.84135	0.339(3)	14.24	7.17	NA	MS

SourceID (*Gaia DR2*)	**RA** (°)	**Dec** (°)	**Dist** (kpc)	r (mag)	n_σ	**FWHM** (Å)	**Class**
428233081505186688	5.22420	59.16597	0.98(5)	16.91	6.30	NA	MS
463122170299415168	45.27165	60.53203	2.4(3)	15.84	10.45	9.14(1)	YSO
464931073088436096	41.95254	60.96403	2.3(1)	13.95	14.50	7.11(3)	CBe
465712207381733888	41.97116	61.56827	6(1)	14.37	9.23	7.11(3)	CBe
510854753206972800	22.63292	62.35899	0.79(5)	16.92	11.45	18.3(2)	CV
512016318518918400	26.73920	63.00894	6(1)	15.64	16.33	13.20(4)	CBe
514896283072146560	30.58104	63.37740	0.277(7)	16.78	8.10	8.1(1)	M-Dwarf
522539950472429312	17.01763	61.66252	0.478(4)	13.63	6.29	NA	MS
522542802330612608	16.68316	61.54304	6(1)	13.95	7.01	NA	RG
522841835128697472	16.42042	62.10566	3.0(2)	13.94	6.43	NA	MS
524238661574474624	13.60656	64.62303	2.10(7)	13.73	10.66	NA	MS
524301986569104768	14.93122	64.91167	0.76(4)	17.26	15.22	25.4(6)	CV
526118620296454528	13.55275	66.29011	0.657(8)	13.92	7.58	NA	MS
527026954341506688	8.95623	64.25796	3.7(6)	16.44	15.88	NA	RG
1823343535657240064	298.95041	20.02398	5.1(8)	14.30	7.58	NA	RG
1833633385113024896	300.48773	23.60346	1.8(2)	17.04	7.69	17.1(4)	CV
1970080784065021184	317.62957	43.13736	1.8(2)	16.69	7.79	NA	MS
1971264992446842368	321.13415	45.06151	1.2(1)	17.02	11.38	25.4(5)	CV
1971760631674787072	318.70753	44.74772	2.3(4)	16.72	7.63	14.2(4)	CV
1978346259653262976	322.95547	48.06728	0.80(3)	16.0	10.81	8.12(5)	YSO
1980077092801857664	331.73643	51.09846	1.09(9)	16.68	22.70	23.3(2)	CV
1997162958039726080	348.79460	55.95792	2.5(3)	15.79	8.97	NA	MS
2007611926287679616	337.32027	57.08722	4.4(4)	13.90	9.17	11.17(4)	CBe
2015167735817519360	345.94170	61.81282	0.75(2)	14.67	13.17	8.12(3)	YSO
2017612358184192768	359.83326	66.38672	0.79(2)	14.63	15.15	7.11(1)	YSO
2018379134410850688	294.93998	22.08596	0.63(2)	16.64	6.96	8.0(1)	T-Tauri
2021418317687630080	293.75924	24.77452	0.417(9)	16.22	9.75	6.0(1)	M-Dwarf
2022704849371681024	290.84948	23.73654	0.53(1)	15.56	6.63	NA	γ Dor
2024391951234118912	292.20999	24.98576	2.1(3)	16.75	8.75	NA	γ Dor
2030374978109714688	298.23984	30.08825	1.42(5)	15.41	12.01	NA	MS
2033218590053883648	294.44810	31.67799	5(1)	16.04	6.24	NA	MS
2034706057485253376	297.86111	32.94033	1.03(7)	17.20	6.80	NA	MS
2034718873668327680	297.07516	32.64354	6(1)	14.83	5.84	NA	RG
2046886550363859968	295.31150	32.87483	3.8(5)	16.08	8.31	NA	MS
2054712221272378496	304.89408	33.57262	1.19(7)	16.41	5.72	NA	MS

SourceID (*Gaia DR2*)	**RA** (°)	**Dec** (°)	**Dist** (kpc)	r (mag)	n_σ	**FWHM** (Å)	**Class**
2061420410434776192	305.63634	39.28241	0.79(1)	13.81	11.67	NA	MS
2066074436989021440	311.37173	41.05052	3.0(3)	15.43	7.12	12.18(6)	YSO
2071104221657306496	309.20076	45.34096	0.93(3)	15.94	13.20	5.08(3)	YSO
2162336164614235648	316.96769	44.09499	0.475(5)	13.94	10.86	21.32(6)	CV
2162850736074155392	316.39512	46.37367	3.3(4)	15.62	7.37	NA	MS
2162947798009095808	312.86245	44.22094	0.79(4)	16.93	20.98	6.09(5)	YSO
2164154516334658560	318.81208	46.45168	3.9(3)	14.61	8.45	NA	MS
2165559073718735872	316.57904	48.68376	3.3(5)	15.92	13.61	NA	RG
2166302068698392960	311.57080	45.42623	3.7(6)	16.14	9.97	10.15(6)	YSO
2168759236665389312	315.91445	50.26468	0.59(1)	15.69	10.23	7.11(6)	YSO
2170925927411884672	323.05921	49.74432	1.6(2)	17.02	13.46	16.2(2)	CV
2171539832855206784	324.25968	51.53298	0.208(3)	16.85	6.96	7.1(1)	M-Dwarf
2201164799270238976	334.85416	59.55555	3.4(4)	15.07	16.14	8.12(1)	YSO
2207626251145725440	345.36781	62.93661	0.89(2)	14.79	6.35	9.14(3)	YSO
2209703675287141504	356.09811	66.02327	1.40(8)	15.89	8.65	NA	MS
2210266522162289536	350.61836	66.02533	0.74(1)	14.77	21.19	9.14(2)	YSO
3105714495538124288	101.47439	-2.90572	6(1)	14.87	12.27	12.05(3)	CBe
3113703547027886720	104.11159	0.38287	1.4(2)	17.31	11.52	NA	MS
3119336516892417024	99.76717	-0.80271	1.9(3)	16.07	9.99	8.03(3)	YSO
3125759898182046080	102.12067	1.51568	3.7(4)	13.84	7.67	10.04(2)	CBe
3126136064298079616	102.23128	1.59818	4.8(8)	13.9	13.62	8.03(2)	CBe
3126247218051915264	103.83627	2.31710	0.72(6)	16.52	6.70	NA	γ Dor
3126354699614424064	102.17797	2.06890	1.3(2)	16.77	6.91	NA	γ Dor
3129126499706430976	103.45549	4.99618	4.5(7)	14.06	4.97	9.03(2)	CBe
3130722814493379968	100.68126	4.59085	0.275(9)	16.75	9.55	6.0(1)	M-Dwarf
3132911705924538112	102.16806	6.64056	3.9(7)	14.76	8.94	15.06(6)	CBe
3324947321590659456	94.65241	6.94508	0.65(3)	16.38	23.69	9.03(3)	T-Tauri
3326685615112100992	100.31993	9.45838	0.73(2)	14.38	33.26	10.04(3)	T-Tauri
3326734470363384320	100.68635	9.98419	0.76(6)	16.87	5.95	7.0(3)	T-Tauri
3326896854488117632	99.87232	9.72769	0.73(3)	15.92	6.40	7.0(1)	T-Tauri
3326904585429362560	100.10781	9.84931	0.73(6)	16.20	17.18	6.02(3)	T-Tauri
3326938262268528128	100.43208	10.13123	0.70(3)	14.68	6.46	8.0(2)	Orion V
3349903246243514752	88.17739	17.18745	0.77(4)	15.64	6.07	8.03(4)	T-Tauri
3350768218295646336	100.70632	10.18121	0.469(9)	14.01	5.80	NA	γ Dor
3355257352475460480	98.03440	12.48440	6(1)	13.99	8.46	8.03(1)	CBe

SourceID (*Gaia DR2*)	**RA** (°)	**Dec** (°)	**Dist** (kpc)	r (mag)	n_σ	**FWHM** (Å)	**Class**
3356132937390349824	99.12068	14.54367	1.2(2)	16.09	7.86	27.1(3)	CV
3369781587542465280	96.72360	17.79049	5.1(8)	14.37	7.41	9.04(3)	CBe
3372309639651922432	97.35764	19.57940	3.0(3)	15.33	7.34	NA	MS
3376955492893794048	93.22946	22.17039	0.78(4)	15.93	8.35	23.1(2)	CV
3424187076448918272	88.42817	21.87315	3.5(5)	14.41	8.30	12.05(5)	CBe
3444277112392781440	86.47305	29.61278	4.0(4)	13.95	8.16	15.23(3)	CBe
3448388529968087040	85.47474	32.60364	4.7(8)	14.43	9.99	10.15(5)	CBe
3451057044687431296	89.29311	32.65179	2.7(3)	15.30	26.23	NA	MS
3455444577118544768	85.92677	34.93837	4.2(6)	14.75	6.41	Inv[1].	CBe
4259907860230511616	281.70494	-1.61207	3.6(7)	14.06	6.52	9.03(3)	YSO
4260073676017333248	281.49998	-1.21673	3.0(5)	15.12	8.02	11.04(3)	YSO
4261708203134890624	286.48896	-2.63683	1.0(1)	16.34	9.33	19.1(4)	CV
4261769672710353920	285.84803	-2.29662	0.227(3)	16.48	14.9	28.1(2)	CV
4265284712670427264	282.72070	-1.15056	3.3(7)	16.78	13.38	NA	RG
4273688795363205760	277.83036	1.11545	0.452(7)	14.61	7.25	8.03(3)	YSO
4278629893245072384	282.28770	1.42557	0.58(1)	14.66	6.41	10.04(7)	YSO
4281419427268084480	286.68511	4.66668	0.71(3)	15.40	9.3	10.04(7)	YSO
4282917305693991936	284.19650	7.01528	0.40(2)	17.12	6.32	26.1(2)	CV
4283915662276175104	279.88333	4.99783	3.1(5)	15.65	10.57	NA	RG
4305857275780220032	286.25677	6.40762	2.3(4)	16.94	5.98	9.04(6)	YSO
4306883326285609728	284.75531	7.12050	1.12(6)	14.66	10.58	NA	γ Dor
4309797032104899328	288.96905	11.20355	1.9(1)	13.88	13.53	12.05(6)	YSO
4311951731334685440	283.65170	10.57241	0.98(3)	14.32	7.16	NA	MS
4311967365016126976	283.85071	10.91176	1.8(1)	15.52	6.72	NA	γ Dor
4316059300586640256	291.39037	12.68881	3.4(6)	15.05	6.23	16.06(9)	YSO
4514006371833731456	286.21437	17.35407	1.8(2)	16.62	9.68	NA	γ Dor

[1]The spectrum of this object presents Hα-inversion, as shown in Fig. 3.4.

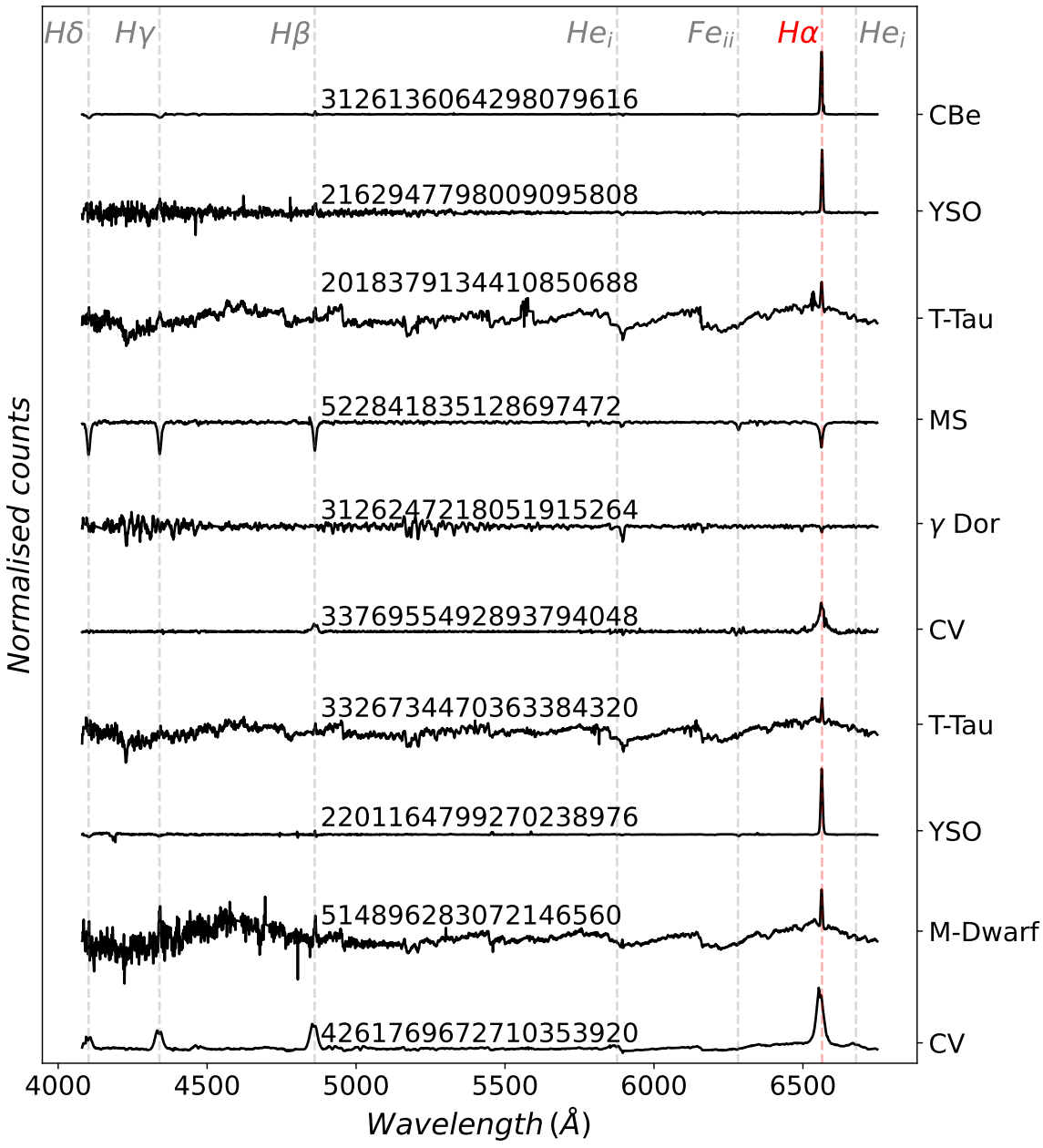


Figure A.1: subset 1-10/114 of inspected spectra, normalised for visual purposes.

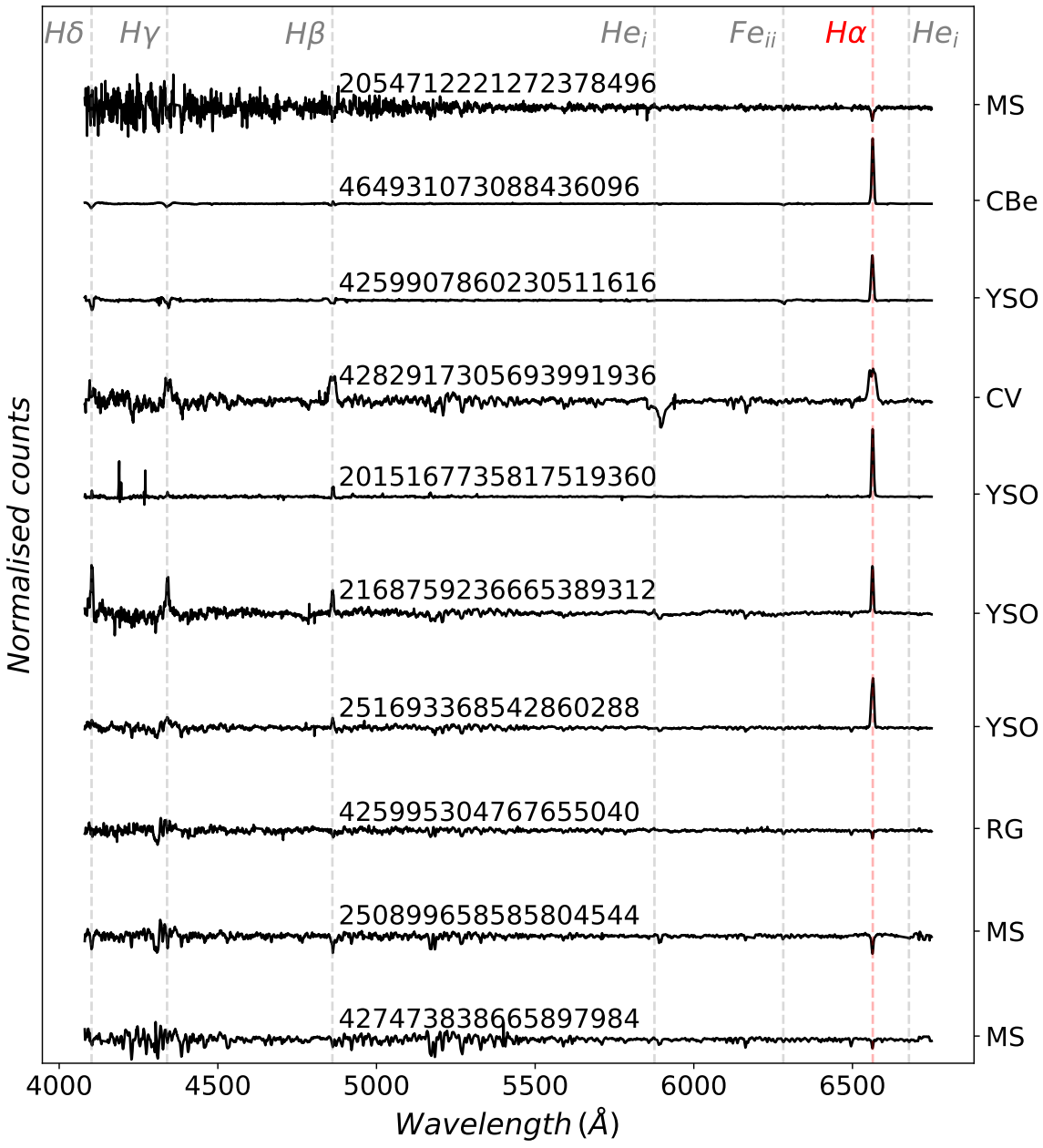


Figure A.2: subset 11-20/114 of inspected spectra, normalised for visual purposes.

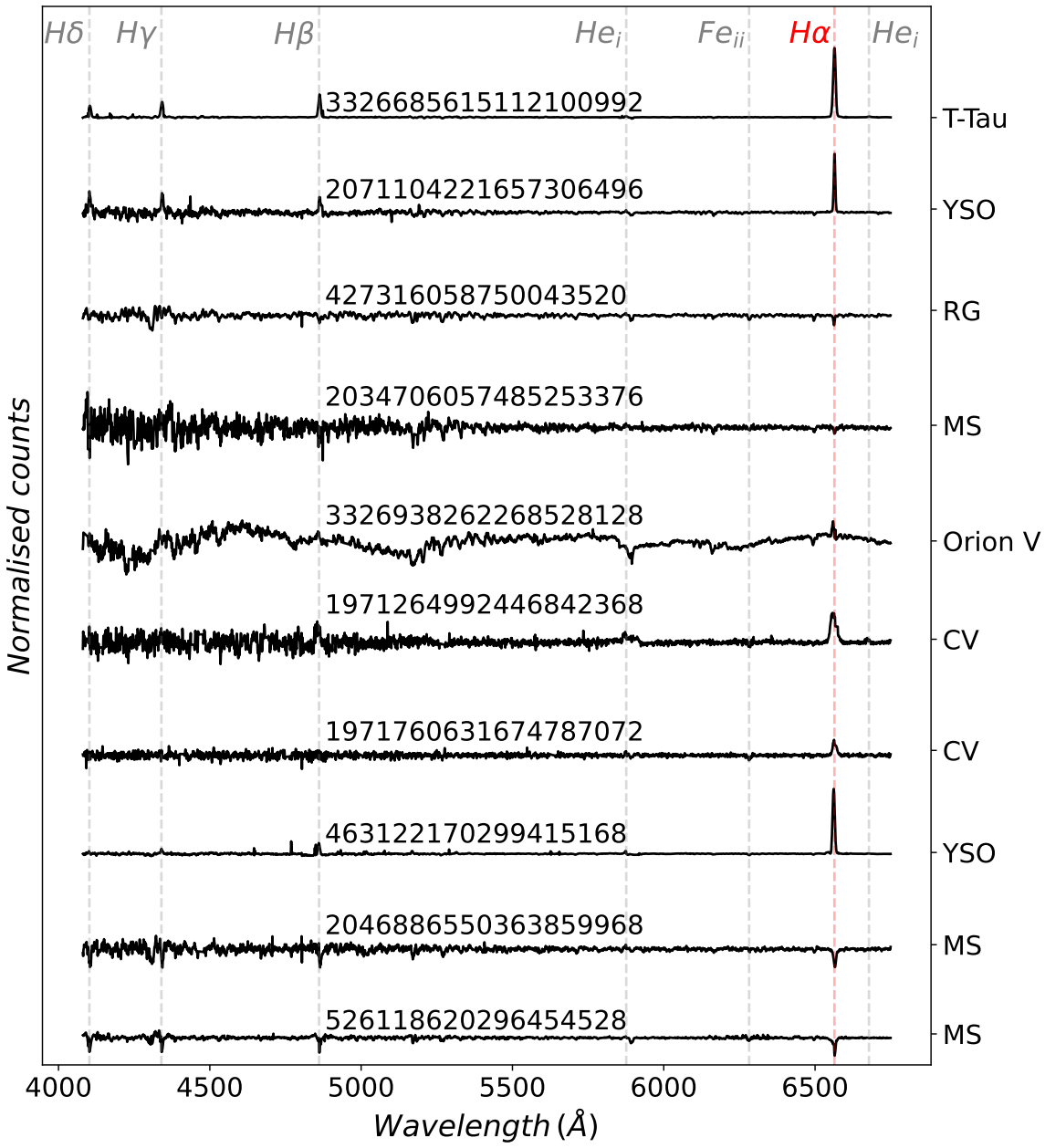


Figure A.3: subset 21-30/114 of inspected spectra, normalised for visual purposes.

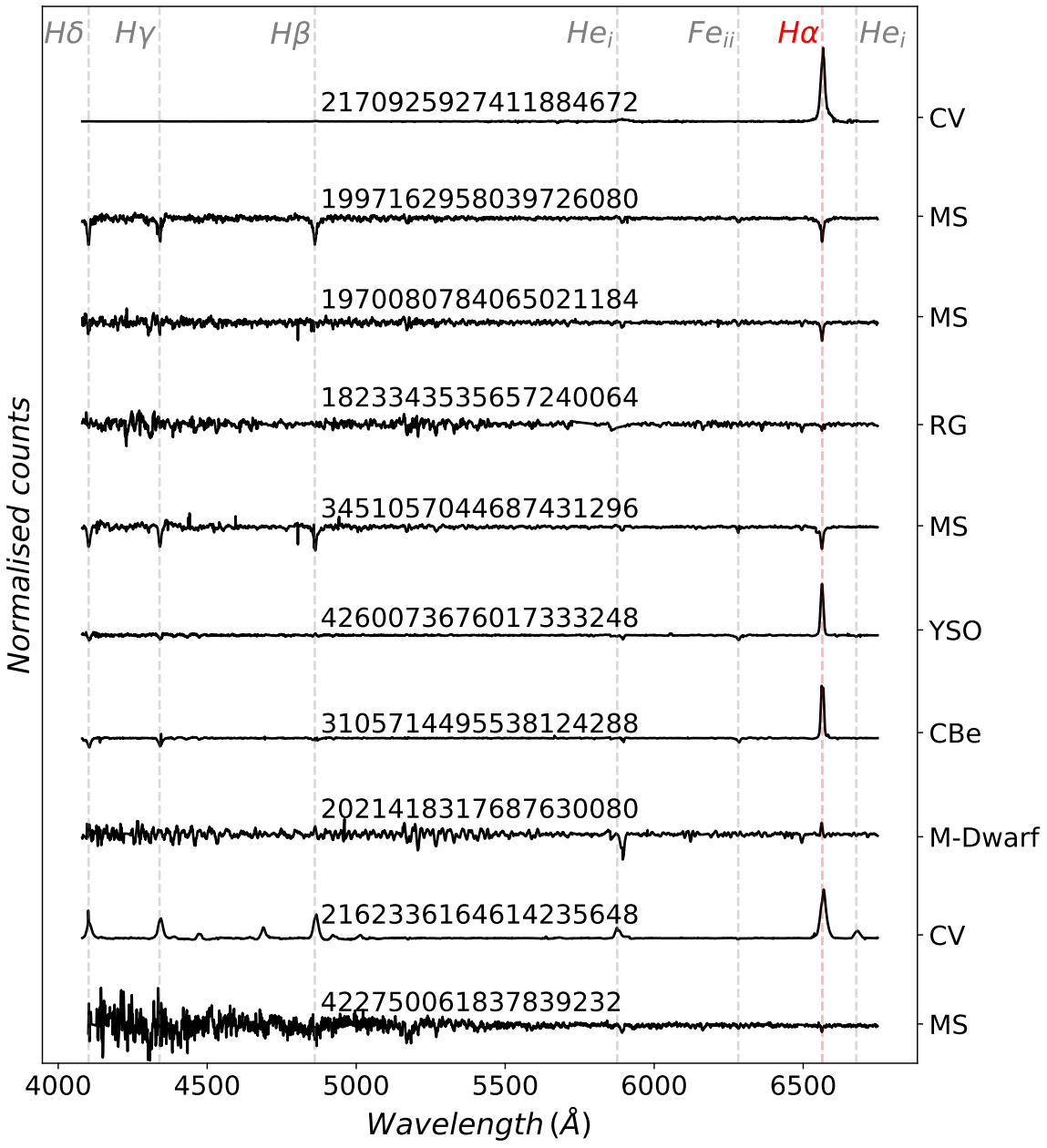


Figure A.4: subset 31-40/114 of inspected spectra, normalised for visual purposes.

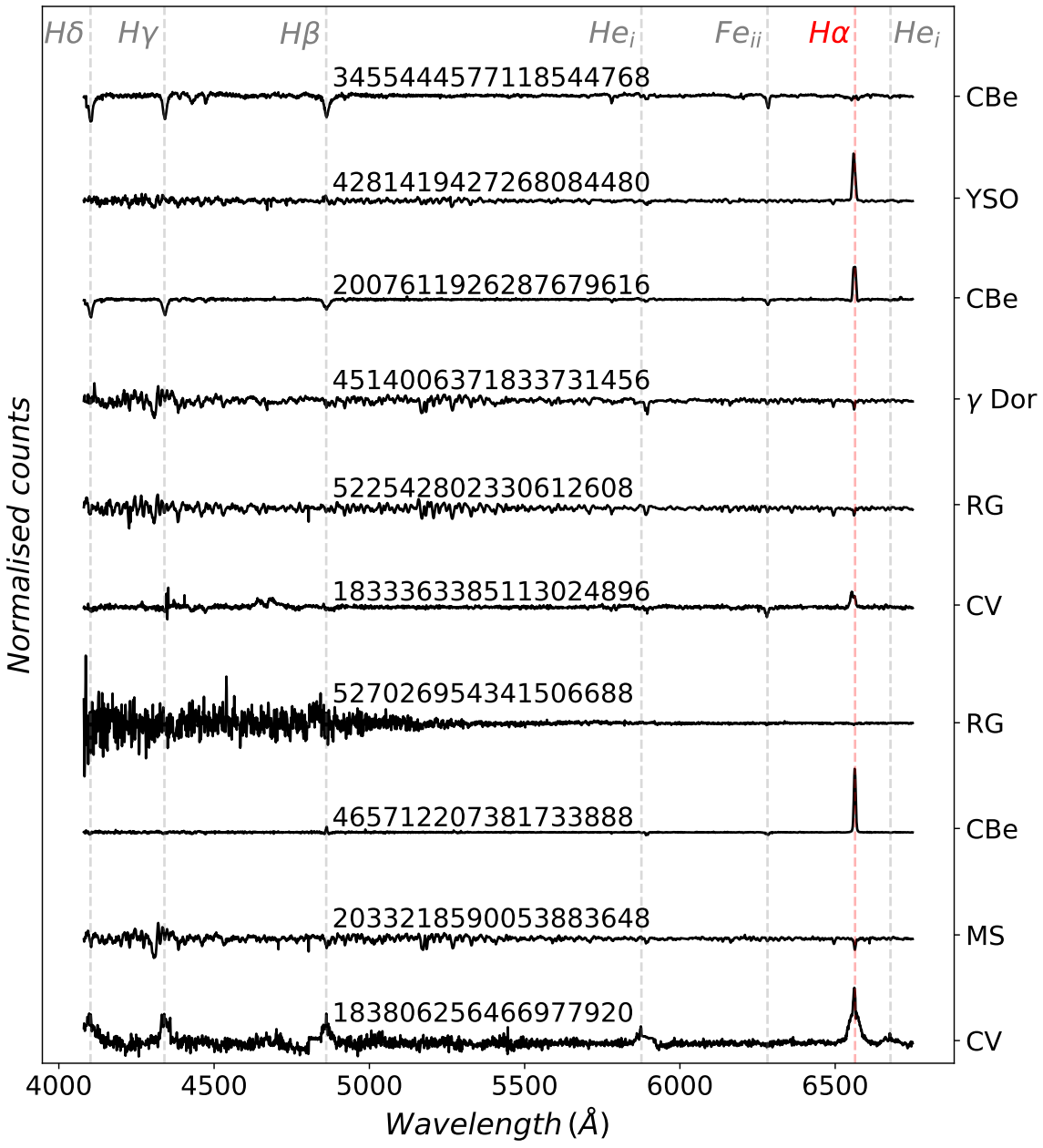


Figure A.5: subset 41-50/114 of inspected spectra, normalised for visual purposes.

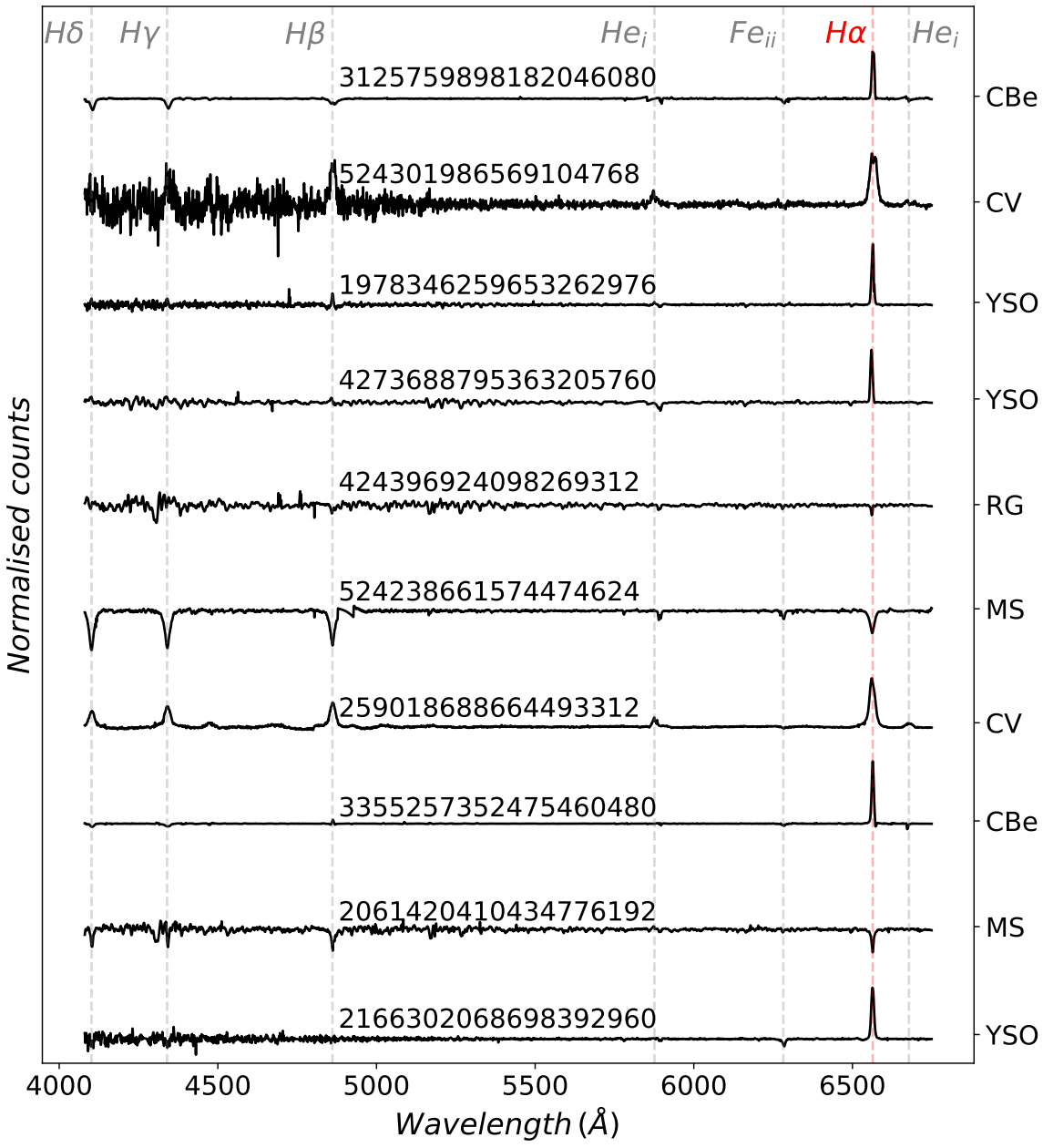


Figure A.6: subset 51-60/114 of inspected spectra, normalised for visual purposes.

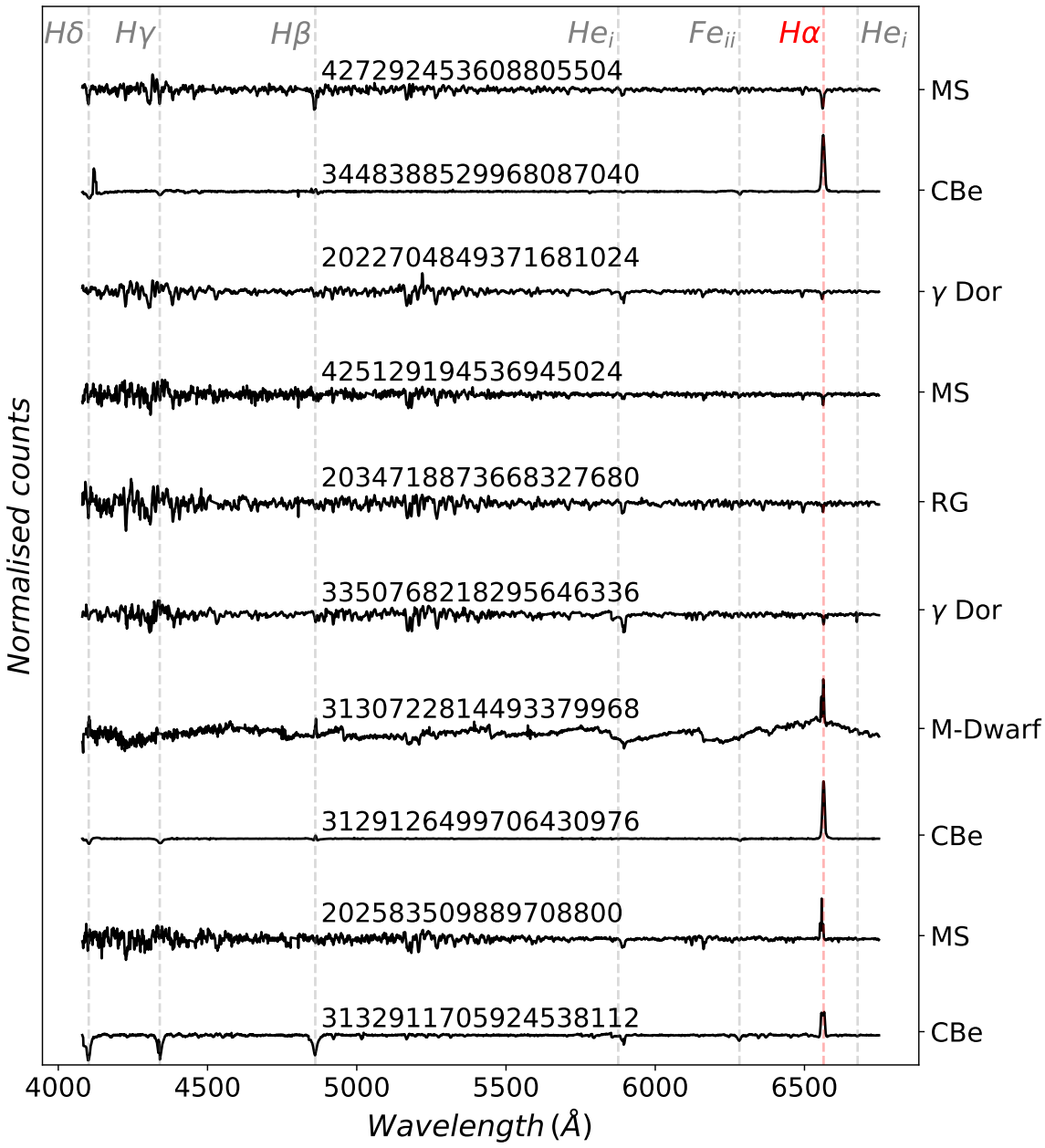


Figure A.7: subset 61-70/114 of inspected spectra, normalised for visual purposes.

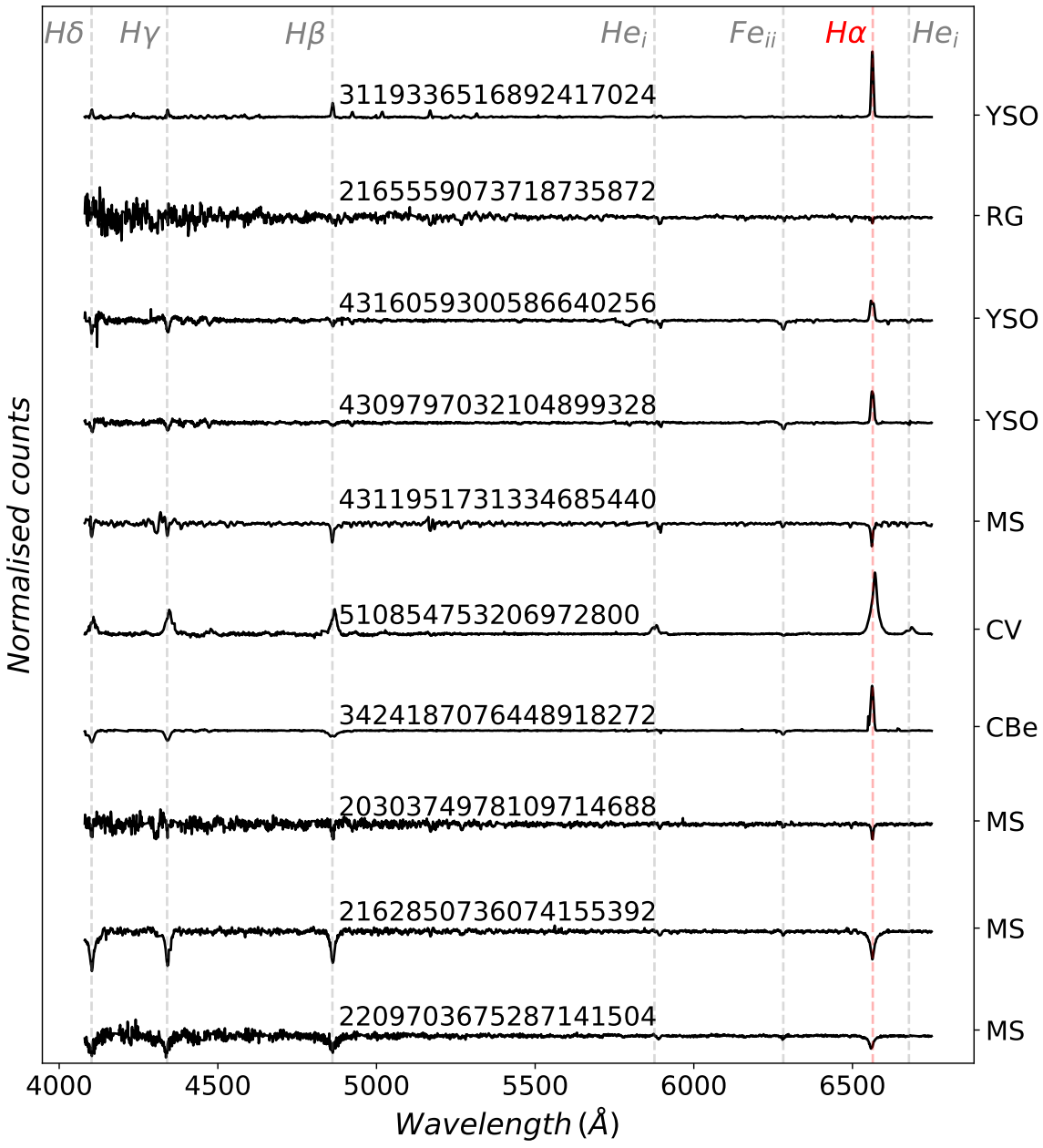


Figure A.8: subset 71-80/114 of inspected spectra, normalised for visual purposes.

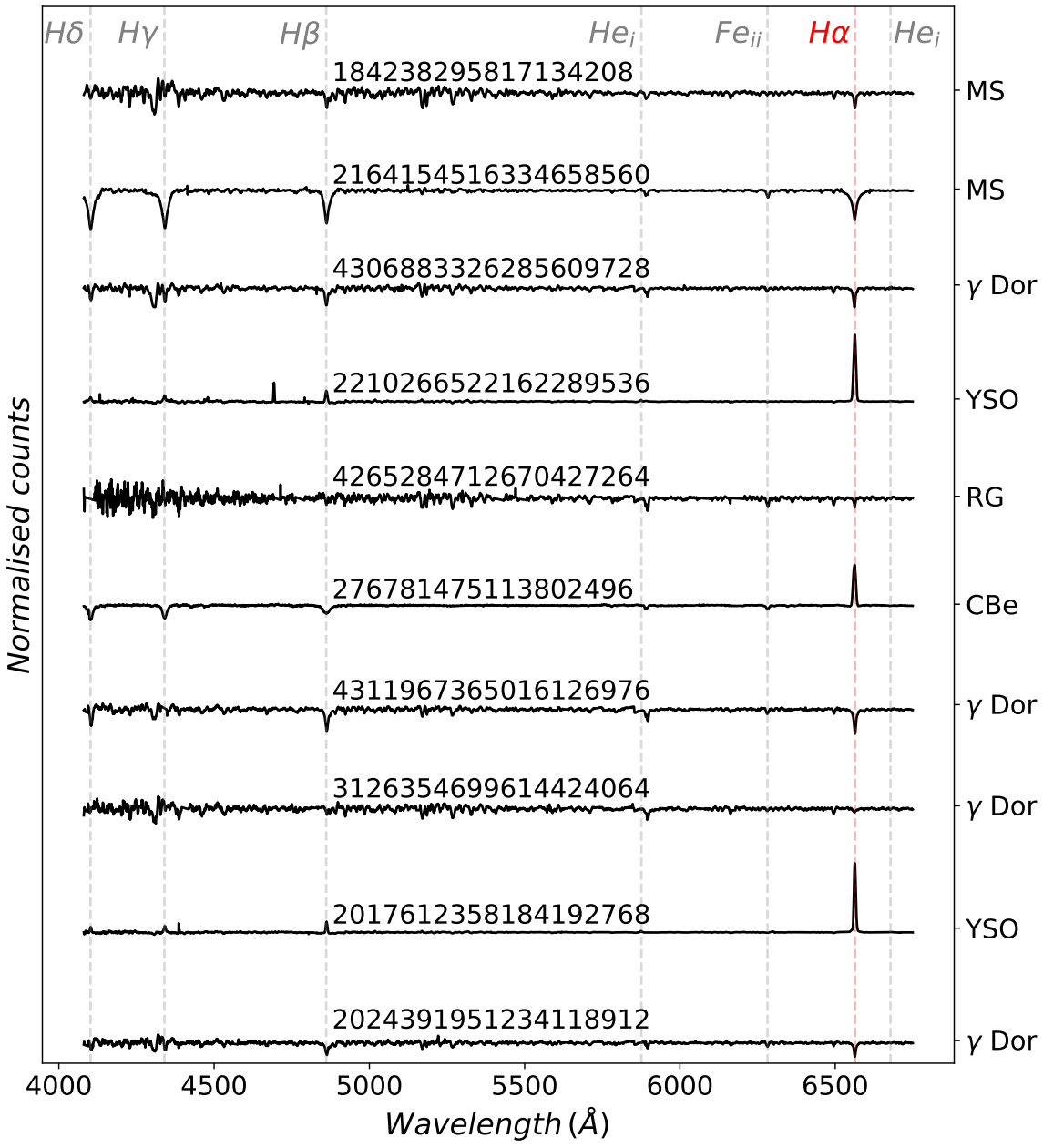


Figure A.9: subset 81-90/114 of inspected spectra, normalised for visual purposes.

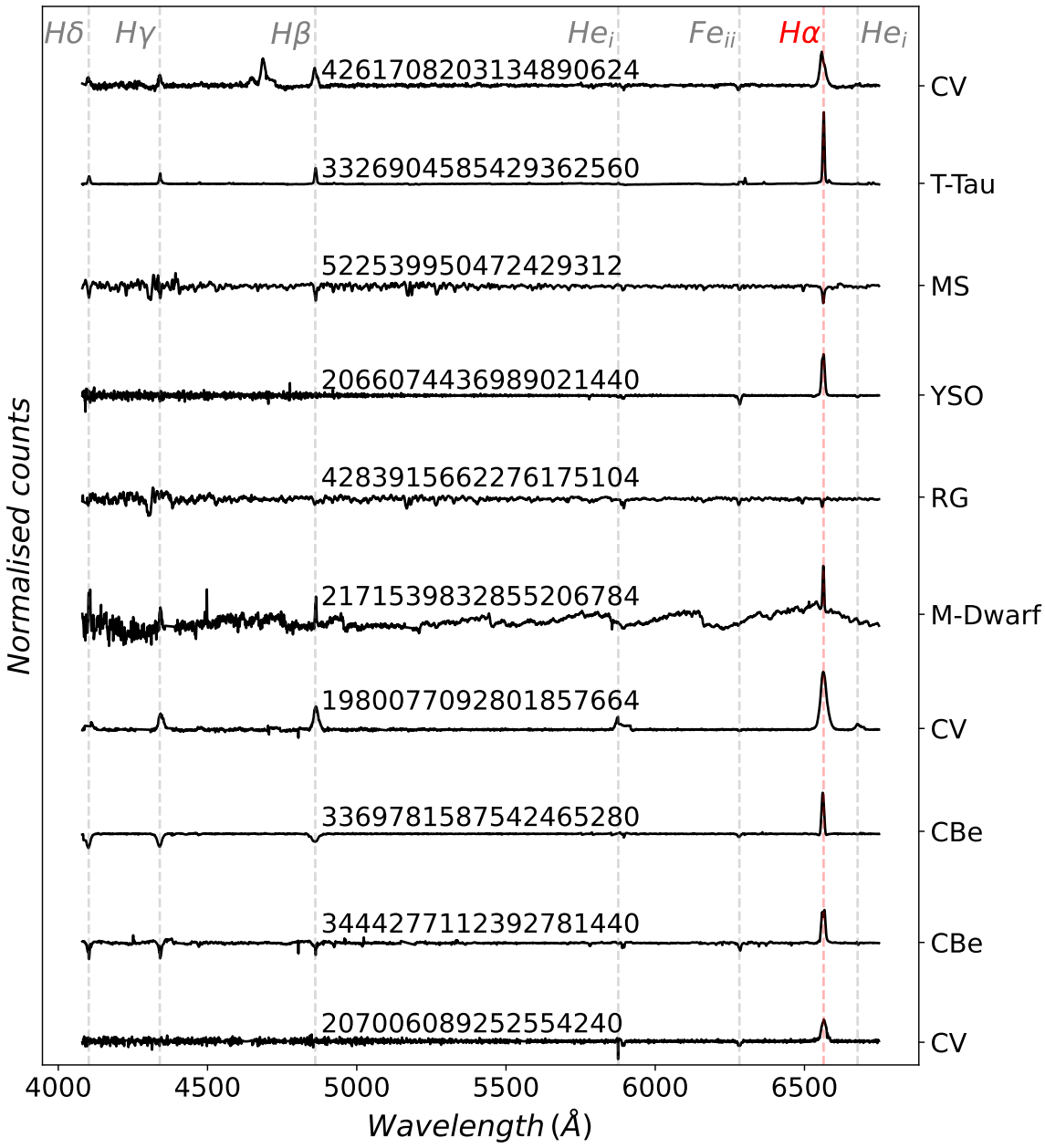


Figure A.10: subset 91-100/114 of inspected spectra, normalised for visual purposes.

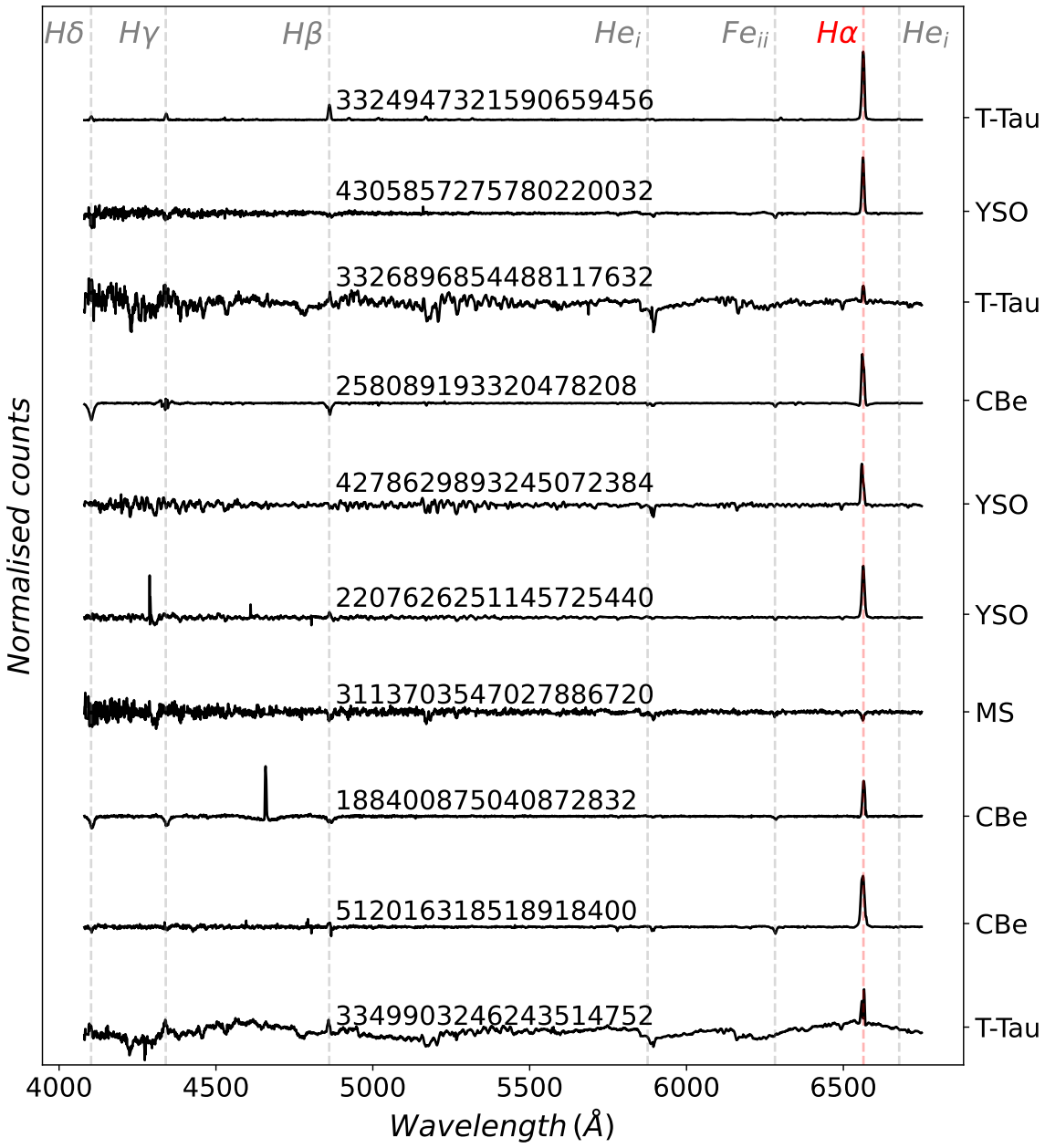


Figure A.11: subset 101-110/114 of inspected spectra, normalised for visual purposes.

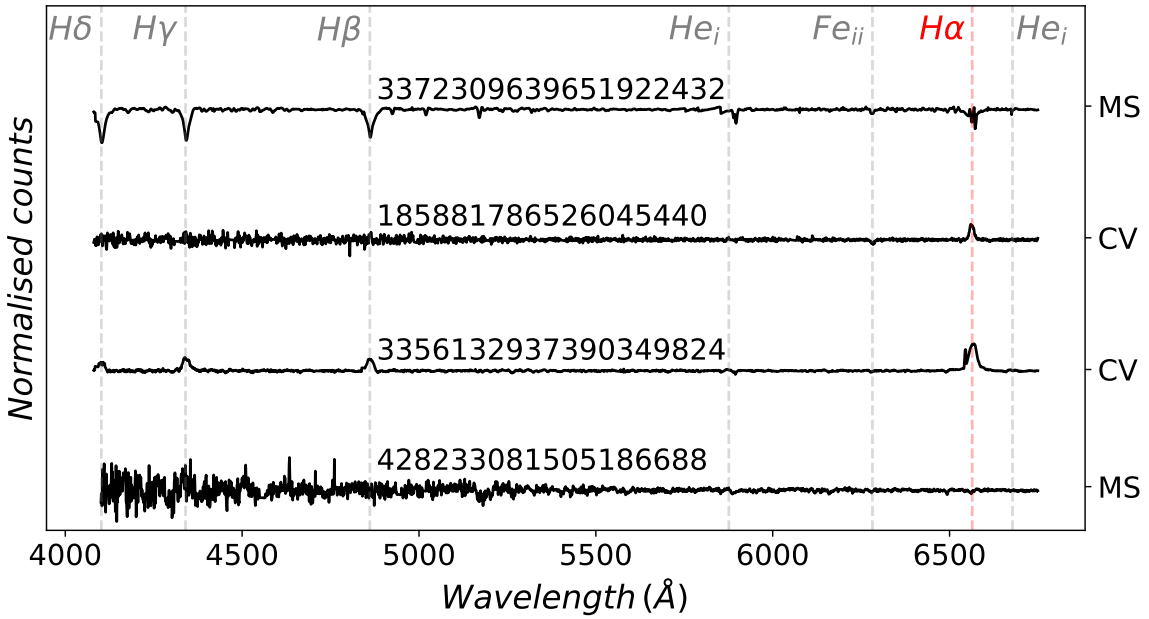


Figure A.12: subset of 111-114/114 inspected spectra, normalised for visual purposes.

Bibliography

The HIPPARCOS and TYCHO catalogues. Astrometric and photometric star catalogues derived from the ESA HIPPARCOS Space Astrometry Mission, volume 1200 of *ESA Special Publication* (1997).

Abbott, B.P., Abbott, R., Abbott, T.D., et al. *Observation of Gravitational Waves from a Binary Black Hole Merger.* Phys. Rev. Lett., **116**(6) (2016), 061102. 1602.03837.

Abrahams, E.S., Bloom, J.S., Mowlavi, N., et al. *Informing the Cataclysmic Variable Donor Sequence from Gaia DR2 Color-Magnitude and Inferred Variability Metrics.* arXiv e-prints (2020), arXiv:2011.12253. 2011.12253.

Abramovici, A., Althouse, W.E., Drever, R.W.P., et al. *LIGO: The Laser Interferometer Gravitational-Wave Observatory.* Science, **256**(5055) (1992), 325.

Allington-Smith, J., Murray, G., Content, R., et al. *Integral Field Spectroscopy with the Gemini Multiobject Spectrograph. I. Design, Construction, and Testing.* PASP, **114**(798) (2002), 892.

Astraatmadja, T.L., Bailer-Jones, C.A.L. *Estimating Distances from Parallaxes. III. Distances of Two Million Stars in the Gaia DR1 Catalogue.* ApJ, **833**(1) (2016), 119. 1609.07369.

Audard, M., Ábrahám, P., Dunham, M., et al. *Protostars and Planets VI* (2014).

Avilés, A., Chávez, C., García, P., et al. *On the Orbital Period of the Cataclysmic Variable V767 Cyg.* Rev. Mexicana Astron. Astrofis., **54** (2018), 389.

Balman, Ş., Godon, P., Sion, E.M. *Swift X-Ray Telescope Observations of the Nova-like Cataclysmic Variables MV Lyr, BZ Cam, and V592 Cas.* ApJ, **794**(1) (2014), 84. 1408.1996.

Barentsen, G., Farnhill, H.J., Drew, J.E., et al. *The second data release of the INT Photometric Hα Survey of the Northern Galactic Plane (IPHAS DR2).* MNRAS, **444**(4) (2014), 3230. 1406.4862.

Barentsen, G., Vink, J.S., Drew, J.E., et al. *Bayesian inference of T Tauri star properties using multi-wavelength survey photometry.* MNRAS, **429**(3) (2013), 1981. 1211.6108.

Bechtold, J., Yee, H.K.C., Elston, R., et al. *Hα Imaging of the Candidate Protogalaxy MS 1512-cB58.* ApJ, **477**(1) (1997), L29.

Belczyński, K., Mikołajewska, J., Munari, U., et al. *A catalogue of symbiotic stars.* A&AS, **146** (2000), 407. astro-ph/0005547.

Beltrán, M. *Disks around O-type young stellar objects.* arXiv e-prints (2020), arXiv:2005.06912. 2005.06912.

Benitez, N., Dupke, R., Moles, M., et al. *J-PAS: The Javalambre-Physics of the Accelerated Universe Astrophysical Survey.* arXiv e-prints (2014), arXiv:1403.5237. 1403.5237.

Berry, M., Ivezić, Ž., Sesar, B., et al. *The Milky Way Tomography with Sloan Digital Sky Survey. IV. Dissecting Dust.* ApJ, **757**(2) (2012), 166. 1111.4985.

Bond, N.A., Ivezić, Ž., Sesar, B., et al. *The Milky Way Tomography with SDSS. III. Stellar Kinematics.* ApJ, **716**(1) (2010), 1. 0909.0013.

Bondi, H., Hoyle, F. *On the mechanism of accretion by stars.* MNRAS, **104** (1944), 273.

Bonoli, S., Marín-Franch, A., Varela, J., et al. *The miniJPAS survey: A preview of the Universe in 56 colors.* A&A, **653** (2021), A31. 2007.01910.

Borucki, W.J., Koch, D., Basri, G., et al. *Kepler Planet-Detection Mission: Introduction and First Results.* Science, **327**(5968) (2010), 977.

Breedt, E., Gänsicke, B.T., Marsh, T.R., et al. *CSS100603:112253-111037: a helium-rich dwarf nova with a 65 min orbital period.* MNRAS, **425**(4) (2012), 2548. 1207.3836.

Breiman, L. *Random Forests.* Machine Learning, **45** (2001), 5.

Bressan, A., Marigo, P., Girardi, L., et al. *PARSEC: stellar tracks and isochrones with the PAdova and TRieste Stellar Evolution Code.* MNRAS, **427** (2012), 127. 1208.4498.

Buder, S., Asplund, M., Duong, L., et al. *The GALAH Survey: second data release.* MNRAS, **478**(4) (2018), 4513. 1804.06041.

Burrows, D.N., Hill, J.E., Nousek, J.A., et al. *Swift x-ray telescope (XRT).* In J.E. Truemper, H.D. Tananbaum, editors, *X-Ray and Gamma-Ray Telescopes and Instruments for Astronomy.*, volume 4851 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series* (2003), pages 1320–1325.

Carlin, J.L., Lépine, S., Newberg, H.J., et al. *An algorithm for preferential selection of spectroscopic targets in LEGUE.* Research in Astronomy and Astrophysics, **12**(7) (2012), 755. 1206.3577.

Carrasco, J.M., Catalán, S., Jordi, C., et al. *Gaia photometry for white dwarfs.* A&A, **565** (2014), A11. 1403.6045.

Charles, P.A., Coe, M.J. *Optical, ultraviolet and infrared observations of X-ray binaries.* arXiv e-prints (2003), astro-ph/0308020. astro-ph/0308020.

Chou, Y. *Measuring the orbital periods of low mass X-ray binaries in the X-ray band.* Research in Astronomy and Astrophysics, **14**(11) (2014), 1367-1382. 1408.6638.

Collins, George W., I. *The Use of Terms and Definitions in the Study of be Stars (review Paper).* In A. Slettebak, T.P. Snow, editors, *IAU Colloq. 92: Physics of Be Stars* (1987), page 3.

Condon, J.J., Cotton, W.D., Greisen, E.W., et al. *The NRAO VLA Sky Survey.* AJ, **115**(5) (1998), 1693.

Cram, L.E. *X-ray heating of the quiescent chromospheres of dMe stars.* ApJ, **253** (1982), 768.

Cui, X.Q., Zhao, Y.H., Chu, Y.Q., et al. *The Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST).* Research in Astronomy and Astrophysics, **12**(9) (2012), 1197.

Cutri, R.M., Wright, E.L., Conrow, T., et al. *Explanatory Supplement to the WISE All-Sky Data Release Products.* Explanatory Supplement to the WISE All-Sky Data Release Products (2012).

Dalton, G., Trager, S.C., Abrams, D.C., et al. *WEAVE: the next generation wide-field spectroscopy facility for the William Herschel Telescope.* In I.S. McLean, S.K. Ramsay, H. Takami, editors, *Ground-based and Airborne Instrumentation for Astronomy IV*, volume 8446 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series* (2012), page 84460P.

de Groot, M. *The most luminous stars in the universe.* Irish Astronomical Journal, **18** (1988), 163.

de Jong, R.S., Agertz, O., Berbel, A.A., et al. *4MOST: Project overview and information for the First Call for Proposals.* The Messenger, **175** (2019), 3. 1903.02464.

Deason, A.J., Belokurov, V., Erkal, D., et al. *The Clouds are breaking: tracing the Magellanic system with Gaia DR1 Mira variables.* MNRAS, **467**(3) (2017), 2636. 1611.04600.

Dias, W.S., Monteiro, H., Caetano, T.C., et al. *Proper motions of the optically visible open clusters based on the UCAC4 catalog.* A&A, **564** (2014), A79.

Drew, J.E., Gonzalez-Solares, E., Greimel, R., et al. *The VST Photometric Hα Survey of the Southern Galactic Plane and Bulge (VPHAS+).* MNRAS, **440**(3) (2014), 2036. 1402.7024.

Drew, J.E., Greimel, R., Irwin, M.J., et al. *The INT Photometric Hα Survey of the Northern Galactic Plane (IPHAS).* MNRAS, **362** (2005), 753. astro-ph/0506726.

Evans, I.N., Primini, F.A., Glotfelty, K.J., et al. *The Chandra Source Catalog.* ApJS, **189**(1) (2010), 37. 1005.4665.

Eyres, S.P.S., Bode, M.F., Taylor, A.R., et al. *The Inner Nebula and Central Binary of the Symbiotic Star HM Sagittae.* ApJ, **551**(1) (2001), 512. astro-ph/0101189.

Farnhill, H.J., Drew, J.E., Barentsen, G., et al. *Calibrated and completeness-corrected optical stellar density maps of the northern Galactic plane.* MNRAS, **457**(1) (2016), 642. 1601.02983.

Fratta, M., Scaringi, S., Drew, J.E., et al. *Population-based identification of H α-excess sources in the Gaia DR2 and IPHAS catalogues.* MNRAS, **505**(1) (2021), 1135. 2105.00749.

Fratta, M., Scaringi, S., Monguió, M., et al. *Spectroscopic follow-up of a sub-set of the Gaia/IPHAS catalogue of Hα-excess sources.* arXiv e-prints (2022), arXiv:2211.01694. 2211.01694.

Gaia Collaboration. *VizieR Online Data Catalog: Gaia EDR3 (Gaia Collaboration, 2020).* VizieR Online Data Catalog (2020), I/350.

Gaia Collaboration, Brown, A.G.A., Vallenari, A., et al. *Gaia Data Release 2. Summary of the contents and survey properties.* A&A, **616** (2018), A1. 1804.09365.

Gaia Collaboration, Brown, A.G.A., Vallenari, A., et al. *Gaia Early Data Release 3. Summary of the contents and survey properties.* A&A, **649** (2021), A1. 2012.01533.

Gaia Collaboration, Prusti, T., de Bruijne, J.H.J., et al. *The Gaia mission.* A&A, **595** (2016), A1. 1609.04153.

Gaia Collaboration, Vallenari, A., Brown, A.G.A., et al. *Gaia Data Release 3: Summary of the content and survey properties.* arXiv e-prints (2022), arXiv:2208.00211. 2208.00211.

Gänsicke, B.T., Dillon, M., Southworth, J., et al. *SDSS unveils a population of intrinsically faint cataclysmic variables at the minimum orbital period.* MNRAS, **397**(4) (2009), 2170. 0905.3476.

Gavazzi, G., Boselli, A., Donati, A., et al. *Introducing GOLDMine: A new galaxy database on the WEB.* A&A, **400** (2003), 451. astro-ph/0212257.

Geier, S., Raddi, R., Gentile Fusillo, N.P., et al. *The population of hot subdwarf stars studied with Gaia. II. The Gaia DR2 catalogue of hot subluminous stars.* A&A, **621** (2019), A38. 1810.09321.

Giampapa, M.S., Golub, L., Rosner, R., et al. *A heating mechanism for the chromospheres of M dwarf stars.* SAO Special Report, **1** (1982), A.73.

Glazebrook, K., Peacock, J.A., Collins, C.A., et al. *An imaging K-band survey - I. The catalogue, star and galaxy counts.* MNRAS, **266** (1994), 65. astro-ph/9307022.

Groot, P.J., Drew, J., Greimel, R., et al. *The European Galactic Plane Surveys: EGAPS.* In *IAU Joint Discussion*, volume 26 of *IAU Joint Discussion* (2006), page 54.

Groot, P.J., Verbeek, K., Greimel, R., et al. *The UV-Excess survey of the northern Galactic plane.* MNRAS, **399**(1) (2009), 323. 0906.3498.

Gunn, J.E., Knapp, G.R. *The Sloan Digital Sky Survey.* In B.T. Soifer, editor, *Sky Surveys. Protostars to Protogalaxies*, volume 43 of *Astronomical Society of the Pacific Conference Series* (1993), page 267.

Gunning, H.C., Schmidt, S.J., Davenport, J.R.A., et al. *Hα Emission From Active Equal-Mass, Wide M Dwarf Binaries*. PASP, **126**(946) (2014), 1081. 1410.1514.

Han, Z., Boonrucksar, S., Qian, S., et al. *Spectroscopic properties of the dwarf nova-type cataclysmic variables observed by LAMOST*. PASJ, **72**(5) (2020), 76. 2007.14016.

Hillenbrand, L.A., Hoffer, A.S., Herczeg, G.J. *An Enhanced Spectroscopic Census of the Orion Nebula Cluster*. AJ, **146**(4) (2013), 85. 1307.5533.

Hillier, D. *Wolf-Rayet Stars*. In P. Murdin, editor, *Encyclopedia of Astronomy and Astrophysics* (2000), page 1895.

Hillier, D.J. *Wolf-Rayet stars and stellar winds*. In C.S. Jeffery, U. Heber, editors, *Hydrogen Deficient Stars*, volume 96 of *Astronomical Society of the Pacific Conference Series* (1996), page 111.

Hook, I.M., Jørgensen, I., Allington-Smith, J.R., et al. *The Gemini-North Multi-Object Spectrograph: Performance in Imaging, Long-Slit, and Multi-Object Spectroscopic Modes*. PASP, **116**(819) (2004), 425.

Hou, W., Luo, A.l., Li, Y.B., et al. *Spectroscopically Identified Cataclysmic Variables from the LAMOST Survey. I. The Sample*. AJ, **159**(2) (2020), 43. 1911.08338.

Hubert, H., Chambon, M.T. *Emission Features of Several Be Stars as Related to Their Luminosity Class and Spectral Type*. In A. Slettebak, editor, *Be and Shell Stars*, volume 70 (1976), page 33.

Humphreys, R.M., Davidson, K. *The Luminous Blue Variables: Astrophysical Geysers*. PASP, **106** (1994a), 1025.

Humphreys, R.M., Davidson, K. *The Luminous Blue Variables: Astrophysical Geysers*. PASP, **106** (1994b), 1025.

Humphreys, R.M., Davidson, K., Smith, N. *η Carinae's Second Eruption and the Light Curves of the η Carinae Variables*. PASP, **111**(763) (1999), 1124.

Ilbert, O., Arnouts, S., McCracken, H.J., et al. *Accurate photometric redshifts for the CFHT legacy survey calibrated using the VIMOS VLT deep survey*. A&A, **457**(3) (2006), 841. astro-ph/0603217.

Iłkiewicz, K., Scaringi, S., Court, J.M.C., et al. *Exploring the tilted accretion disc of AQ Men with TESS*. MNRAS, **503**(3) (2021), 4050. 2103.03041.

Irwin, M., Lewis, J. *INT WFS pipeline processing*. New A Rev., **45**(1-2) (2001), 105.

Itoh, Y., Gupta, R., Oasa, Y., et al. *Optical Spectroscopy of Candidates of Young Stellar Objects in NGC 1333*. PASJ, **62** (2010), 1149.

Ivezić, Ž., Sesar, B., Jurić, M., et al. *The Milky Way Tomography with SDSS. II. Stellar Metallicity*. ApJ, **684**(1) (2008), 287. 0804.3850.

Jurić, M., Ivezić, Ž., Brooks, A., et al. *The Milky Way Tomography with SDSS. I. Stellar Number Density Distribution.* ApJ, **673**(2) (2008), 864. astro-ph/0510520.

Kassim, N.E., Lazio, T.J.W., Erickson, W.C., et al. *The Low-Frequency Array (LOFAR): opening a new window on the universe.* In H.R. Butcher, editor, *Radio Telescopes*, volume 4015 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series* (2000), pages 328–340.

Kaye, A.B., Handler, G., Krisciunas, K., et al. *Gamma Doradus Stars: Defining a New Class of Pulsating Variables.* PASP, **111**(761) (1999), 840. astro-ph/9905042.

Keller, S.C., Wood, P.R., Bessell, M.S. *Be stars in and around young clusters in the Magellanic Clouds.* A&AS, **134** (1999), 489. astro-ph/9809072.

Kennicutt, R. C., J. *The rate of star formation in normal disk galaxies.* ApJ, **272** (1983), 54.

Kennicutt, Robert C., J., Lee, J.C., Funes, J.G., et al. *An Hα Imaging Survey of Galaxies in the Local 11 Mpc Volume.* ApJS, **178**(2) (2008), 247. 0807.2035.

Kharchenko, N.V., Piskunov, A.E., Schilbach, E., et al. *Global survey of star clusters in the Milky Way. II. The catalogue of basic parameters.* A&A, **558** (2013), A53. 1308.5822.

Kiman, R., Faherty, J.K., Cruz, K.L., et al. *Calibration of the Hα Age-Activity Relation for M Dwarfs.* AJ, **161**(6) (2021), 277. 2104.01232.

Knigge, C., Baraffe, I., Patterson, J. *The Evolution of Cataclysmic Variables as Revealed by Their Donor Stars.* ApJS, **194**(2) (2011), 28. 1102.2440.

Kochanek, C.S., Shappee, B.J., Stanek, K.Z., et al. *The All-Sky Automated Survey for Supernovae (ASAS-SN) Light Curve Server v1.0.* PASP, **129**(980) (2017), 104502. 1706.07060.

Kohoutek, L., Wehmeyer, R. *Catalogue of stars in the Northern Milky Way having H-alpha in emission.* Astronomische Abhandlungen der Hamburger Sternwarte, **11** (1997), 1.

Kohoutek, L., Wehmeyer, R. *Catalogue of H-alpha emission stars in the Northern Milky Way.* A&AS, **134** (1999), 255.

Kolb, U. *A model for the intrinsic population of cataclysmic variables.* A&A, **271** (1993), 149.

Kuhn, M.A., Hillenbrand, L.A., Sills, A., et al. *Kinematics in Young Star Clusters and Associations with Gaia DR2.* ApJ, **870**(1) (2019), 32. 1807.02115.

Lawrence, A., Warren, S.J., Almaini, O., et al. *The UKIRT Infrared Deep Sky Survey (UKIDSS).* MNRAS, **379**(4) (2007), 1599. astro-ph/0604426.

Lee, B., Chary, R.R. *Improved photometric redshifts with colour-constrained galaxy templates for future wide-area surveys.* MNRAS, **497**(2) (2020), 1935. 1910.08588.

Lee, J.C., Gil de Paz, A., Tremonti, C., et al. *Comparison of Hα and UV Star Formation Rates in the Local Volume: Systematic Discrepancies for Dwarf Galaxies.* ApJ, **706**(1) (2009), 599. 0909.5205.

Lépine, S., Gaidos, E. *An All-sky Catalog of Bright M Dwarfs.* AJ, **142**(4) (2011), 138. 1108.2719.

Lindegren, L., Hernández, J., Bombrun, A., et al. *Gaia Data Release 2. The astrometric solution.* A&A, **616** (2018), A2. 1804.09366.

Linsky, J.L., Bornmann, P.L., Carpenter, K.G., et al. *Outer atmospheres of cool stars. XII. A survey of IUE ultraviolet emission line spectra of cool dwarf stars.* ApJ, **260** (1982), 670.

Loebman, S.R., Ivezić, Ž., Quinn, T.R., et al. *The Milky Way Tomography with Sloan Digital Sky Survey. V. Mapping the Dark Matter Halo.* ApJ, **794**(2) (2014), 151. 1408.5388.

Lomb, N.R. *Least-Squares Frequency Analysis of Unequally Spaced Data.* Ap&SS, **39**(2) (1976), 447.

LSST Science Collaboration, Abell, P.A., Allison, J., et al. *LSST Science Book, Version 2.0.* arXiv e-prints (2009), arXiv:0912.0201. 0912.0201.

Luna, G.J.M., Sokoloski, J.L., Mukai, K., et al. *Symbiotic stars in X-rays.* A&A, **559** (2013), A6. 1211.6082.

Magnier, E.A., Schlafly, E., Finkbeiner, D., et al. *The Pan-STARRS 1 Photometric Reference Ladder, Release 12.01.* ApJS, **205**(2) (2013), 20. 1303.3634.

Marton, G., Ábrahám, P., Szegedi-Elek, E., et al. *Identification of Young Stellar Object candidates in the Gaia DR2 x AllWISE catalogue with machine learning methods.* MNRAS, **487**(2) (2019), 2522. 1905.03063.

Michelson, P.F., Atwood, W.B., Ritz, S. *Fermi Gamma-ray Space Telescope: high-energy results from the first year.* Reports on Progress in Physics, **73**(7) (2010), 074901. 1011.0213.

Mohr-Smith, M., Drew, J.E., Napiwotzki, R., et al. *The deep OB star population in Carina from the VST Photometric Hα Survey (VPHAS+).* MNRAS, **465**(2) (2017), 1807. 1610.07609.

Monguió, M., Greimel, R., Drew, J.E., et al. *IGAPS: the merged IPHAS and UVEX optical surveys of the northern Galactic plane.* A&A, **638** (2020), A18. 2002.05157.

Mortier, A., Oliveira, I., van Dishoeck, E.F. *Spectroscopic properties of young stellar objects in the Lupus molecular clouds.* MNRAS, **418**(2) (2011), 1194. 1108.1087.

Neiner, C., de Batz, B., Cochard, F., et al. *The Be Star Spectra (BeSS) Database.* AJ, **142**(5) (2011), 149.

Neugebauer, G., Habing, H.J., van Duinen, R., et al. *The Infrared Astronomical Satellite (IRAS) mission.* ApJ, **278** (1984), L1.

Paczynski, B. *Evolution of cataclysmic binaries.* Acta Astron., **31** (1981), 1.

Pala, A.F., Gänsicke, B.T., Breedt, E., et al. *A Volume-limited Sample of Cataclysmic Variables from Gaia DR2: Space Density and Population Properties.* MNRAS, **494**(3) (2020), 3799. 1907.13152.

Pala, A.F., Schmidtobreick, L., Tappert, C., et al. *The cataclysmic variable QZ Lib: a period bouncer.* MNRAS, **481**(2) (2018), 2523. 1809.02135.

Parker, Q.A., Phillipps, S., Pierce, M.J., et al. *The AAO/UKST SuperCOSMOS Hα survey.* MNRAS, **362**(2) (2005), 689. astro-ph/0506599.

Podsiadlowski, P., Han, Z., Rappaport, S. *Cataclysmic variables with evolved secondaries and the progenitors of AM CVn stars.* MNRAS, **340**(4) (2003), 1214.

Porter, J.M., Rivinius, T. *Classical Be Stars.* PASP, **115**(812) (2003), 1153.

Predehl, P., Hasinger, G., Böhringer, H., et al. *eROSITA.* In M.J.L. Turner, G. Hasinger, editors, *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 6266 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series* (2006), page 62660P.

Pringle, J.F. *Circumstellar Discs (Invited Paper).* In L. Drissen, C. Leitherer, A. Nota, editors, *Nonisotropic and Variable Outflows from Stars*, volume 22 of *Astronomical Society of the Pacific Conference Series* (1992), page 14.

Raddi, R., Drew, J.E., Fabregat, J., et al. *First results of an Hα based search of classical Be stars in the Perseus Arm and beyond.* MNRAS, **430**(3) (2013), 2169. 1301.1298.

Rajpurohit, A.S., Allard, F., Rajpurohit, S., et al. *Exploring the stellar properties of M dwarfs with high-resolution spectroscopy from the optical to the near-infrared.* A&A, **620** (2018), A180. 1810.13252.

Rajpurohit, A.S., Reylé, C., Schultheis, M., et al. *Stellar parameters of M dwarfs from low and high-resolution spectra together with new model atmospheres.* In S. Boissier, P. de Laverny, N. Nardetto, R. Samadi, D. Valls-Gabaud, H. Wozniak, editors, *SF2A-2012: Proceedings of the Annual meeting of the French Society of Astronomy and Astrophysics* (2012), pages 383–388.

Rappaport, S., Verbunt, F., Joss, P.C. *A new technique for calculations of binary stellar evolution application to magnetic braking.* ApJ, **275** (1983), 713.

Ricker, G.R., Winn, J.N., Vanderspek, R., et al. *Transiting Exoplanet Survey Satellite (TESS).* Journal of Astronomical Telescopes, Instruments, and Systems, **1** (2015), 014003.

Ritter, H., Kolb, U. *Catalogue of cataclysmic binaries, low-mass X-ray binaries and related objects (Seventh edition).* A&A, **404** (2003), 301. astro-ph/0301444.

Rivinius, T., Carciofi, A.C., Martayan, C. *Classical Be stars. Rapidly rotating B stars with viscous Keplerian decretion disks.* A&A Rev., **21** (2013), 69. 1310.3962.

Sale, S.E., Drew, J.E., Barentsen, G., et al. *A 3D extinction map of the northern Galactic plane based on IPHAS photometry.* MNRAS, **443**(4) (2014), 2907. 1406.0009.

Scargle, J.D. *Studies in astronomical time series analysis. II. Statistical aspects of spectral analysis of unevenly spaced data.* ApJ, **263** (1982), 835.

Scaringi, S., Groot, P.J., Still, M. *Kepler observations of the eclipsing cataclysmic variable KIS J192748.53+444724.5.* MNRAS, **435** (2013a), L68. 1307.6558.

Scaringi, S., Groot, P.J., Verbeek, K., et al. *Spectroscopic identifications of blue-Hα-excess sources in the Kepler field of view.* MNRAS, **428**(3) (2013b), 2207. 1210.3038.

Scaringi, S., Knigge, C., Drew, J.E., et al. *The Gaia/IPHAS and Gaia/KIS value-added catalogues.* MNRAS, **481** (2018), 3357. 1809.04086.

Shappee, B.J., Prieto, J.L., Grupe, D., et al. *The Man behind the Curtain: X-Rays Drive the UV through NIR Variability in the 2013 Active Galactic Nucleus Outburst in NGC 2617.* ApJ, **788**(1) (2014), 48. 1310.2241.

Sheets, H.A., Thorstensen, J.R., Peters, C.J., et al. *Spectroscopy of Nine Cataclysmic Variable Stars.* PASP, **119**(855) (2007), 494. 0704.0948.

Smak, J. *Structure of Accretion Disks and the Visibility of White Dwarfs in Cataclysmic Binaries.* Acta Astron., **42** (1992), 323.

Smith, B.J., Price, S.D., Baker, R.I. *The COBE DIRBE Point Source Catalog.* ApJS, **154**(2) (2004), 673. astro-ph/0406177.

Smithsonian Astrophysical Observatory. *SAOImage DS9: A utility for displaying astronomical images in the X11 window environment* (2000). 0003.002.

Steinmetz, M., Zwitter, T., Siebert, A., et al. *The Radial Velocity Experiment (RAVE): First Data Release.* AJ, **132**(4) (2006), 1645. astro-ph/0606211.

Strom, K.M., Strom, S.E., Edwards, S., et al. *Circumstellar Material Associated with Solar-Type Pre-Main-Sequence Stars: A Possible Constraint on the Timescale for Planet Building.* AJ, **97** (1989), 1451.

Szkody, P. *The Impact of Accretion of the Secondary; the White Dwarf and the Disk ill Quiescent CVs.* In N. Vogt, editor, *Cataclysmic Variable Stars*, volume 29 of *Astronomical Society of the Pacific Conference Series* (1992), page 42.

Tacchella, S., Smith, A., Kannan, R., et al. *H α emission in local galaxies: star formation, time variability, and the diffuse ionized gas.* MNRAS, **513**(2) (2022), 2904. 2112.00027.

Tauris, T.M., van den Heuvel, E.P.J. *Formation and evolution of compact stellar X-ray sources.* In *Compact stellar X-ray sources*, volume 39 (2006), pages 623–665.

Taylor, M.B. *TOPCAT & STIL: Starlink Table/VOTable Processing Software.* In P. Shopbell, M. Britton, R. Ebert, editors, *Astronomical Data Analysis Software and Systems XIV*, volume 347 of *Astronomical Society of the Pacific Conference Series* (2005), page 29.

Thomas, S.J., Barr, J., Callahan, S., et al. *Vera C. Rubin Observatory: telescope and site status.* In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 11445 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series* (2020), page 114450I.

Tody, D. *The IRAF Data Reduction and Analysis System.* In D.L. Crawford, editor, *Instrumentation in astronomy VI*, volume 627 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series* (1986), page 733.

Voges, W., Aschenbach, B., Boller, T., et al. *The ROSAT all-sky survey bright source catalogue.* A&A, **349** (1999), 389. astro-ph/9909315.

Voges, W., Aschenbach, B., Boller, T., et al. *VizieR Online Data Catalog: ROSAT All-Sky Survey Faint Source Catalog (Voges+ 2000).* VizieR Online Data Catalog (2000), IX/29.

Vučetić, M.M., Arbutina, B., Urošević, D. *Optical supernova remnants in nearby galaxies and their influence on star formation rates derived from Hα emission.* MNRAS, **446**(1) (2015), 943. 1404.3942.

Walter, F.M. *Naked T-Tauri Stars - the Low-Mass Pre-Main Sequence Unveiled.* PASP, **98** (1986), 1100.

Wargau, W.F. *Nova-like objects and dwarf novae during outburst - a comparative study.* The Messenger, **40** (1985), 7.

Warner, B. *Cataclysmic Variable Stars* (2003).

Waters, L.B.F.M., Waelkens, C. *Herbig Ae/Be Stars.* ARA&A, **36** (1998), 233.

Webster, B.L., Allen, D.A. *Symbiotic stars and dust.* MNRAS, **171** (1975), 171.

Weis, K., Bomans, D.J. *Luminous Blue Variables.* Galaxies, **8**(1) (2020), 20. 2009.03144.

Wenger, M., Ochsenbein, F., Egret, D., et al. *The SIMBAD astronomical database. The CDS reference database for astronomical objects.* A&AS, **143** (2000), 9. astro-ph/0002110.

Whitelock, P.A. *Symbiotic Miras.* PASP, **99** (1987), 573.

Witham, A.R., Knigge, C., Drew, J.E., et al. *The IPHAS catalogue of Hα emission-line sources in the northern Galactic plane.* MNRAS, **384** (2008), 1277. 0712.0988.

Wolf, B. *Empirical amplitude-luminosity relation of S Doradus variables and extragalactic distances.* A&A, **217** (1989), 87.

Wood, M.A., Montgomery, M.M., Simpson, J.C. *Smoothed Particle Hydrodynamics Simulations of Apsidal and Nodal Superhumps.* ApJ, **535**(1) (2000), L39. astro-ph/0004159.

Wood, M.A., Thomas, D.M., Simpson, J.C. *SPH simulations of negative (nodal) superhumps: a parametric study.* MNRAS, **398**(4) (2009), 2110. 0906.2713.

Wright, E.L., Eisenhardt, P.R.M., Mainzer, A.K., et al. *The Wide-field Infrared Survey Explorer (WISE): Mission Description and Initial On-orbit Performance.* AJ, **140**(6) (2010), 1868. 1008.0031.

Yanny, B., Rockosi, C., Newberg, H.J., et al. *SEGUE: A Spectroscopic Survey of 240,000 Stars with g = 14-20.* AJ, **137**(5) (2009), 4377. 0902.1781.

Young, A., Skumanich, A., Stauffer, J.R., et al. *A Study of Excess H alpha Emission in Chromospherically Active M Dwarf Stars.* ApJ, **344** (1989), 427.

Zhao, B., Tomida, K., Hennebelle, P., et al. *Formation and Evolution of Disks Around Young Stellar Objects.* Space Sci. Rev., **216**(3) (2020), 43. 2004.01342.