Astronomy & Astrophysics manuscript no. output
January 17, 2025

Mapping Hα-Excess Candidate Point Sources in the Southern Hemisphere Using S-PLUS Data

L. A. Gutiérrez-Soto[1,2,⋆], R. Lopes de Oliveira[2,3,4], S. Akras[5], D. R. Gonçalves[6], L. F. Lomelí-Núñez[6], C. Mendes de Oliveira[2], E. Telles[4], A. Alvarez-Candal[7,8], M. Borges Fernandes[4], S. Daflon[4], C. E. Ferreira Lopes[9,10], M. Grossi[6], D. Hazarika[9,10], P. K. Humire[2], C. Lima-Dias[11], A. R. Lopes[1], J. L. Nilo Castellón[12], S. Panda[13] A. Kanaan[14], T. Ribeiro[15], W. Schoenell[16]

(Affiliations can be found after the references)

Received September 15, 1996; accepted March 16, 1997

ABSTRACT

Context. We use the Southern Photometric Local Universe Survey (S-PLUS) Fourth Data Release (DR4) to identify and classify Hα-excess point source candidates in the Southern Sky. This approach combines photometric data from 12 S-PLUS filters with machine learning techniques to improve source classification and advance our understanding of Hα-related phenomena.

Aims. Our goal is to enhance the classification of Hα-excess point sources by distinguishing between Galactic and extragalactic objects, particularly those with redshifted emission lines, and to identify sources where the Hα excess is associated with variability phenomena, such as short-period RR Lyrae stars.

Methods. We selected Hα-excess candidates using the $(r - J0660)$ versus $(r - i)$ color-color diagram from the S-PLUS main survey (MS) and Galactic Disk Survey (GDS). Dimensionality reduction was achieved using UMAP, followed by HDBSCAN clustering. We refined this by incorporating infrared data, improving the separation of source types. A Random Forest model was then trained on the clustering results to identify key color features for the classification of Hα-excess sources. **New, effective color-color diagrams were constructed by combining data from S-PLUS MS and infrared data. These diagrams, alongside tentative color criteria, offer a preliminary classification of Hα-excess sources without the need for complex algorithms.**

Results. Combining multiwavelength photometric data with machine learning techniques significantly improved the classification of Hα-excess sources. We identified 6956 sources with excess in the $J0660$ filter, and cross-matching with SIMBAD allowed us to explore the types of objects present in our catalog, including emission-line stars, young stellar objects, nebulae, stellar binaries, cataclysmic variables, variable stars, and extragalactic sources such as QSOs, AGNs, and galaxies. The cross-match also revealed X-ray sources, transients, and other peculiar objects. Using S-PLUS colors and machine learning, we successfully separated RR Lyrae stars from both other Galactic stars and extragalactic objects. Additionally, we achieved a clear separation between Galactic and extragalactic sources. However, distinguishing cataclysmic variables from QSOs at specific redshifts remained challenging. Incorporating infrared data refined the classification, enabling us to separate Galactic from extragalactic sources and to distinguish cataclysmic variables from QSOs. The Random Forest model, trained on HDBSCAN results, highlighted key color features that distinguish the different classes of Hα-excess sources, providing a robust framework for future studies **such as follow-up spectroscopy.**

Key words. surveys – techniques: photometric – stars: novae, cataclysmic variables – quasars: emission lines

1. Introduction

Hydrogen Balmer emission lines are primarily produced by radiative processes, particularly radiative excitation and ionization, which dominate over collisional excitation under typical nebular conditions. For example, the Einstein A-coefficient for the Hα transition ($A_{32} \approx 4.41 \times 10^7\,s^{-1}$) is significantly larger than the typical collisional excitation rate coefficient ($\sim 10^{-9}$ to $10^{-8}\,cm^3\,s^{-1}$) at electron temperatures around 10 000 K. While collisional excitation can become more important in shock-heated or very dense environments, it generally remains secondary in the diffuse conditions of most nebulae. Universe being hydrogen-abundant, the observation of those electronic transitions offers an important window into the study of astrophysical objects. Among all the possible electronic transitions, the Balmer series represents an extremely useful tool in Astronomy, because it falls in the commonly used optical spectral range. In particular, the Hα emission line – rest-frame wavelength of 6564.614 Å in vacuum – that corresponds to the electron transition from the $n = 3$ to the $n = 2$ energy level, is the strongest one, in both emission or absorption, and the most widely used to identify various types of objects, for instance: star-forming regions, H II regions, planetary nebulae (PNe), supernovae, novae, young stellar objects (YSO), Herbig-Haro objects, circumstellar disks, post-asymptotic and asymptotic giant stars (AGB), red giant branch (RGB), active late-type dwarfs. Amongst massive stars, emission lines are observed in Be stars with decretion disks, B[e] supergiants, Luminous Blue Variables (LBVs), Wolf-Rayet (WR) stars, and interacting binary systems that are experiencing mass exchange, like symbiotic stars (SySt), cataclysmic variables (CVs), among others.

In high-redshift sources, such as starburst galaxies and quasi-stellar objects (QSOs), Hα emission is present, but redshifted to longer wavelengths. However, when we detect emission near 6563Å from high redshift sources, the recombination of Hα is not the cause, instead, it is the outcome of UV emission lines that have shifted towards the visible spectrum.

⋆ E-mail: gsotoangel@fcaglp.unlp.edu.ar

Most existing databases of the aforementioned classes of objects are not homogeneous and remain far from complete, even in the local universe. Some classes are highly populated, while others are significantly underrepresented. For example, there are ∼ 300 known SySts in the Milky Way but only 75 in nearby galaxies (Akras et al. 2019b; Merc et al. 2019) with constantly new discoveries being made every year (e.g. Merc et al. 2020; Akras et al. 2021; Merc et al. 2021, 2022; Munari et al. 2021, 2022; Akras 2023). The number of known PNe in our galaxy is on the order of ∼3500 (Parker et al. 2016), which may represent only 15-30% of the total population (Frew 2008; Jacoby et al. 2010).

Hα surveys have been conducted with varying angular resolutions, sky coverage, and sensitivity. Some surveys, despite having modest spatial resolutions, have successfully resolved extended nebular emissions, enabling the study of supernova remnants, galaxy groups, and star-forming regions (e.g. Davies et al. 1976; Blair & Long 2004; Jaiswal & Omar 2016; Cook et al. 2019). Others, with higher spatial resolution, have revealed compact emission-line sources in the Milky Way and nearby galaxies. Examples of them are the INT Photometric Hα survey (IPHAS; Drew et al. 2005; Barentsen et al. 2014), the SuperCOSMOS Hα survey with the UK Schmidt Telescope (UKST) of the Anglo-Australian Observatory (Parker et al. 2005), and the VST Photometric Hα Survey (VPHAS+; Drew et al. 2014).

Colour-colour diagrams from photometric surveys are also used to identify possible Hα emitters. For example, the (r - Hα) versus ($r-i$) colour-colour and similar diagrams has been used to find CVs (Witham et al. 2006, 2007), YSOs (Vink et al. 2008), SySt (Corradi et al. 2008; Corradi & Giammanco 2010; Corradi et al. 2011; Miszalski & Mikołajewska 2014; Mikołajewska et al. 2014, 2017; Akras et al. 2019c), early-type emission-line stars (Drew et al. 2008), and PNe (Miszalski et al. 2009; Viironen et al. 2009; Sabin et al. 2010; Akras et al. 2019a). Additionally, other combinations of broadband filters have been tailored to distinguish AGNs, QSOs, and compact PNe based on their distinct photometric signatures (Peters et al. 2015; Gutiérrez-Soto et al. 2024).

In particular, the class of Be stars is the more common, nearly 50%, in the total sample of Ha emitters in IPHAS with only a moderate $r - H\alpha$ excess, well limited near-infrared colors ($J - H, H - K$; Corradi et al. 2008; Raddi et al. 2015 and moderate mid-infrared colours (e.g. $W1 - W4, W3 - W4$; Akras et al. 2019c). Besides the identification of different classes of Hα emitters, the $r - H\alpha$ excess derived from Ha photometric surveys such as the IPHAS and VPHAS+, it also provides a new more automatic way to derive the accretion rate in large numbers of YSOs (Barentsen et al. 2011; Kalari et al. 2015).

Witham et al. (2008) developed a method to select Hα emission line sources in the IPHAS survey by implementing the aforementioned colour-colour diagram (r - Hα) versus ($r - i$). Objects with Hα excess line were identified by iteratively fitting the stellar locus and considering as candidates those objects that fall several sigma above this stellar locus in the r - Hα color. This conservative method yields a total of 4853 point sources in tha IPHAS catalog that exhibit strong photometric evidence for Hα emission. They obtained spectra from around 300 sources, confirming more than 95 percent of them as genuine emission-line stars.

Monguió et al. (2020) developed the INT Galactic Plane Survey (IGAPS) by merging the IPHAS and UVEX optical surveys. The IGAPS catalog includes 295.4 million photometric measurements in the i, r, narrow-band Hα, g, and U_{RGO} filters. It identifies 8292 candidate emission line stars and over 53 000 variable stars with confidence greater than 5σ.

More recently, Fratta et al. (2021) introduced a technique using Gaia data to identify Hα-bright sources in the IPHAS catalog. They partitioned the data based on Gaia color-absolute magnitude and Galactic coordinates to minimize contamination and then applied the strategy from Witham et al. (2008) to these partitions.

Two ongoing multi-band surveys are observing the sky in a systematic, complementary way, with 5 broad and 7 narrow-band filters, including Hα: the Javalambre Photometric Local Universe Survey (J-PLUS[1]; Cenarro et al. 2019), covering the Northern celestial hemisphere, and the Southern-Photometric Local Universe Survey (S-PLUS[2]; Mendes de Oliveira et al. 2019), covering the southern sky with a twin 83 cm telescope and filter system. The first survey is paving the way for an even more ambitious survey, the Javalambre Physics of the Accelerating Universe Astrophysical Survey (J-PAS; Benitez et al. 2014 and miniJ-PAS; Bonoli et al. 2021), which will observe the Northern sky with 56 narrow-band filters. As source hunters, the spectral energy distributions provided by these surveys enable an unprecedented source classification using photometry only. However, in the Big Data era, efficient investigation tools are required to deal with their massive imaging and catalogues production, and machine learning techniques have been increasingly used to explore these data sets (e.g Bom et al. 2021; Yang et al. 2022).

Here we present a census of Hα-excess point-like sources from the S-PLUS DR4, identified using the (r - $J0660$) versus (r - i) color-color diagram. Advanced machine learning techniques are employed to improve the identification and classification of these sources from the S-PLUS DR4 dataset. Specifically, we use Uniform Manifold Approximation and Projection (UMAP; Becht et al. 2018; McInnes et al. 2020) for dimensionality reduction followed by Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN; Campello et al. 2013) clustering to group sources based on their multi-wavelength photometric signatures. This approach allows us to handle high-dimensional data effectively and uncover patterns that traditional methods might overlook. Additionally, we incorporate Wide-Field Infrared Survey Explorer (WISE; Wright et al. 2010) data and apply a Random Forest (Breiman 2001) model to refine our classification and identify key features that distinguish different types of Hα-excess sources.

Section 2 describes the observations related to the S-PLUS project, including important information on the fourth data release, photometry, and data handling. Section 3.3 presents the technique implemented to select the Hα-feature sources. Section 4 includes the analysis of the results. In Section 5, we present the machine learning methods used to analyze and make a more accurate classification of the Hα sources. Finally, Section 6 discusses our main results and conclusions.

2. S-PLUS Survey Overview

S-PLUS surveys the southern sky using the 12 filters from the Javalambre filter system (Marín-Franch et al. 2012), a spanning the wavelength range from 3 000Å to 10 000Å. This system comprises seven narrow-band filters ($J0378$, $J0395$, $J0410$, $J0430$, $J0515$, $J0660$, $J0861$, and five broad-band Sloan-like (Fukugita et al. 1996) filters (see Figure 1). The narrow-band $J0660$ filter used in S-PLUS is centered at λ 6614Å and has a width of ≈

[1] https://www.j-plus.es
[2] http://www.splus.iag.usp.br

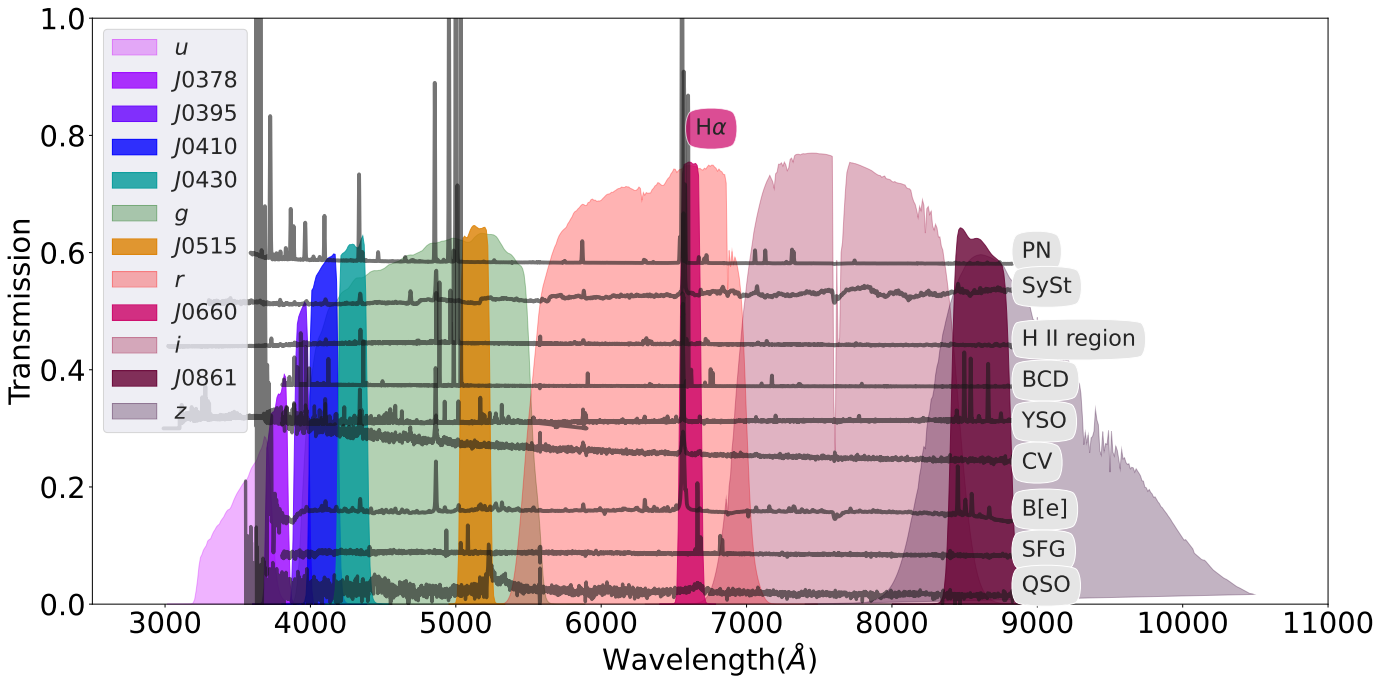

Fig. 1. Transmission curves of the S-PLUS filter set. The narrowband filter *J*0660 includes the Hα emission line. Over-plotted is spectra of different classes of emission line objects. From top to bottom: a PN, a symbiotic star, an extragalactic H II region, a blue compact/H II galaxy, a YSO, a CV star, a B[e] star, a star-forming galaxy and a QSO at a redshift of ∼3.31.

147Å (Table 2 of Mendes de Oliveira et al. 2019). Consequently, it covers both the Hα and the doublet [N II] $\lambda\lambda$ 6548,6584 spectral lines for sources up to a redshift of ≈ 0.02. S-PLUS is conducted using a dedicated 0.83 m robotic telescope located at Cerro Tololo, Chile (Mendes de Oliveira et al. 2019).

This work uses data from S-PLUS DR4 (Herpich et al. 2024). DR4 encompasses 171 fields at very low galactic latitudes ($|b| < 15°$), an additional 341 fields carried over from DR3 spanning the Main Survey (MS) footprint (with $|b| > 30°$), and 150 fields within the Magellanic Clouds (MC) region. This accumulation results in a total of 1629 fields in DR4, covering an expansive area of 3022.7 square degrees. Notably, this coverage includes 347.4 square degrees within the disk regions and 289.5 square degrees within the Magellanic clouds survey. Here, we explore the MS and Galactic Disk Survey (GDS), we primary goal of identify objects with Hα excess in S-PLUS DR4.

2.1. Flux Calibration

The flux calibration of the S-PLUS survey was performed using a combination of external and internal calibration steps to ensure uniformity and accuracy across the entire survey footprint. The calibration process begins with an external calibration, where synthetic photometry is integrated with a reference catalog to derive the calibrated magnitudes for the different filters. Zero points (ZPs) are determined as the difference between the predicted magnitudes from the synthetic models and the instrumental magnitudes observed in the survey. The synthetic spectral library for this step was constructed by convolving the library of Coelho (2014) with the transmission curves of multiple reference catalogs and the S-PLUS filter system.

In regions where external calibration data are unavailable, such as for the *u*, *J*0410, and *J*0430 filters, a stellar locus method is applied. This technique calibrates these filters by leveraging the stellar locus, a relationship between specific filter magnitudes observed for a population of stars. This step is crucial when external reference data are insufficient. Once the external calibration is complete, an internal calibration step further refines the ZPs by using pre-calibrated narrowband filters, which better constrain the synthetic models and improve the calibration accuracy to 0.01–0.02 mag. This internal calibration is particularly valuable for cases where external calibration might be lacking or less precise.

Finally, the calibration is aligned to the Gaia system by applying an average offset derived from synthetic photometry, ensuring consistency across the entire survey region. This final alignment guarantees that the flux calibration is homogeneous and compatible with the Gaia photometric system, as outlined by Herpich et al. (2024).

2.2. Filter Sequence and Observational Strategy

In the S-PLUS survey, the filters are observed in the following fixed sequence: *u*, *J*0378, *J*0395, *J*0410, *J*0430, *J*0515, *g*, *r*, *i*, *z*, *J*0660, *J*0861. The total time for observing each field is approximately 90 minutes, with each filter observed through three exposures, including readout times and filter changes. This sequence enables the study of variable sources, such as RR Lyrae stars and eclipsing binaries, by constructing light curves on timescales shorter than 30 minutes, which are suitable for investigating short-period variability in these sources.

are you sure about that
or > 30 min.

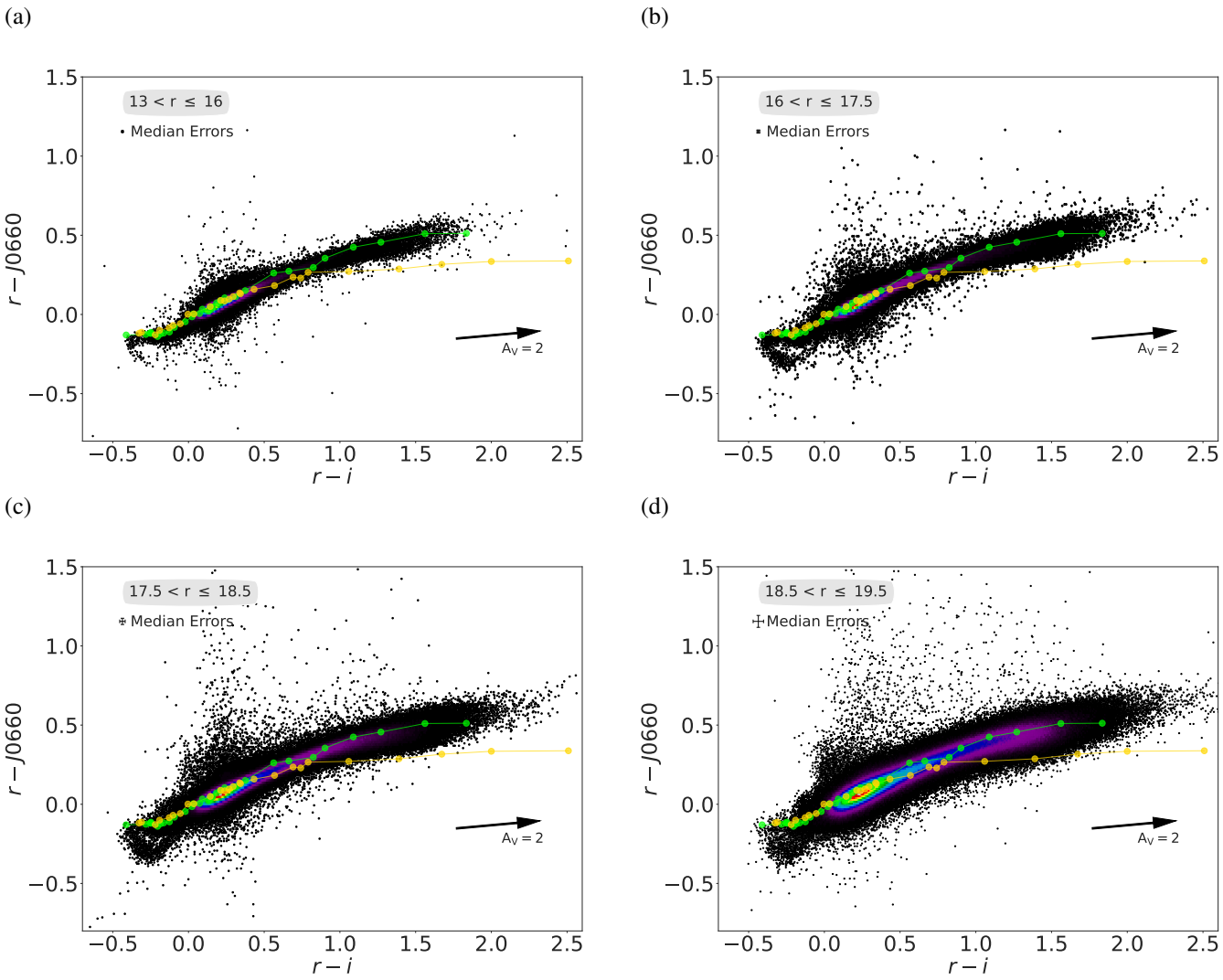

Fig. 2. The $(r - J0660)$ versus $(r - i)$ color-color plots used to select objects with Hα excess. These plots display data for all stars from the S-PLUS DR4 MS, representing the `PStotal` photometry in these colors. The data are divided into four magnitude bins: *(a)* $13 < r \leq 16$, *(b)* $16 < r \leq 17.5$, *(c)* $17.5 < r \leq 18.5$, and *(d)* $18.5 < r \leq 19.5$. Objects with Hα excess are expected to be located towards the top of these diagrams. Lighter green and yellow points connected by lines represent the tracks for main-sequence and giant stars, respectively. These tracks are derived from the synthetic spectra library of Pickles (1998). The background color gradient represents the density of objects, with red indicating the highest concentration of points, followed by green, blue, and purple, which represent progressively lower concentrations.

3. Selection Procedure

The selection of Hα candidates is based on applying a series of restrictions to catalogs provided by S-PLUS DR4 in the r, i, and $J0660$ bands. We have reinforced this information in the Galactic disk region by generating catalogs with PSF photometry using SExtractor + PSFEx.

3.1. Main Survey (MS) Data

Amongst the different aperture photometry available in the S-PLUS DR4 catalog, the `PStotal`[3] photometry is used (Almeida-Fernandes et al. 2022). To acquire data with high-quality photometry and identify compact objects in the MS, several criteria were applied:

- Objects must exhibit an r magnitude within the range of $13 < r \leq 19.5$.
- $J0660$ magnitude < 19.4 and i magnitude < 19.2, which are average photometric depths for a S/N >10 threshold (see Table 4 of Almeida-Fernandes et al. 2022).
- Errors less than 0.2 mag in the r, $J0660$, and i filters.
- The signal-to-noise ratio (S/N) in the respective filter should be higher than 10.
- Objects should have `SEX_FLAGS_DET` < 4. **The `SEX_FLAGS_DET` parameter is a bit-flag generated by SExtractor, indicating potential issues during photometry. The value corresponds to the sum of all flags for the respective observation tile, and objects with a `SEX_FLAGS_DET` value less than 4 were selected to ensure high-quality photometry.**
- **Objects must satisfy `CLASS_STAR_r` = 1 and `CLASS_STAR_i` = 1, corresponding to the binary classification in the r and i filters, where a value of 1 indicates that the source is classified as a point source (star) in each filter, and a value of 0 denotes non-stellar**

[3] `PStotal` refers to photometry obtained using a 3-arcsecond circular aperture, with corrections applied to account for the fraction of flux that falls outside this aperture. This method is intended to provide the best estimate of the total magnitude of point sources.

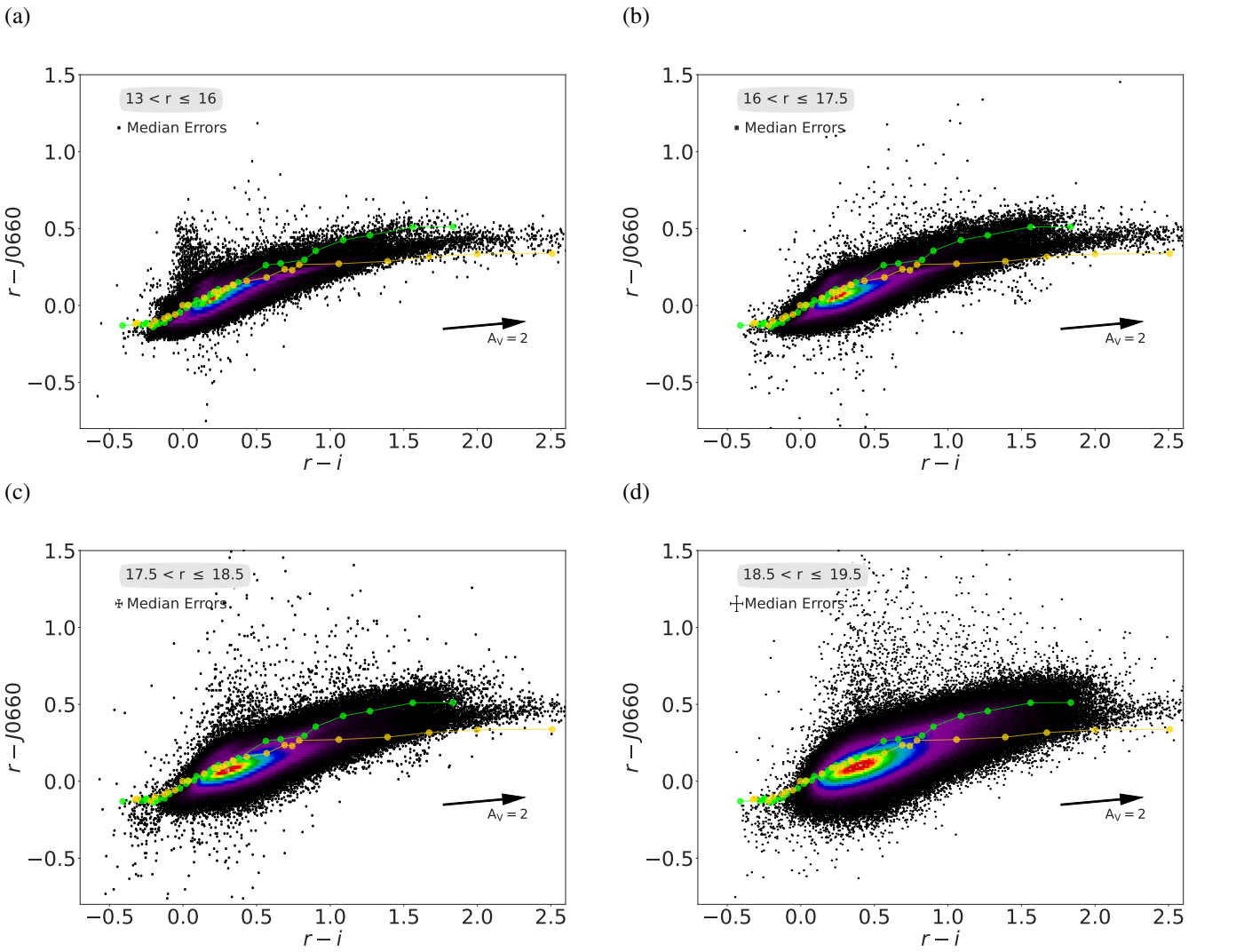

Fig. 3. Same as Fig. 2, but for the GDS and using PSF photometry.

or extended sources. The `CLASS_STAR` parameter in SExtractor represents a probability value ranging from 0 to 1, with higher values indicating a greater likelihood that the source is a point source. In our selection, we applied this binary classification for `CLASS_STAR_r` and `CLASS_STAR_i` to ensure a higher likelihood that the sources are stars.

Additional criteria were implemented. These criteria are systematically chosen to ensure the robustness and reliability of the selected sample, considering various photometric and morphological properties of the sources.

- We consider the morphological properties of the sources by imposing a threshold on ellipticity. Sources with ellipticity values greater than 0.2 are likely to have non-galactic types (e.g., AGN, QSOs, galaxies, radio sources) or irregular shapes and are therefore excluded.
- We select sources with compact morphology by constraining the radius enclosing 50% of the total flux, setting `FLUX_RADIUS_50` < 3. Sources with a flux radius exceeding 3 pixels are likely to have extended morphology and are thus excluded from the sample.

These constraints led to the selection of 6 655 139 sources. The data were obtained by querying the project's database using the `splusdata` Python package, accessible via `S-PLUS Cloud`[4].

3.2. Galactic Disk Survey (GDS)

We used a combination of SExtractor[5] (Bertin & Arnouts 1996) and PSFex[6] (Bertin 2011) for source detection and posterior photometric measurements. We performed a serie of proofs with different SExtractor (e.g. DETECT_MINAREA, DETECT_THRESH, PHOT_APERTURES) and PSFex (e.g. PSF_SIZE, PHOTFLUX_KEY, PSFVAR_DEGREES) parameters plus test images (e.g. BACKGROUND, BACKGROUND_RMS, -BACKGROUND, APERTURES) to detect the largest number of objects with the best measurement possible of PSF-magnitude, MAG_PSF. The crucial parameters for PSF photometry are listed in Table1. The detection was performed on images from which their median-filtered version was subtracted; faint sources are detected more easily in a median-subtracted image (González-Lópezlira et al. 2017). All median images were produced with a 11×11 pix^2 median filter.

[4] `https://splus.cloud/`
[5] `https://www.astromatic.net/software/sextractor`
[6] `https://www.astromatic.net/software/psfex`

Table 1. SExtractor and PSFex input parameters.

SExtractor	
Parameter	Value
DETECT_MINAREA	3
DETECT_THRESH	1.5
ANALYSIS_THRESH	1.5
PIXEL_SCALE	0.55
BACK_SIZE	64
BACK_FILTERSIZE	3
PSFex	
PSF_SIZE	18
PSFVAR_DEGREES	3

The PSF photometry method is described in González-López lira et al. (2017), Lomelí-Núñez et al. (2022), González-López lira et al. (2022) and Lomelí-Núñez in prep. A brief description of the photometric method is given below. a) *First run of SExtractor*: we run SExtractor for the first time for the detection and selection of point sources based on their brightness versus compactness, as measured by the parameters of SExtractor MAG_AUTO (a Kron-like elliptical aperture magnitude; Kron 1980) and FLUX_RADIUS (similar to the effective radius). For the creation of the PSF, we selected sources in the space MAG_AUTO vs FLUX_RADIUS, in a range to: $12 \lesssim$ MAG_AUTO $\lesssim 21.5$ and $1 \lesssim$ FLUX_RADIUS $\lesssim 3.5$. Because we are observing towards the Galactic disk, the number of sources to creation each PSF can reach ∼20000 sources, which was not possible in the previous works because they were focused on extragalactic sources far away from the Galactic disk. b) *PSF creation:* we used PSFex to create the PSF using the point sources selected in the last step. **The spatial variations of the PSF were modeled using third-degree polynomials as a function of the pixel coordinates (X, Y)**. For PSF creation, the flux of each star was measured in an aperture of 9 pixels of radius in all bands (equivalent to 4″.95×4″.95); such aperture, determined through the growth-curve method for each passband, is large enough to measure the total flux of the stars, but small enough to reduce the likelihood of contamination by external sources. c) *Second run of SExtractor:* we run SExtractor again this time using the PSF created in the last step as an input parameter to measure the magnitude of the PSF (MAG_PSF). In this work we always used the MAG_PSF, for simplicity only the name of each band is written.

The same constraints described in Section 3.1 were applied to the GDS to ensure high-quality data, resulting in the selection of 7 007 778 sources. To ensure the reliability of the data, five fields from the GDS were excluded from the analysis due to apparent calibration issues. These fields showed systematic offsets in the $(r-J0660)$ color when compared to other fields, indicating potential zero-point calibration problems. Excluding these fields minimizes the impact of systematic errors and enhances the robustness of the results.

3.3. Selection of Hα Excess Sources

Before searching for potential sources of Hα excess sources hidden in the S-PLUS DR4 footprint, we first divided our sample into four subsamples based on their magnitudes in the r band: (i) $13 \leq r < 16$, (ii) $16 \leq r < 17.5$, (iii) $17.5 \leq r < 18.5$, and (iv) $18.5 \leq r < 19.5$. This way, we avoided mixing up bright and faint sources with low and high uncertainties, respectively. Otherwise, the selection criteria could be affected by the intrinsic scatter in the measurement of faint objects. Figures 2 and 3 display the (r-$J0660$) versus (r-i) color-color diagrams for the sources from the MS of S-PLUS and the sub-survey of the GDS, respectively. The lighter green and yellow points connected by lines represent the tracks of main sequence and giant stars, respectively. These loci for main sequence and giant stars were derived from the synthetic spectra library by Pickles (1998), convolved with the S-PLUS transmission curves in the AB magnitude system (Oke & Gunn 1983). It is important to note that in these diagrams, the magnitudes for the MS correspond to `PStotal`, while for the GDS sources they correspond to PSF photometry.

The identification of objects is based on the method successfully applied by Witham et al. (2006, 2008) to the IPHAS catalog, since similar filters are also available in S-PLUS: r, $J0660$, and i. Similar technique was also used by Scaringi et al. (2013); Wevers et al. (2017); Monguió et al. (2020); Fratta et al. (2021) to reveal Hα excess sources.

We first generated (r - $J0660$) versus (r - i) diagrams for each magnitude bin in each field and then attempted to fit the regions predominantly occupied by main-sequence and giant stars using a linear regression model. After this, we applied an iterative σ-clipping technique, where data points more than several σ away from the fitted line were excluded in successive iterations to refine the fit. This process primarily aimed to remove outliers, ensuring that the final fit closely follows the bulk of the non-emitting stars, and was applied to the MS fields. Objects with Hα emission typically exhibit an excess in (r - $J0660$), causing them to appear above the main stellar loci in these plots. Therefore, it is expected that objects with Hα signatures will be located above these fitted lines. For fields in the MS with low stellar density, mostly those outside the Galactic plane, this fit often works well (as illustrated in Figure 4). However, many fields of the GDS display (at least) two distinct stellar loci in the color–color plane, resulting from differential reddening and/or contributions from both main-sequence stars and giants, where the fit is likely to align with the reddened locus (also illustrated in Figure 5).

To address this aspect in the GDS, we followed the procedure implemented by Witham et al. (2008): we selected the objects above the initially fitted line and iteratively adjusted the fit, moving it upwards towards the uppermost locus of points in the color–color diagram. As shown in Figure 5, this upper locus generally corresponds to the unreddened main sequence. In cases where the final fit is poorer than the initial one (e.g., in fields containing only a single stellar locus), we reverted to the initial fit. Once the appropriate fit for each magnitude bin was established, we identified objects significantly above the fit as likely Hα excess candidates. During this process, we examined the color–color diagram for each field and bin to ensure the fit was suitable, and found that, in general, 2 to 3 iterations were sufficient to locate the upper locus. This method ensures that objects exhibiting excess in Hα emission should adhere to the specified criterion:

$$(r - J0660)_{\rm obs} - (r - J0660)_{\rm fit} \geq C \times \sigma_{\rm est}, \quad (1)$$

where $(r - J0660)_{\rm obs}$ denotes the observed color difference between the r and $J0660$ bands, $(r - J0660)_{\rm fit}$ represents the color difference predicted by the linear regression fit, C is a constant parameter set to 5, and $\sigma_{\rm est}$ is the estimated standard deviation of the residuals around the fit, defined as:

$$\sigma_{\rm est} = \sqrt{\sigma_{\rm s}^2 + (1-m)^2 \times \sigma^2_{(r-J0660)} + m^2 \times \sigma^2_{(r-i)}}, \quad (2)$$

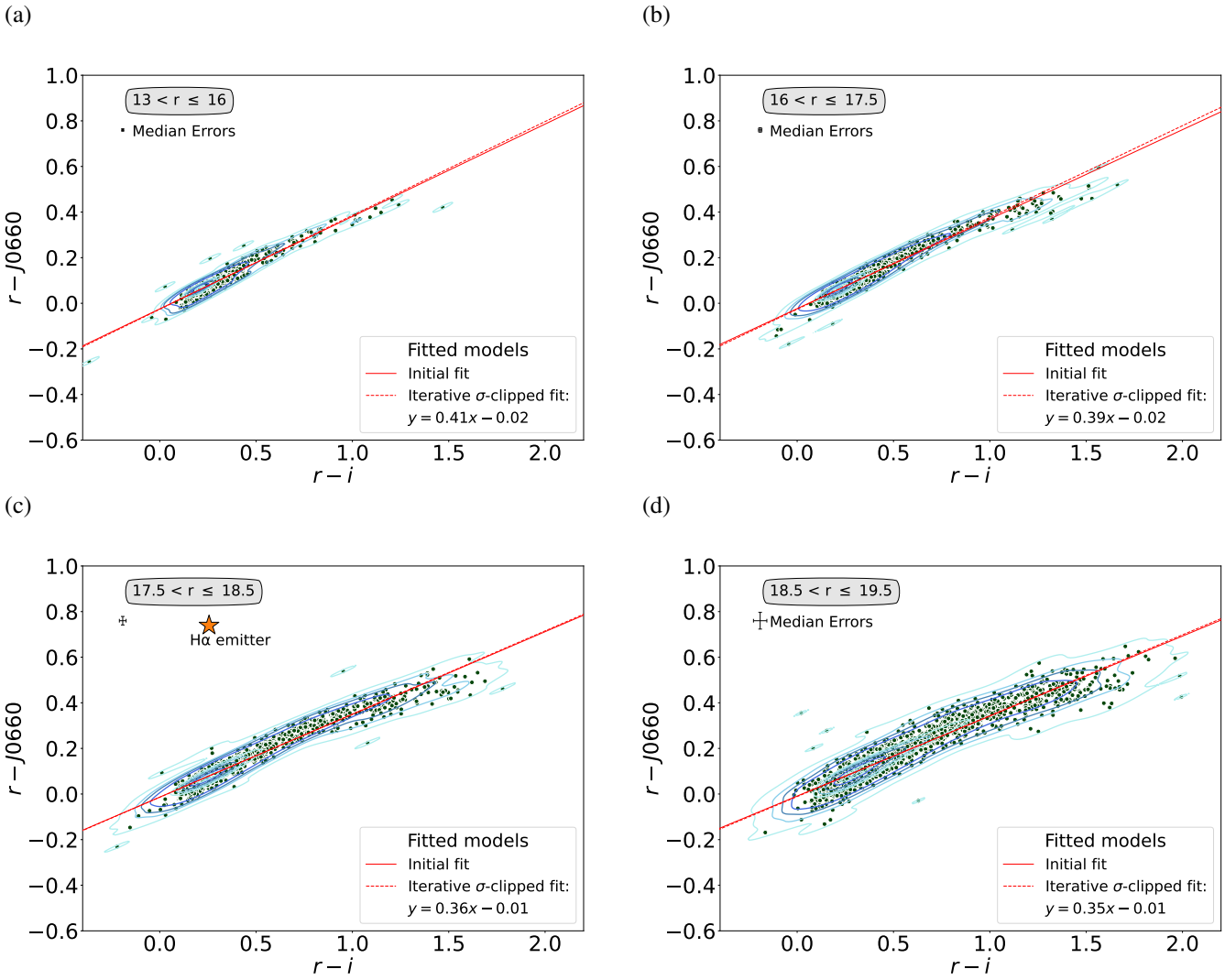

Fig. 4. Illustration of the selection criteria used to identify strong emission-line objects via color-color plots. The data shown here are from the S-PLUS field STRIPE82-0142, split into four magnitude bins, as displayed in the four panels. The thin red continuous lines show the initial linear fit to all data points (in green), while the dashed red line is the fit after applying iterative σ-clipping. Objects selected as Hα emitters are located above the dashed line. The orange star in panel (*c*) represents the cataclysmic variable (CV) FASTT 1560 (S-PLUS ID: DR4_3_STRIPE82-0142_0021237), highlighted here as an example of an Hα emitter identified by our criteria.

where σ_s represents the root mean squared value of the residuals around the fit, $\sigma_{(r-i)}$ denotes the error in the color index between the r and i bands, $\sigma_{(r-J0660)}$ denotes the error in the color index between the r and $J0660$ bands, and m represents the slope of the linear regression fit. The fits were performed using the `astropy.modeling` library [7].

Figure 4 illustrates the procedure applied to one field in the MS (STRIPE82-0142). The iterative approach was used for each individual field, with solid red lines indicating the initial fit. Sources showing $J0660$ excess or lying significantly above the stellar locus were identified as deviations from these fitted lines. The large orange star in panel c of Figure 4 represents a known Hα emitter (CV, FASTT 1560, Abril et al. 2020) that lies significantly above the stellar locus, with $(r - J0660) > 0.5$. Figure 5 shows the same procedure applied to the GDS. The red lines indicate the initial fit, while the black dashed lines represent the final iterative fits.

[7] `https://docs.astropy.org/en/stable/modeling/index.html`

4. Results and Analysis

Our objective is the identification of Hα excess sources within the S-PLUS footprint, leveraging the unique filter system of the survey. This effort resulted in 3 637 outliers for the MS and 3 319 for the GDS. The distribution of the sources with excess Hα emission in the $(r - J0600)$ versus $(r - i)$ color-color plane is depicted in Figure 6. Square light orange symbols represent objects with Hα excess identified in the MS, while greenish circle symbols denote those found in the GDS. All the sources placed above the locus of the main and giant stars exhibit an excess in the $J0600$ filter, attributed to the Hα excess. The broad distribution of sources on the color-color diagram of $(r - J0660)$ and $(r - i)$ indicates the selection of several types of Hα sources. These sources are likely associated with PNe, CVs, SySt, YSOs, Be stars, as well as extragalactic compact objects like QSOs and galaxies, among others (see Figure 2 of Gutiérrez-Soto et al. 2020).

The fractional contribution of different classes of sources to the overall sample was evaluated by cross-matching the ob-

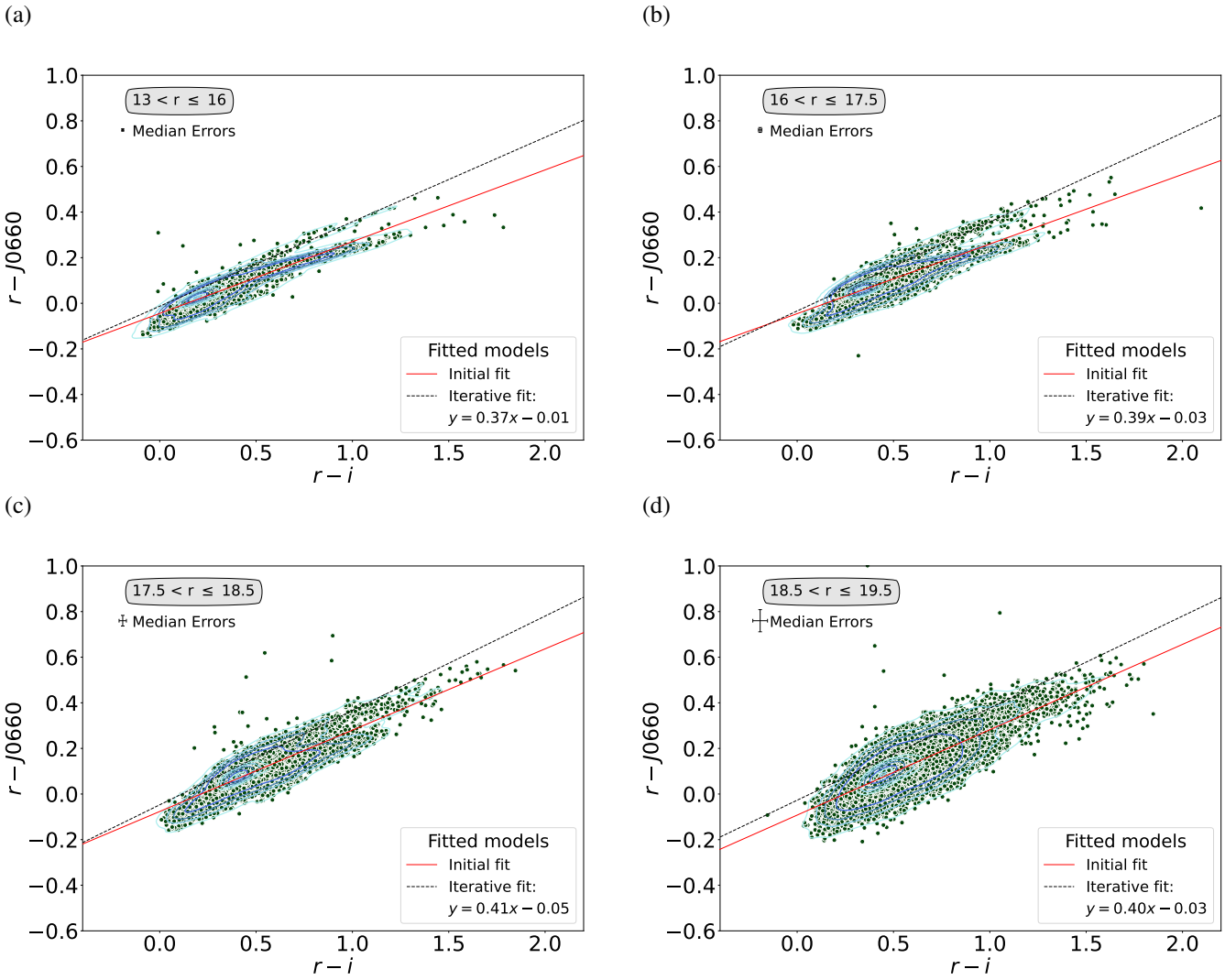

Fig. 5. Color-color diagram of the stars locate at field SPLUS-d288 like those found in Fig. 4. The red lines represent the original fit to all data, while the black dashed lines represent the final fits to the upper locus of points, obtained by applying an iterative fitting process to the initial fit.

jects' list with the SIMBAD database[8]. Optical spectra available in the Sloan Digital Sky Survey (SDSS; York et al. 2000) and in the Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST; Wu et al. 2011) were also explored. In all cases, positive matches between the different catalogs were considered for those sources that have an angular distance on the sky-plane within a given limit ($d_{max,proj}$). Verification of the photometry and assessment of Hα excess in the selected objects within the disk area were conducted by cross-matching the Hα source list identified in S-PLUS with photometric data from VPHAS+ DR2.

4.1. Matches with SIMBAD sources

We identified a total of 1 263 positive matches between our catalogs of Hα compact excess sources and the SIMBAD database, assuming a search radius of $d_{\rm max,proj} = 2$,arcsec for the MS and 1 arcsec for the GDS. In the MS, the identified objects primarily fall into categories such as variable stars, predominantly cataclysmic variables and/or candidates (CataclyV*), eclipsing binaries and/or candidates (EB*), and RR Lyrae Variables (RRLyr), as well as various kinds of stars including normal stars, white dwarfs, and/or candidates (WD*). Additionally, extragalactic compact sources that exhibit redshifted lines coinciding with the *J*0660 filter, simulating the Hα emission line, are also present. These include AGNs, Seyfert galaxies, QSOs, and other objects. It is important to note that the presence of redshifted emission lines in extragalactic sources may contribute to the identification of some of these objects as Hα-excess candidates (see Table 2 for details).

For the GDS, the identified categories include emission-line stars (Em*), young stellar objects (YSO) and candidates, which encompass T Tauri (TTau*) and Herbig Ae/Be (Ae*) star candidates. Additionally, variable stars such as cataclysmic variables (CataclyV*), eclipsing binaries (EB*), and RR Lyrae variables (RRLyr) are found, along with objects exhibiting nebular components, such as planetary nebula (PN) candidates, novae, and reflection nebulae (RfNeb), among others. As shown in Table 2, the highest number of sources in the disk belong to the Em* and young stellar objects category, reflecting the active star formation processes in the Galactic disk.

An important consideration regarding the SIMBAD matches is that in the MS, numerous extragalactic sources with emission lines are selected due to the mapping of high latitudes

[8] http://simbad.u-strasbg.fr/simbad/

Table 2. Summary of the positional cross-match results between the S-PLUS list of Hα source and the SIMBAD database. A search radius of 2 arcsec was used for the MS, while 1 arcsec was used for the GDS. The first column indicates main object categories, the second column lists SIMBAD object types, and the third column indicates the number of objects in each category.

Main Type	Associated SIMBAD Types	Number of S-PLUS Objects with SIMBAD Match
	Main Survey	
Stellar Binary System	CataclyV*, CV*_Candidate, RSCVn, EB*, EB*_Candidate, SB*_Candidate	353
Variable Star	PulsV*, V*, PulsV*delSct, RotV*, RRLyr	139
Star	Star, Blue, low-mass*, WD*, WD*_Candidate, PM*, BlueStraggler	47
Radio Source	Radio, Radio(cm), RadioG	9
Active Galactic Nucleus (AGN)	AGN, AGN_Candidate, Seyfert_1	23
Quasar	QSO, QSO_Candidate	143
Galaxy	Galaxy	9
Other	Hsd_Candidate, Pec*, AGB*, MIR	8
Total		**731**
	Disk	
Emission-line star	Em*, Be*	125
Young stellar object	YSO, YSO_Candidate, Orion_V*, TTau*_Candidate	102
Stellar Binary System	CataclyV*, CV*_Candidate, RSCVn, EB*, EB*_Candidate, SB*	146
Variable star	PulsV*delSct, PulsV*, LPV*, LP*_Candidate, Mira, RRLyr, V*, V*?_Candidate, BYDra	43
Star	Star, **, RGB*, C*, WD*_Candidate	104
Nebula	PN?_Candidate, RfNeb, Nova	3
Other	EmObj, Hsd_Candidate, deltaCep, Cepheid_Candidate, Transient, X	9
Total		**532**

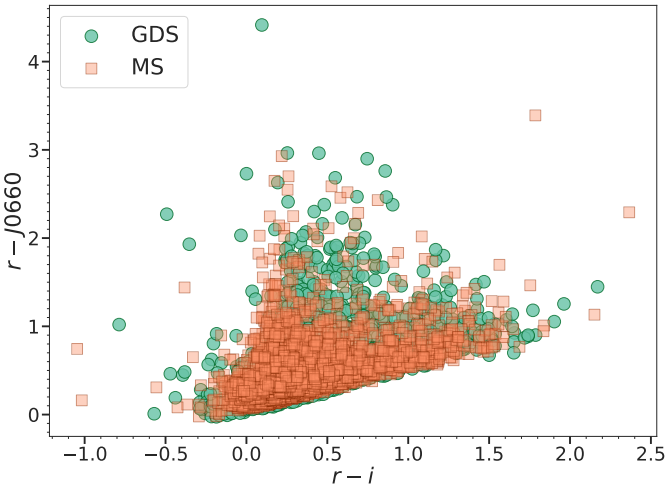

Fig. 6. The color-color diagram shows the distribution of Hα-feature sources in the $(r-J0660)$ versus $(r-i)$ color-color space. The data is divided into two populations: GDS and MS representing distinct galactic components. The GDS population, depicted by filled circles in greenish color, corresponds to Hα excess sources associated with disk structures. In contrast, the MS population, represented by square symbols in light orange, includes Hα excess sources primarily located in the direction of the Galactic halo and extragalactic sources.

in the southern sky. Conversely, for the GDS, no extragalactic sources have been selected. While the MS emphasizes extragalactic sources and diverse stellar populations, the disk region primarily showcases young stellar objects and variable stars, indicative of ongoing star formation and stellar evolution processes. In both regions, variable stars such as EB*, among others, are also present. The results are described below and listed in Table 2.

In our analysis of Hα-excess sources, variable stars such as RR Lyrae stars and eclipsing binaries are frequently detected due to their tendency to exhibit significant photometric deviations in the Hα-related bands. It is important to highlight that RR Lyrae stars, which are known for their characteristic spectral features, often show Hα absorption lines. This occasionally causes them to be identified as outliers in our selection, as our criteria are sensitive to any significant deviation from expected stellar colors, whether it involves emission or absorption features. Moreover, the use of the S-PLUS filter system, with its 12 sequential filters, plays a role in detecting these short-period variables. Since both RR Lyrae and eclipsing binaries have short periods (typically hours to days), the sequential observation through S-PLUS's filters can capture these stars at different phases of their variability. This effect can lead to apparent Hα-excess due to the changes in brightness across different bands during the observation sequence. In particular, eclipsing binaries can display Hα emission due to complex interactions between the stellar components and their surrounding material. This phenomenon has been observed in systems such as the eclipsing binary VV Cephei, where periodic variations in Hα emission occur during different phases of the eclipse (Pollmann et al. 2018).

An important observation is that our selection criteria have predominantly excluded extended sources. In the MS, only 23 AGN and 9 galaxies were identified, making up approximately 3.1% and 1.2% of the total 731 SIMBAD matches (see Table 2), respectively. Additionally, we identified 143 QSOs, representing about 19.6% of the total matches. These percentages highlight the effectiveness of our selection criteria in isolating compact sources with significant Hα excess, while also illustrating the relative proportions of different astrophysical categories identified in our survey.

4.1.1. Redshifted Lines Mimicking the Hα Emission

According to the classification in the literature, near 20% of the Hα sources in our sample are classified as QSOs. It is important to note that the excess observed in the $J0660$ filter is due to QSOs whose emission lines are redshifted to the wavelength range of this filter. For instance, lines such as Hβ, Mg II 2798 Å, C III] 1909 Å, and C IV 1550 Å can contribute to this excess (see Gutiérrez-Soto et al. 2020 and the bottom of Figure 1 of Nakazono et al. 2021, which shows the main emission lines of a quasar at different redshifts and indicates which of those fall within the $J0660$ filter).

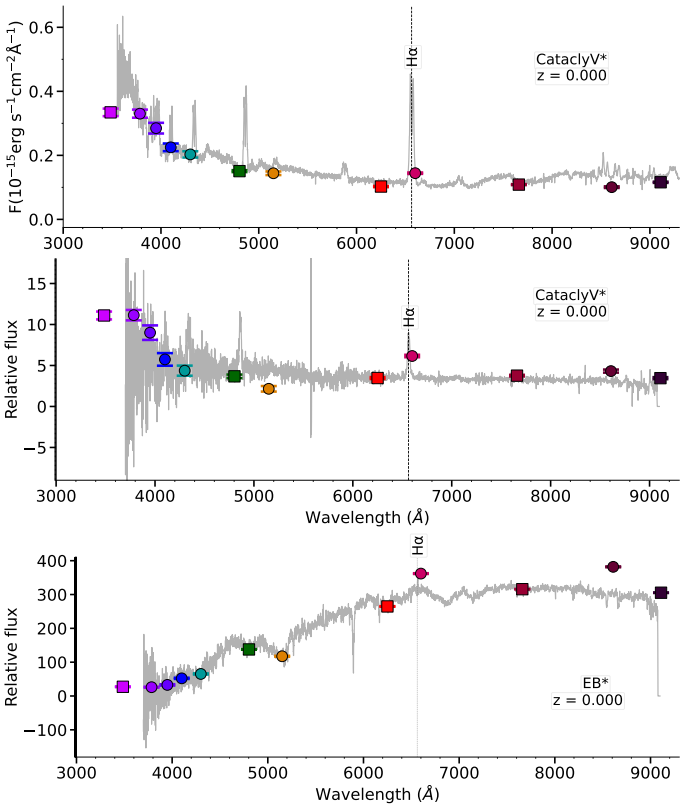

Fig. 7. Spectra of three objects identified as Hα excess sources using our methodology. The top panel displays the SDSS spectrum, while the middle and bottom panels show LAMOST spectra. The colored symbols correspond to S-PLUS photometry in flux units for the following filters (from left to right): *u*, *J*0378, *J*0395, *J*0410, *J*0430, *g*, *J*0515, *r*, *J*0660, *i*, *J*0861, and *z*. Square symbols represent broad-band filters, while circle symbols denote narrow-band filters. According to SIMBAD, the objects in the top and middle panels–SDSS ID J113722.24+014858.5 and LAMOST ID J232551.47-014023.5–are classified as cataclysmic variables. The bottom panel shows an eclipsing binary star with weak Hα emission (LAMOST ID: J012119.09-001950.0). The dashed line marks the position of the Hα wavelength.

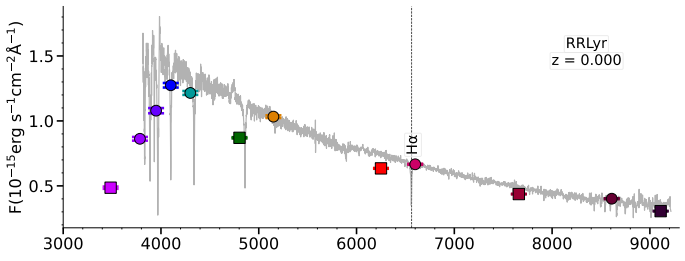

Fig. 8. SDSS spectrum and S-PLUS photometry of the RR Lyrae star SDSS J010045.13-010212.2, showing an Hα absorption line.

This particular population of apparent Hα emitters includes AGNs, Seyfert 1 galaxies, and other emission-line galaxies. In particular, within the redshift range $0.306 < z < 0.376$, lines such as Hβ and [O III] 4959, 5007 Å are redshifted into the *J*0660 filter.

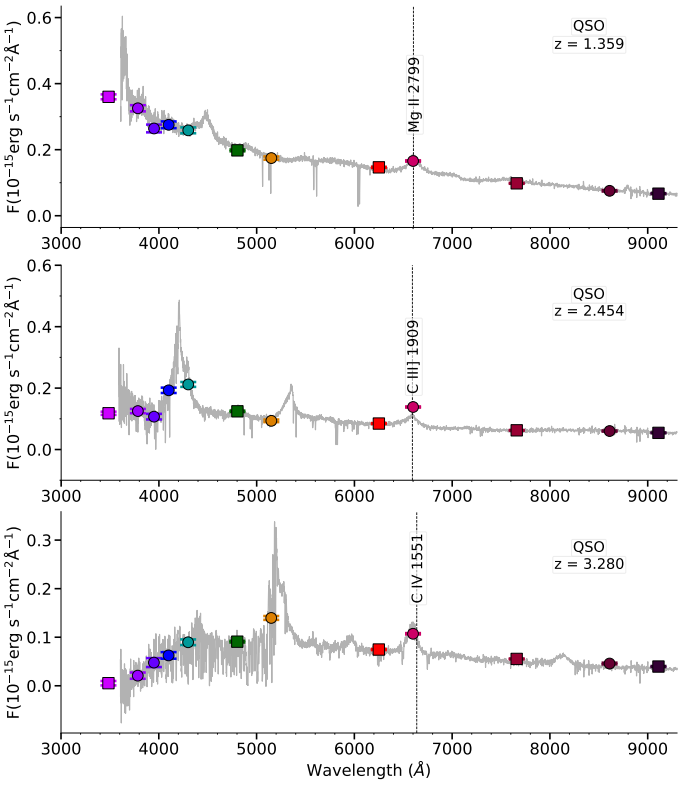

Fig. 9. S-PLUS photometry and SDSS spectra of three QSOs with redshifts of 1.359, 2.454 and 3.280 (top to bottom) selected as Hα excess sources. At these redshifts, the emission lines Mg II λ2799, C III] λ1909 and C IV λ1551 are detected in the *J*0660 filter. The SDSS IDs of the sources are: J235157.58+003610.5, J220529.34-003110.7, and J224539.94-002419.6.

4.2. Matches with SDSS and LAMOST

Our list of Hα-excess sources identified in the MS was cross-matched with the DR18 SDSS catalog (Ahumada et al. 2020) and the DR9 LAMOST catalog, using a 2 arcsec radius. These cross-matching identified 212 common sources (138 from SDSS and 74 from LAMOST). The procedure was restricted to the MS due to its overlap with SDSS and LAMOST areas, unlike the S-PLUS Galactic disk survey. It is noteworthy that some Hα-excess sources detected by our algorithm may exhibit transient behavior, meaning that Hα-excess features might be present in spectra from one survey (SDSS or LAMOST) but not in others (S-PLUS), or vice versa. This variability is attributed to differences in observational epochs and conditions across the surveys. Upon spectroscopic examination, approximately 60% of these sources exhibited emission lines, which might include redshifted lines other than Hα, while about 30% showed Hα-related absorption features.

Most of the objects with available spectroscopic information in SDSS and LAMOST correspond to CVs, QSOs, AGN, and variable stars. A more detailed spectroscopic characterization of these sources is out the scope of this paper. Also, it is worth noticing that there is a number of objects without a conclusive classification.

Figure 7 presents the SDSS (upper) and LAMOST (lower) spectra, along with the corresponding S-PLUS photometry (colored symbols) for two known cataclysmic variables (CVs) and one eclipsing binary, respectively. The excess in the *J*0660 filter is evidently produced by the Hα line. Note that the bluer emis-

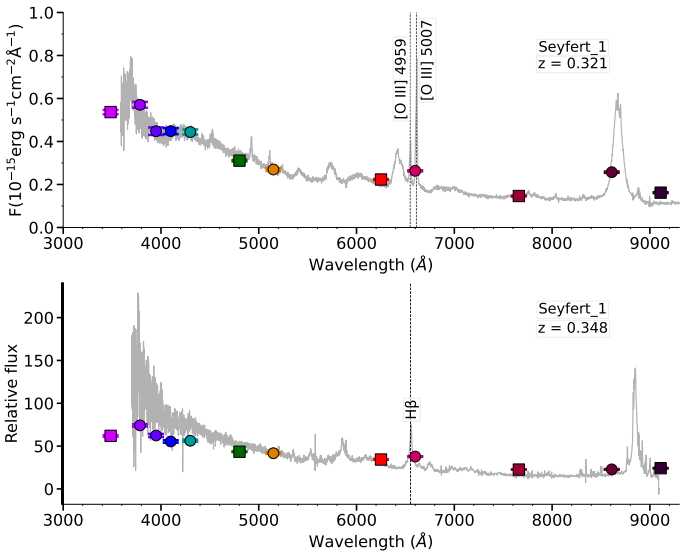

Fig. 10. As Figure 9 but for sources with redshifts of $z = 0.321$ and $z = 0.348$. At these redshifts, the lines [O III] $\lambda\lambda 4959, 5007$ doublet and Hβ line are detected in the narrow band filter, generating an Hα excess. The spectra of the sources are from SDSS (upper) and LAMOST (bottom) with IDs SDSS J231742.60+000535.1 and LAMOST J033429.44+000611.0, respectively.

sion tends to be more intense, which is consistent with the expected behavior of CVs showing strong Balmer series emission. Bottom panel of Figure 7 displays the LAMOST spectrum and S-PLUS photometry of an eclipsing binary. The spectra exhibit weak Hα emission, which is effectively captured by the narrow $J0660$ filter of S-PLUS.

Figure 8 shows the SDSS spectra and S-PLUS photometry of an RR Lyrae star with Hα in absorption. The absorption feature in Hα affects both the r-band and the $J0660$ filter. The apparent Hα excess observed in the $(r - J0660)$ color index for sources with Hα absorption is due to the differential effect of the absorption feature on the broadband r filter and the narrowband $J0660$ filter. The Hα line lies within the r-band, so the absorption feature reduces the total flux detected, making the r band appear fainter. In contrast, the $J0660$ filter, shows a less pronounced reduction in flux. This difference results in a more negative $(r - J0660)$ color index, creating an apparent Hα excess. This photometric effect is important for identifying Hα excess sources, as it indicates the presence of Hα variations, even in absorption, within various stellar objects (Fratta et al. 2021). Furthermore, most of the Hα absorption line objects in the MS (relatively high latitude) are RR Lyrae stars, as confirmed by SIMBAD, which lists 111 RR Lyrae stars in our sample. We also explore the distribution of RR Lyrae stars in the $(r - J0660)$ versus $(r - i)$ color diagram, finding that these variable stars span $(r - J0660)$ values between -0.4 and 0.6. This means that a population of these stars have $(r - J0660) > 0$, indicating their selection. For more details, see Figure A.1 in the Appendix A.

Figure 9 presents examples of SDSS spectra for three QSOs, where the $J0660$ filter captures emission from different redshifted lines. For the QSO in the upper panel (redshift ∼1.36), the excess corresponds to the Mg II 2798 Å line. In the middle panel (redshift ∼2.45), the excess is due to the C III] 1909 Å line. Finally, for the QSO in the bottom panel (redshift ∼3.28), the C IV 1550 Å line produces the observed excess. These plots demonstrate how the $J0660$ filter captures redshifted emission lines for QSOs at various redshifts.

Other extragalactic objects for which we found spectra in SDSS and LAMOST include AGNs. For example, Figure 10 displays the spectra of two nearby AGNs with redshifts of approximately $z \simeq 0.35$ (top) and $z \simeq 0.32$ (bottom). In the first, the Hβ emission line falls within our narrowband filter. For the second source, with $z \simeq 0.32$, the doublet [O III] 4959, 5007 Å emission lines lie in the $J0660$ filter, resulting in an observed excess.

The analysis of individual spectra reveals distinct Hα line features, including both emission and absorption at expected wavelengths, offering valuable insights into the physical characteristics and evolutionary stages of the objects. The spectral confirmation rates we present provide a conservative estimate of the selection purity. This is because our algorithm targets Hα-excess sources, not strictly Hα emitters. Thus, objects with excess in the $J0660$ filter are selected as outliers, even if they lack a prominent Hα emission line. By referring to "Hα-excess" rather than "Hα-emitters," we highlight that our selection is based on photometric excess in the $J0660$ filter, rather than solely on strong Hα emission.

4.3. Evaluation of Photometric Color Consistency Between S-PLUS and VPHAS+

We performed a comparative analysis of PSF photometric colors between the S-PLUS data from the GDS and those provided by VPHAS+ DR2[9]. For the crossmatching, we considered a radius of 1" and ended up with a number of 793 matches. We computed the differences in two key color indices: $r - i$ and $r - H\alpha$. Specifically, we investigated the median difference and the median absolute deviation (MAD) of these colors to assess the consistency and agreement between the two surveys. It is worth noting that VPHAS+, like S-PLUS, employs the r, i, and a narrowband filter (NB-659) designed to detect the Hα line, facilitating a meaningful comparison of Hα emission.

The comparison of colors reveals important insights into the consistency and reliability of S-PLUS photometry (see Figure 11). The median difference in the $r-i$ color between S-PLUS and VPHAS+ was −0.21, with a MAD of 0.07. For the $r - H\alpha$ color, the median difference was 0.02 with a MAD of 0.27. These results indicate a systematic offset between the photometric colors of the two surveys, which is within the expected range considering differences in instrumentation and filter systems.

A key factor contributing to the differences in the $r - H\alpha$ color index is the distinct characteristics of the Hα filters used in S-PLUS and VPHAS+. The S-PLUS Hα filter ($J0660$) has an effective wavelength of 6614 Å and a width of 147 Å whereas the VPHAS+ NB-659 filter has an effective wavelength of 6588 Å and a width of 107 Å. These differences can significantly affect the measurement of Hα excess, as the narrower VPHAS+ filter captures a more restricted range of wavelengths, potentially leading to higher precision. The broader S-PLUS filter, on the other hand, may include additional continuum emission, affecting the photometric measurement. Additionally, the exposure times in the two surveys differ, with VPHAS+ using a 120-second exposure and S-PLUS using a 290-second exposure. The longer exposure time in S-PLUS allows for greater sensitivity to faint sources and potentially higher signal-to-noise ratios (SNR), contributing to the observed differences in photometric colors.

Despite the observed systematic differences, the MAD values suggest that the photometric measurements from both surveys exhibit good agreement. This consistency is crucial for

[9] More detailed information about the VPHAS+ survey can be found at: `https://www.vphasplus.org/`

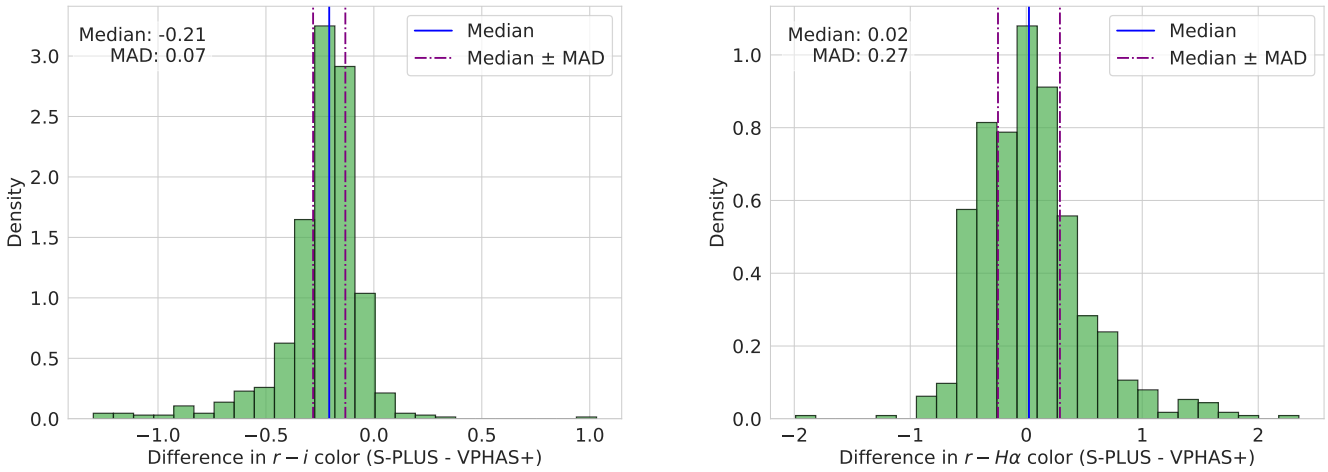

Fig. 11. Histograms illustrating the discrepancies in the photometric colors $r - i$ and r - Hα between the S-PLUS and VPHAS+ surveys. The left panel depicts the differences in the $r - i$ color, while the right panel shows the differences in the (r - Hα) color. Both histograms reveal significant differences for the stars included in both surveys. The median value and median absolute deviation (MAD) for each color discrepancy are provided, offering insights into the agreement between the two datasets.

take out the parentheses

cross-referencing and integrating datasets from different surveys for comprehensive astrophysical studies. The observed differences in photometric colors may result from various factors, including differences in filter characteristics, photometric calibration, and data processing techniques. Further investigations are warranted to better understand these factors' contributions to the observed discrepancies.

4.4. Hα Excess Source Distributions

The upper panel of Figure 12 presents a histogram of the r-band magnitude distribution for all objects in our study from the MS. The normalized density facilitates comparison between different subsets. The blue curve represents Hα excess objects, while the salmon curve represents all stars from the MS. The magnitude distribution for Hα excess sources shows a higher concentration at intermediate magnitudes. The lower panel of Figure 12 focuses on the r-band magnitude distribution sources for the subset of Hα excess objects in the disk. A noticeable large number of sources with Hα excess have magnitudes in the r-band between 13 and 13.5, something that we do not see in the stars of GDS. This implies that Hα excess objects could be intrinsically more luminous or closer to us than the general population of all stars. However, these stars are closer to the saturation limit. Therefore, we recommend exercising caution with all sources in our sample that have an r-band magnitude less than 13.5.

Figure 13 shows the distribution of all Hα excess sources in Galactic latitude and longitude, along with a zoomed-in view of the GDS in the bottom panel. The distribution of objects in Galactic longitude for the MS (left panel of Figure 14) indicates that the blue bars, representing Hα excess sources, are relatively evenly spread across the Galactic longitude, similar to the general population of stars from the MS (pink bars). Peaks are observed around Galactic longitudes of 15°, 50°, and 270°, which are also present in the general star population of the MS.

The bottom panel of Figure 13 and the right panel of Figure 14 show the distribution of objects in Galactic longitude specifically within the Galactic disk. There is a noticeable concentration of Hα excess sources at specific longitudes, particularly around 243°. Additionally, there are small peaks around 225°

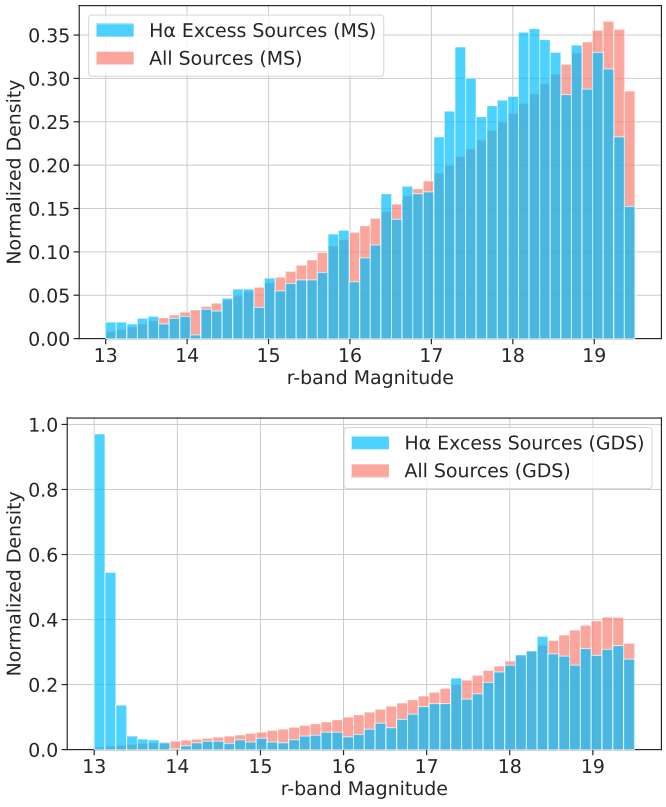

Fig. 12. *Upper panel*: Distribution of r-band magnitudes for Hα excess sources (blue curve) compared to all the stars (salmon curve) in the MS. *Lower panel*: Distribution of r-band magnitudes for Hα excess sources (blue curve) in the GDS compared to all stars (salmon curve).

and 268° in Galactic longitude. While Hα excess sources follow a distribution similar to that of all stars, the peaks are more pronounced for Hα excess sources.

It should be noted that the observed concentrations of Hα excess sources in certain Galactic regions may be influenced

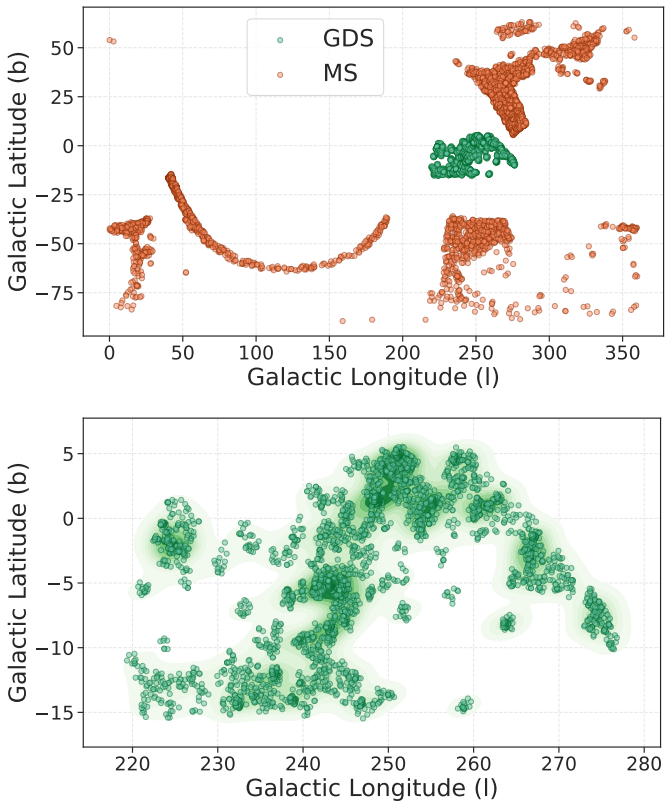

Fig. 13. Distribution of Hα excess objects in Galactic longitude and latitude coordinates. The *upper panel* shows all the Hα sources selected, and the *lower panel* is a zoomed-in view of the GDS.

Galactic disk survey

by the uneven sky coverage of the S-PLUS survey (see Figure 6 in Herpich et al. 2024). In the MS, this effect may be more pronounced due to the lower star formation activity in high-latitude regions. For the GDS, the concentrations may be influenced by both the presence of star-forming regions and the patchy sky coverage of the survey. This limitation may lead to uneven sampling of the Galactic plane, potentially over- or under-representing the observed numbers in specific regions. Therefore, these peaks should be interpreted with this survey limitation in mind.

5. Machine Learning Approaches

In this section, inspired by the goal of separating Galactic sources from extragalactic ones in our Hα excess list, we applied machine learning approaches. Our list of Hα excess sources selected in the MS of S-PLUS naturally includes extragalactic compact objects with redshifted lines detected in the *J*0660 filter. To classify the sources in our Hα excess list, we utilized the multi-band coverage provided by S-PLUS optical photometry. To achieve this, we employed two unsupervised machine learning algorithms: UMAP and HDBSCAN. UMAP is used to reduce the dimensions of our data and perform a feature extraction, while HDBSCAN classifies the data based on the results from UMAP. We conducted two experiments: one using the 66 colors generated from the 12 S-PLUS filters, and a second one by adding filters from the Wide-Field Infrared Survey Explorer (WISE Wright et al. 2010). **This classification helps identify specific types of objects for subsequent spectroscopic follow-up.** Additionally, we used a Random Forest algorithm to identify important features and construct color-color diagrams to separate the classes of objects identified by HDBSCAN. This methodology is applied to the list of Hα excess sources obtained from the MS of S-PLUS. **The classification results also provide the basis for defining tentative color criteria, which can be used to refine the separation between different classes of Hα excess sources, based on the new color-color diagrams proposed here.**

5.1. Dimensionality Reduction and Clustering

5.1.1. UMAP

Uniform Manifold Approximation and Projection (UMAP; Becht et al. 2018; McInnes et al. 2020) is a dimensionality reduction algorithm designed to handle high-dimensional data while preserving its underlying structure. Unlike some other techniques, UMAP is based on a mathematical framework that combines aspects of Riemannian geometry and algebraic topology. This enables UMAP to capture both local and global relationships within the data. UMAP aims to create a low-dimensional representation that retains the intricate nonlinear relationships present in the original high-dimensional features. This process involves constructing a high-dimensional graph representation of the data and then optimizing a low-dimensional graph to match it. By doing so, UMAP effectively preserves the essential information and structure encoded in the data. This makes UMAP particularly well-suited for datasets where parameters exhibit complex nonlinear behavior. In our analysis, we use UMAP to reduce the dimensionality of our input space, consisting of 66 colors and additional WISE bands, while retaining essential information encoded in the data

For the implementation of the algorithm, we used the Python package `umap`[10]. UMAP has three key hyperparameters: `n_neighbors`, `n_components`, and `min_dist`.

The `n_neighbors` parameter balances local versus global structures in the data by setting the number of neighboring points UMAP considers for each data point when learning the manifold structure. Low values of `n_neighbors` cause UMAP to focus on very local structures, while higher values make UMAP look at larger neighborhoods, potentially losing fine details in favor of capturing broader patterns.

The `n_components` parameter, similar to the parameter used in standard dimension reduction algorithms in the `scikit-learn` package (Pedregosa et al. 2011), allows us to set the number of dimensions in the reduced space into which we will embed the data. **`scikit-learn` is a widely used Python library for machine learning, built on top of SciPy, and distributed under the 3-Clause BSD license. It provides implementations for many state-of-the-art machine learning techniques, making it a versatile tool for data analysis and modeling.** great!

The `min_dist` parameter controls how closely UMAP can pack points together in the low-dimensional representation. Lower values result in clumpier embeddings, which are useful for clustering and capturing fine topological structures, while higher values focus on preserving broader topological structures.

[10] For more details, see `https://umap-learn.readthedocs.io/en/latest/index.html`

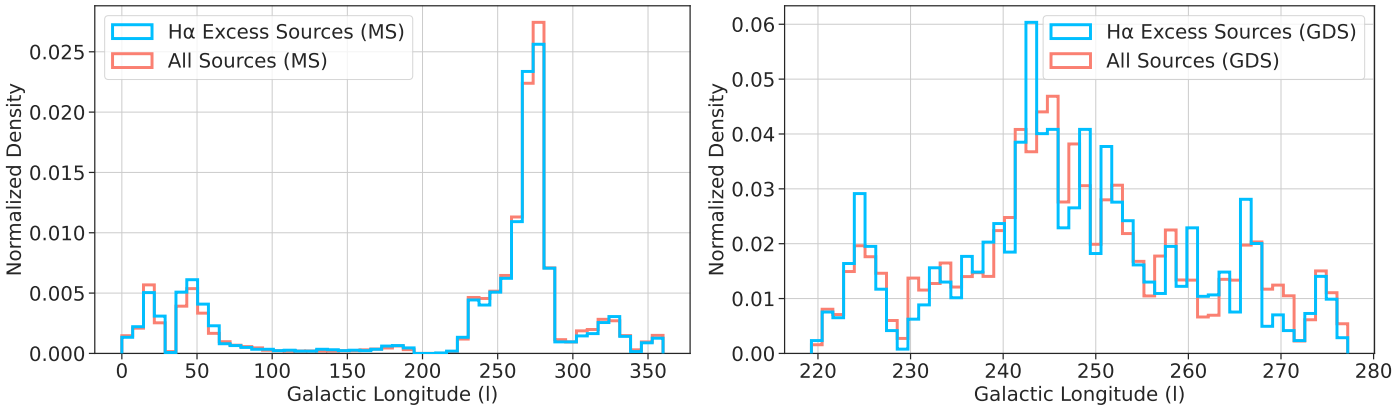

Fig. 14. Distribution of the objects in galactic longitude for Hα excess sources (blue bars) and all stars (salmon bars) for the MS (*left panel*) and the GDS (*right panel*).

5.1.2. HDBSCAN

After obtaining a new system of reduced variables that condenses all the information from the original variables, we utilized HDBSCAN to identify clusters within the data. This clustering approach complements the reduction achieved by UMAP, allowing for a comprehensive understanding of the underlying structure of the dataset.

Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN; Campello et al. 2013) is an unsupervised machine learning algorithm for clustering. It builds on the density-based spatial clustering of applications with noise (DBSCAN; Ester et al. 1996) by introducing a hierarchy to the clustering process, which allows for the extraction of "persistent" clusters from the hierarchical tree. HDBSCAN's main advantage over DBSCAN is its ability to find clusters of varying densities and shapes.

For this task, we adopted the Python implementation of `HDBSCAN`[11] (McInnes et al. 2017). The two most critical parameters are the "minimum cluster size" (`min_cluster_size`) and "minimum number of samples" (`min_samples`). The "minimum cluster size" refers to the smallest group size that is considered a cluster. The "minimum number of samples" determines how conservative the clustering will be; larger values result in more points being classified as noise, restricting clusters to denser areas.

HDBSCAN can also classify sources as noise if they do not fit well into any cluster based on these parameters. Additionally, the algorithm relies on a distance metric, such as Euclidean distance, to measure the distance between points and determine their density. The choice of metric can significantly affect the clustering results, as it influences how distances are computed and, consequently, how clusters are formed.

5.2. Classification Results

Our unsupervised UMAP model projects the data, and HDBSCAN identifies the clusters. To ensure high-quality photometry, we required errors below 0.2 mag in all filters, reducing the sample to 2181 MS objects. This step minimizes the impact of noisy measurements, improving the performance of UMAP and HDBSCAN. By focusing on reliable photometric data, we enhance the accuracy and robustness of the clustering, ensuring more reliable classifications of Hα excess sources.

[11] `https://hdbscan.readthedocs.io/en/latest/`

To perform cross-validation for selecting the optimal `n_neighbors` and `n_components` parameters in UMAP, we systematically explored a range of values for these parameters. The selection of parameters n_neighbors and n_components in UMAP is critical as it directly influences the quality of the reduced-dimensional representation. Initially, we conducted exploratory data analysis to visualize the dataset in reduced dimensions using various combinations of n_neighbors and n_components. This allowed us to qualitatively assess how well UMAP preserved the underlying structure of the data.

We employed two quantitative metrics to objectively evaluate the performance of different parameter combinations: the Silhouette Score (Rousseeuw 1987) and the Davies-Bouldin Index (Davies & Bouldin 1979).

The Silhouette Score measures how well-defined the clusters are in the reduced space. It quantifies how similar an object is to its own cluster (cohesion) compared to other clusters (separation). Higher scores indicate better separation between clusters, meaning that objects are well matched to their own cluster and poorly matched to neighboring clusters.

great !!!

The Davies-Bouldin Index evaluates the average similarity between each cluster and its most similar cluster. This internal evaluation metric assesses how well clustering has been performed using features inherent to the dataset. Lower values indicate better-defined clusters.

A grid of tests was constructued over a range of n_neighbors (5, 10, 15, 20, 30, 50, 70, 100) and n_components (2, 3, 4, 5, 10, 20, 50) values. For each combination, UMAP was applied followed by clustering using **KMeans (Lloyd 1982), and the metrics were computed to determine the optimal parameter set. The KMeans algorithm clusters data by attempting to separate samples into n groups of equal variance by minimizing a criterion known as inertia, or the within-group sum of squares. It requires the number of groups to be specified, for which we used the `n_components` from each UMAP computation. KMeans scales well to large numbers of samples and has been widely used in various fields for clustering tasks.**

do you know 1 or 2 ref from astrophysics, where it used.

5.2.1. Initial Analysis Using S-PLUS Photometry

For the first experiment, we used the 66 S-PLUS colors as input parameters and applied the metric evaluation method described above. These metrics include the silhouette score, which measures cluster cohesion, and the Davies-Bouldin

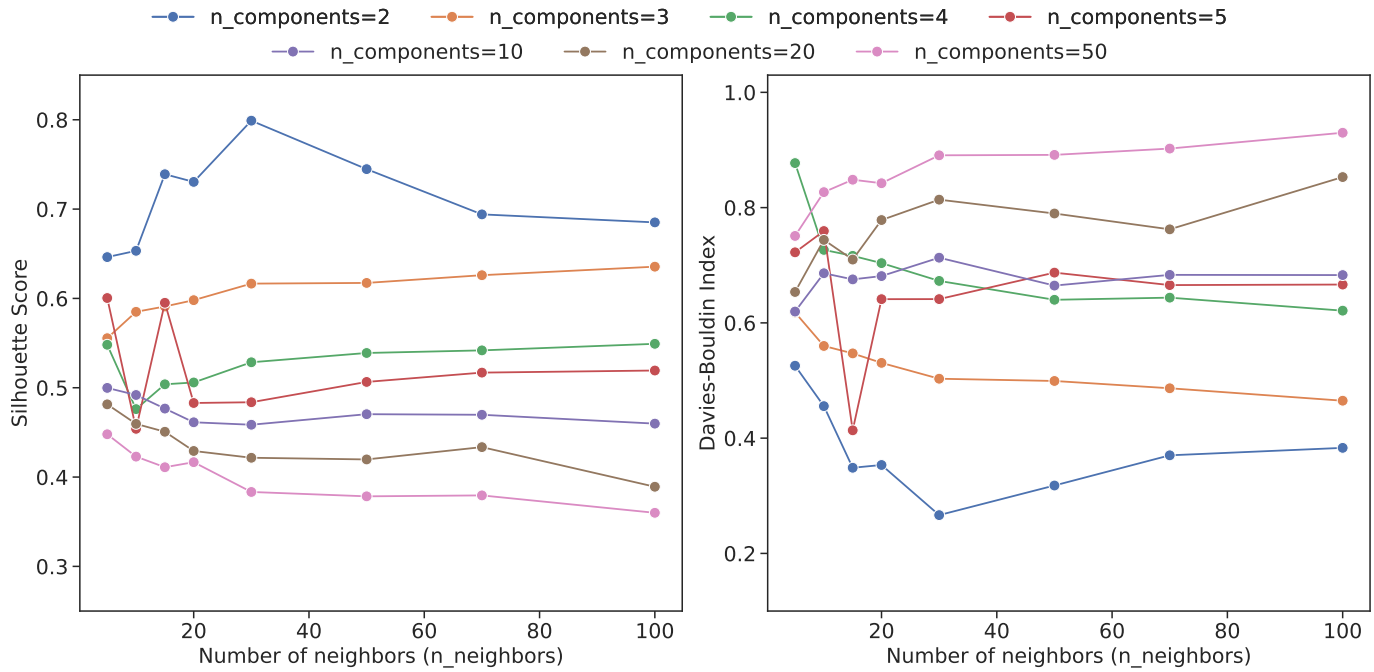

Fig. 15. Silhouette Score (left panel) and Davies-Bouldin Index (right panel) as functions of the number of neighbors (*n_neighbors*) for different values of UMAP components (*n_components*). Higher Silhouette Score values and lower Davies-Bouldin Index values indicate better clustering performance.

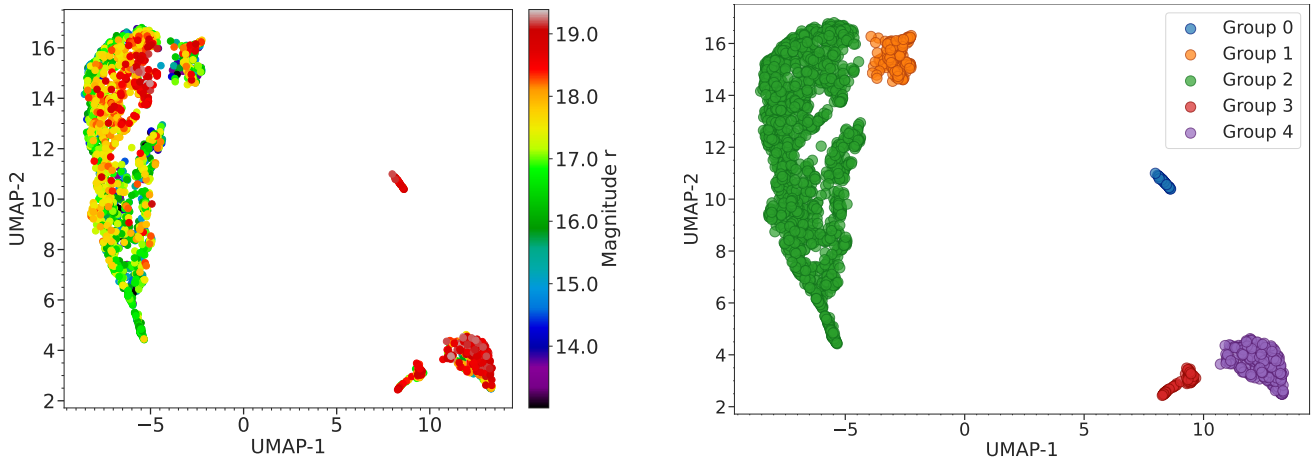

Fig. 16. UMAP dimension reduction applied to the MS from S-PLUS data. The left panel shows the UMAP result using only the S-PLUS colors as input parameters, with the color bar indicating the *r* magnitude. The right panel displays the result after applying HDBSCAN clustering, revealing five distinct groups.

Index, which assesses cluster separation. Lower values of the Davies-Bouldin Index and higher silhouette scores indicate better clustering performance. After evaluating various hyperparameter combinations, we identified that setting `n_neighbors` = 30 and `n_components` = 2 yielded the highest silhouette score (0.799) and the lowest Davies-Bouldin Index (0.266). These values were subsequently adopted as the optimal hyperparameters for our analysis. For the `min_dist` parameter, we used the default value of 0.1. Figure 15 illustrates the behavior of the silhouette score (left panel) and the Davies-Bouldin Index (right panel) as functions of the `n_neighbors` for each `n_components`. In the left panel, we observe that the silhouette score varies with `n_neighbors`, with the best performance achieved at `n_neighbors` = 30 and `n_components` = 2. The right panel shows the Davies-Bouldin Index, which decreases consistently for these hyperparameters, confirming their optimality.

Following dimensionality reduction with UMAP, the resultant variables were utilized to construct HDBSCAN models. We experimented various combination for the "minimum cluster size" and "minimum number of samples" parameters. **We ended up with the optimal value of 2 and 50, respectively. Euclidean metric was employed for distance calculations throughout.**

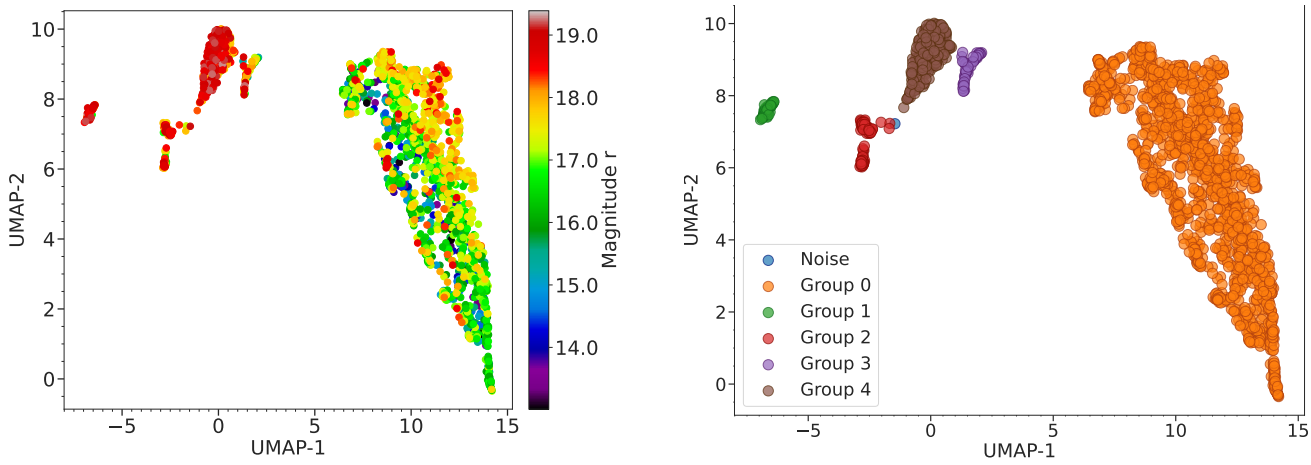

Fig. 17. Similar to Fig. 16, but with additional features created using the W1 and W2 bands from WISE. The left panel shows the UMAP result, with the color bar indicating the *r* magnitude, and the right panel shows the HDBSCAN clustering result, identifying five distinct groups.

The left panel of Figure 16 shows the distribution of the new variables in UMAP space, resulting from applying it to the 66 S-PLUS colors of the Hα excess objects for the MS. The color bar indicates the *r* magnitude, highlighting the bright and faint sources. Visually, it is possible to distinguish at least four groups, with the small clusters located in the upper left of the diagram tending to be fainter. The right panel of the figure shows the same plot but with the results of applying HDBSCAN using the parameters mentioned above. HDBSCAN identified four groups. Table 3 provides the number of objects in each group. To further understand the nature of each group, we examined their SIMBAD counterparts, which are also detailed in the Table 3.

Group 0 contains 58 objects, 22 of which are matched in SIMBAD. The majority are QSOs (19), with the remaining objects including one galaxy, one radio source, and one QSO candidate. This composition suggests that Group 0 primarily consists of extragalactic sources, with a redshift distribution peak around 2.45. It is worth noting that the spectral characteristics of QSOs (a type of AGN) differ significantly from typical galaxies; while starburst galaxies show strong extinction at blue wavelengths, the QSO spectrum rises sharply toward the blue, indicative of the high-energy processes associated with active galactic nuclei.

Group 1 contains 166 objects, 149 of which have entries in the database. This group is predominantly composed of RR Lyrae stars (107), followed by eclipsing binaries (19), various types of pulsating variables (9), and a few other stellar objects, including 2 QSOs. This group appears to represent objects with Hα in absorption, as it is well known that RR Lyrae stars exhibit Hα absorption lines.

Group 2 includes 1539 objects, 323 of which are cataloged in SIMBAD. The majority are eclipsing binaries (275), followed by a few stars (10), QSOs (9), and a small number of cataclysmic variables and RR Lyrae stars. In this context, the QSOs are AGNs without detectable Hα emission within the S-PLUS wavelength range (or where the SDSS spectra do not cover the Hα line). This group is thus characterized by the significant presence of binary star systems and various types of variable stars.

Group 3 consists of 93 objects, 42 of which are matched in the database. According to SIMBAD, the majority are labeled as QSOs (17), along with Seyfert 1 galaxies (10) and other classifications such as AGN candidates, radio sources, and a few galaxies. However, upon inspecting the spectra of several of these QSOs, we find that the line observed within the S-PLUS *J*0660 filter corresponds to [O III] and/or Hβ, rather than Hα. Given the narrow redshift range (0.31 to 0.37), these characteristics suggest that the objects in this group are better classified as AGNs, not QSOs. This may indicate a misclassification in SIMBAD, where sources labeled as QSOs in this group likely correspond to AGNs.

Group 4 includes 325 objects, 143 of which are recorded in the database. This group has a high concentration of QSOs (78) and cataclysmic variables (25). Additionally, it features a mix of blue stars, AGNs, radio sources, and white dwarf candidates. The extragalactic objects in this group show a peak in the redshift distribution around 1.35. It is expected that CVs are located closer to the QSOs than to Galactic sources in the UMAP variable space due to their photometric characteristics, which can resemble those of QSOs in certain features, despite the spectral differences (Scaringi et al. 2013).

In summary, our application of UMAP and HDBSCAN to the Hα excess sources has effectively identified distinct groups with varying astrophysical characteristics using S-PLUS photometry. The classification successfully differentiates extragalactic sources, such as QSOs and AGNs, from galactic sources, including variable stars and binary systems. However, distinguishing Galactic cataclysmic variables from QSOs with redshifts around 1.35 remains challenging. Importantly, our results suggest that objects with (*J*0660 - *r*) color excess due to emission lines can be distinguished from those with excess caused by Hα absorption lines, mainly RR Lyrae stars.

5.2.2. Integration of S-PLUS and WISE Photometry

The second experiment incorporated the W1 and W2 filters from the WISE survey. These filters were selected because they provide the best sensitivity and reliability for detecting sources with infrared excess or thermal emission (Nakazono et al. 2021). To include these data, we crossmatched the Hα sources from the Main Survey of S-PLUS with the ALLWISE catalog (Cutri et al. 2013) using a search radius of 2 arcsec. This radius was chosen considering the broader point-spread function (PSF) of WISE compared to S-PLUS, as discussed in Nakazono et al. (2021). This process initially yielded 3173 matches, which were reduced to 1910 after applying photometric quality cuts, including errors

smaller than 0.5 magnitudes in W1 and W2 and equivalent constraints for S-PLUS filters.

Additional colors were constructed by combining WISE bands (W1 and W2) with S-PLUS broadband filters, such as W1 - W2, W1 - u, W2 - u, W1 - g, and so on. This expanded the parameter space from 66 to 77 variables, enriching the dataset and enhancing the performance of machine learning models in characterizing the physical properties of the Hα sources. **We identified optimal parameters for UMAP as `n_neighbors` = 50 and `n_components` = 2, based on the silhouette score and the Davies-Bouldin Index. For HDBSCAN, we employed `min_cluster_size` = 50 and `min_samples` = 5.**

Figure 17 shows the results of the reduction in dimensionality and the groups identified by applying UMAP followed by HDBSCAN, using the input parameters described in the previous paragraph. On this occasion, HDBSCAN found five groups and one objects that were classified as noise. Table 3 summarizes these results:

Group 0 contains 1,437 objects, 424 of which correspond to entries in the SIMBAD database. Among these, 262 are eclipsing binaries (EB*), followed by 98 RR Lyrae stars (RRLyr). Other objects include EB* candidates, stars, pulsating variables, and a few QSOs. This group predominantly consists of variable stars and a small number of extragalactic sources.

Group 1 includes 59 objects, 23 of which are matched with SIMBAD. The majority are QSOs (20), with a few other objects like a galaxy, a radio source, and a QSO candidate. This group mainly represents extragalactic sources, particularly active galactic nuclei. The redshift distribution has a peak around 2.45. This group resembles **Group 0** from the previous case (without WISE analysis), albeit with one additional QSO.

Group 2 consists of 93 objects, with 43 identified in the database. The group is primarily composed of QSOs (18), Seyfert 1 galaxies (10), and AGN candidates, with some galaxies and radio sources. This indicates a strong presence of active galactic nuclei and other extragalactic objects. The redshift distribution for extragalactic objects in this group ranges approximately from 0.31 to 0.37. This group is analogous to **Group 3** from the previous analysis (without WISE data).

Group 3 includes 51 objects, with 36 matches in SIMBAD. The majority are cataclysmic variables (24), with a few CV candidates, hot subdwarf candidates, and white dwarf candidates. This group is largely composed of cataclysmic variables and related stellar objects.

Group 4 contains 269 objects, 100 of which are matched with the database. The majority are QSOs (83), with a mix of blue stars, AGNs, radio sources, stars, and galaxies. This group shows a variety of astrophysical phenomena, both stellar and extragalactic, with a redshift distribution peaking around 1.35. It is similar to **Group 4** from the previous analysis using only S-PLUS data. In the S-PLUS-only group, the majority of objects are QSOs and cataclysmic variables, while the S-PLUS + WISE group contains more QSOs but no cataclysmic variables. The photometric characteristics of CVs in S-PLUS resemble those of QSOs, explaining why they cluster closer to QSOs than to Galactic sources in the UMAP variable space. The inclusion of WISE data likely contributed to the increase in QSOs by providing infrared information that helps differentiate extragalactic objects.

In summary, the inclusion of WISE filters in our analysis has significantly enhanced the clustering of Hα excess sources. The integration of WISE data has allowed for a more precise differentiation between galactic and extragalactic sources, enriching our understanding of the objects in our dataset. Notably, it has facilitated the separation of cataclysmic variables from QSOs with redshifts around 1.35. For detailed insights, refer to Section 4 where the redshifted emission lines of extragalactic objects are highlighted in the $J0660$ filter. However, it is important to note that the addition of WISE data has introduced challenges in identifying the group of RR Lyrae stars using HDBSCAN.

Uncertainties in photometric colors of variable stars based on single, random observations are inherently biased due to the stars' intrinsic variability. This effect is more pronounced with an increase in the amplitude of variability, as seen in classes of stars like RR Lyrae and Mira variables, which typically exhibit amplitudes higher than 0.3 to 2 magnitudes for RR Lyrae stars Chandra X-ray Observatory. The S-PLUS survey offers a significant advantage in this regard, as its 12 photometric wavebands are observed nearly simultaneously within approximately 1.5 hours SPLUS. Consequently, these observations are closely spaced in phase for variable stars.

For instance, RR Lyrae stars (RRab subtype), which have periods of approximately 0.5 days Chandra X-ray Observatory, will have all 12 S-PLUS wavebands captured within a phase range of $\simeq$0.1. This minimizes the variability effects on the observed photometric colors and ensures a more precise and reliable measurement of stellar parameters derived from these data. In contrast, random or non-simultaneous observations are likely to result in larger uncertainties due to phase mismatches, particularly for variable stars with significant amplitude changes over short timescales.

When S-PLUS wavebands are combined with external data, such as WISE wavebands, the phase mismatch becomes a critical source of uncertainty. WISE observations, which are not time-synchronized with S-PLUS, can introduce errors because the observed phases of variable stars in the combined dataset will be random. As a result, the derived colors and parameters will suffer from increased scatter and reduced precision. Therefore, we attribute the improved performance of models relying solely on S-PLUS observations to the reduced uncertainties in colors achieved by observing all wavebands in a near-simultaneous manner. This emphasizes the importance of phase-coherent photometric observations for the precise characterization of variable stars, especially those with significant amplitude variability.

5.3. Extracting Main Features: Color Analysis

In this section, we focus on the colors derived from the S-PLUS and WISE filters, which are effective in distinguishing the different groups of Hα-excess objects identified by the combined UMAP and HDBSCAN analysis of the S-PLUS data.

In the MS Hα-excess list, we identified extragalactic sources with higher redshifts, where blueward emission lines are redshifted to wavelengths near Hα, resulting in an apparent Hα excess in the $J0660$ filter. By incorporating the WISE filters to create additional colors for the unsupervised machine learning models, we achieved better separation of extragalactic sources from Galactic sources (see Sect. 5.2 for more details).

We used the classifications made by combining UMAP and HDBSCAN to create Random Forest (Breiman 2001) models and identified the most important features, specifically the colors that contribute to the separation or classification of the classes of objects. **The Random Forest algorithm is an ensemble learning method that builds multiple decision trees during the training phase. Each tree is trained on a random subset of the data and a random subset of features, which helps reduce overfitting and improves model generalization. During prediction, the results from all trees are aggregated by voting**

Table 3. Summary of clustering outcomes achieved using the UMAP and HDBSCAN unsupervised machine learning methods applied to Hα excess sources of the MS. Clustering is performed using S-PLUS and S-PLUS + WISE filter combinations for the MS. The table displays the number of objects allocated to each cluster, providing insights into the distribution of sources identified through the clustering process.

	Group	Number of Objects	Number with SIMBAD Match	Comments about SIMBAD Match
			Main Survey	
			Only S-PLUS Filters	
	Group 0	58	22	QSO (19), QSO_Candidate (1), Galaxy (1), Radio (1)
	Group 1	166	149	RRLyr (107), EB* (19), EB*_Candidate (1), PulsV* (9), PulsVdelSct (6), Star (2), QSO (2), RotV* (1), SB*_Candidate (1), BlueStraggler (1)
	Group 2	1539	323	EB* (275), EB*_Candidate (11), Star (10), QSO (9), CataclyV* (1), CV*_Candidate (3), V* (3), RotV* (1), Pec* (2), low-mass* (2), RRLyr (2), AGB* (1), PulsV* (1), PulsVdelSct (1), RSCVn (1)
	Group 3	93	42	QSO (17), Seyfert_1 (10), AGN (3), AGN_Candidate (6), Galaxy (3), Radio (2), RadioG (1)
	Group 4	325	143	QSO (78), CataclyV* (25), CV*_Candidate (6), Blue (7), Star (6), Hsd_Candidate (4), AGN (3), Radio (3), WD* (2), WD*_Candidate (3), RRLyr (2), Galaxy (2), EB* (1), Seyfert_1 (1)
Total		**2181**	**679**	
			S-PLUS + WISE Filters	
	Group 0	1437	424	EB* (262), EB*_Candidate (23), RRLyr (98), Star (13), PulsV* (8), V* (4), RotV* (3), QSO (3), PulsVdelSct (2), low-mass* (2), Pec* (2), CataclyV* (1), CV*_Candidate (1), AGB* (1), SB*_Candidate (1)
	Group 1	59	23	QSO (20), QSO_Candidate (1). Galaxy (1), Radio (1)
	Group 2	93	43	QSO (18), Seyfert_1 (10), AGN (3), AGN_Candidate (6), Galaxy (3), Radio (2), RadioG (1)
	Group 3	51	36	CataclyV* (24), CV*_Candidate (3), Hsd_Candidate (3), WD*_Candidate (3), RRLyr (1), Seyfert_1 (1), Star (1)
	Group 4	269	100	QSO (83), AGN (3), Blue (7), Radio (3), Star (2), Galaxy (2)
	Noise	1	–	–
Total		**1910**	**626**	

(for classification) or averaging (for regression), providing more stable and accurate predictions. Random forests are widely used for classification and regression tasks, known for their ability to handle complex data and offer reliable results. We implemented Random Forest algorithm, using 66 S-PLUS colors plus 11 additional colors generated with the W1 and W2 filters as input parameters, and labels generated by HDBSCAN.

The dataset used in this study exhibited a class imbalance: cluster 0 (1437 points), cluster 1 (59 points), cluster 2 (93 points), cluster 3 (51 points), and cluster 4 (269 points). To address this imbalance, we used the `class_weight='balanced'` parameter in the Random Forest algorithm. The classifier achieved an F1 Macro Average of 0.95 (±0.08) during 5-fold cross-validation. This high score, along with low variability, indicates that the model effectively handles the imbalance and consistently classifies the different clusters. **The Macro F1 score calculates the average F1 score across all classes, giving equal weight to each class regardless of its frequency. See (Sokolova & Lapalme 2009) for more details. The Random Forest algorithm and Macro F1 score were implemented using the `scikit-learn` package.** great info!!!

After performing the model, we accessed the feature importances using `feature_importances` from the Random Forest package. Figure 18 shows the top 20 feature importances and their respective scores, indicating the colors that contributed

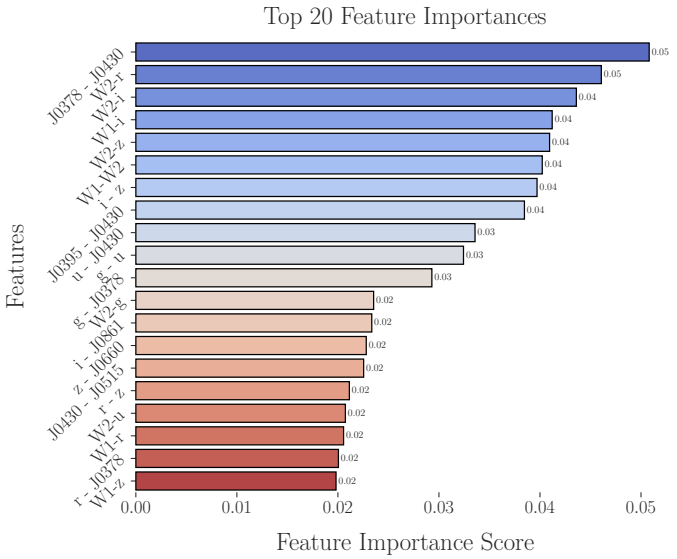

Fig. 18. Top 20 feature importances identified by the Random Forest model, showing the colors that contributed most significantly to the clustering of Hα-excess objects using UMAP + HDBSCAN. The importance scores indicate the relative impact of each color on the classification of different object classes.

most to clustering the different classes of objects identified by UMAP + HDBSCAN.

Now that we have identified the important colors, we used the `pairplot` routine in the `seaborn` package (Waskom 2021) to generate all possible color-color diagrams using the top 20 features. **Seaborn is a Python library designed to simplify the creation of statistical graphics. It extends Matplotlib and integrates seamlessly with pandas, making it particularly effective for handling and visualizing structured datasets. The `pairplot` function is especially useful for creating scatterplot matrices, allowing the simultaneous visualization of pairwise relationships between multiple features in the data.** This allowed us to identify the color-color diagrams that best separate the different classes of objects and choose the most effective ones.

Figure 19 shows nine of color-color diagrams that we selected for their ability to better separate the groups found in our Hα-excess sources list. These diagrams are based on the key features of importance, and we aimed to utilize nearly all of the 20 colors. **Tentative color cuts are presented in the figure to differentiate the various classes of Hα sources.**

This exercise demonstrates that using specific color-color diagrams with selected filters can effectively classify objects. By relying on a few key colors instead of all 12 S-PLUS filters and 2 WISE filters required for the machine learning in Section 5.2, we can reduce the number of necessary observations. This approach is advantageous because not all objects have complete photometry in all filters, and some magnitudes may not meet the clean criteria, reducing the number of objects available for classification. Consequently, using a few specific color criteria enables the classification of more objects, as it circumvents the need for complete data across all filters.

This analysis provides a practical framework for classifying Hα-excess sources without relying on complex algorithms. The proposed color-color diagrams enable the direct application of selection criteria, offering an effective method to distinguish between different classes based on the key features identified. Building on the methodology of Corradi et al. (2008) for the $r - i$ vs. r-Hα diagram, we extend it with new and effective color combinations. These criteria provide a valuable tool for identifying distinct classes of Hα-excess sources, as shown in Table 3, and can aid in targeted spectroscopic follow-ups.

6. Conclusions

In this study, we have leveraged the S-PLUS project to analyze and classify Hα-excess sources in the Southern Sky, resulting in the following key conclusions:

1. We identified 6 956 Hα-excess candidates by using the narrow *J*0660 filter in combination with the broad *r* and *i* filters from S-PLUS. This included 3637 candidates from the high-latitude MS and 3319 from the GDS.
2. Cross-referencing with the SIMBAD database enabled us to explore the types of objects in our list, identifying various emission line objects such as EM stars, YSOs, Be stars, CVs, PNe, and others. We also identified QSOs, non-local galaxies, and objects with Hα in absorption, including RR Lyrae stars, primarily within the MS. The higher detection of RR Lyrae stars (111) in the MS compared to the GDS (8), based on SIMBAD, aligns with their expected distribution in older stellar populations.
3. Validation with spectroscopic data from LAMOST and SDSS showed that approximately 60% of the spectra exhibit Hα emission lines, while around 30% show Hα in absorption in the MS. This comparison indicates the general accuracy of our classifications and supports the reliability of our Hα-excess source identifications. Furthermore, the VPHAS+ data for the GDS are consistent with our findings.
4. The S-PLUS 12-filter system facilitates the detection of RR Lyrae stars and eclipsing binaries as Hα-excess sources, capturing their short-period variability through sequential filter exposures. This makes S-PLUS uniquely suited for identifying and studying such variables, allowing for detailed analysis of their photometric behavior and potential Hα features.
5. The use of machine learning techniques, specifically UMAP for dimensionality reduction and HDBSCAN for clustering, significantly enhanced our analysis of Hα-excess sources. The 12 S-PLUS filters allowed for effective differentiation between Galactic Hα-emission objects and extragalactic sources, as well as those with Hα in absorption, such as RR Lyrae stars. However, the classification of CVs versus QSOs or AGNs, particularly with redshifts around 1.35, remained challenging. The similarity in the photometric characteristics of these objects made the boundaries between them less distinct, highlighting the inherent complexity in separating these sources, even when applying machine learning techniques with S-PLUS colors.
6. The integration of WISE filter data significantly improved the clustering process, leading to a more accurate separation between extragalactic and Galactic sources. In particular, it facilitated the differentiation between CVs and QSOs, especially for objects with redshifts around 1.35, where their photometric characteristics previously overlapped. This enhancement enabled the clear separation of CVs from QSOs, refining the classification of these sources. Additionally, the infrared data allowed the identification of specific groups corresponding to AGNs or QSOs at particular redshifts, clearly separating them from Galactic sources. Certain groups identified in the clustering process were distinctly

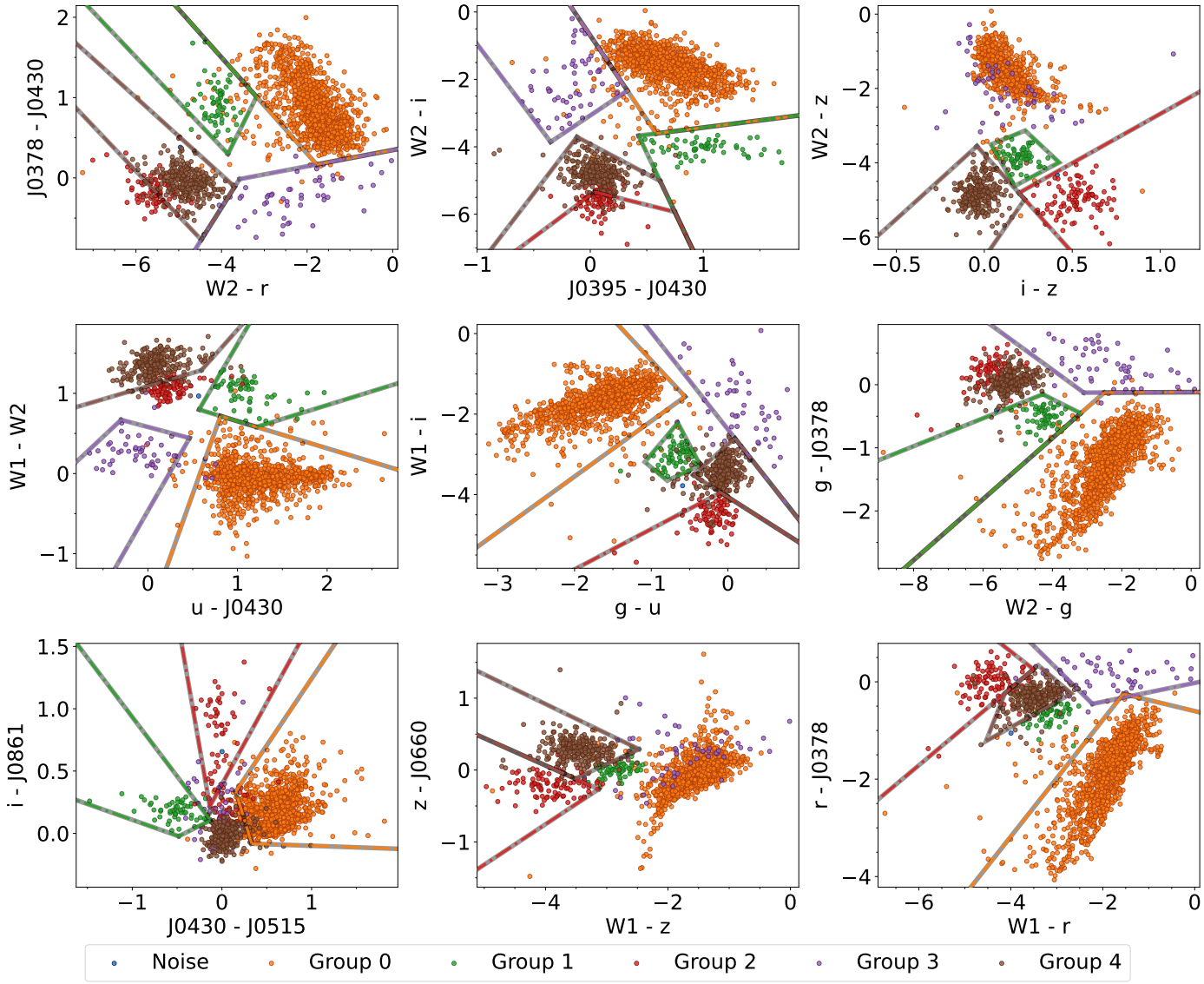

Fig. 19. Examples of color-color diagrams using the top 20 features identified by the Random Forest model. These diagrams show the separation of different classes of objects from the Hα-excess sources list. The selected diagrams illustrate effective clustering achieved through UMAP + HDBSCAN, highlighting key colors that contribute to classification. **The colored lines represent tentative color cuts for separating different classes of Hα-excess sources.**

linked to AGNs or QSOs at specific redshifts, underscoring the value of combining optical and infrared data to resolve subtle differences in photometric signatures. However, the integration of WISE data also introduced challenges in classifying RR Lyrae stars, as the combination of optical and infrared data introduced noise in the clustering algorithm, complicating their precise classification.

7. **Using data from the S-PLUS Main Survey (MS) and WISE, we constructed new, effective color-color diagrams. By applying a Random Forest model to the results of clustering with UMAP and HDBSCAN, we identified key photometric features that differentiate the various classes of Hα-excess sources. Additionally, tentative color criteria are proposed within these color-color diagrams, enabling a preliminary classification of sources without the need for complex algorithms.**

Our study used observational and analytical techniques to gain valuable insights into Hα-excess sources. Although challenges were encountered, particularly with RR Lyrae stars and certain extragalactic objects, our methods provided a robust framework for understanding Hα-excess phenomena. Our findings underscore the versatility of S-PLUS for identifying and understanding Hα-excess phenomena, demonstrating its potential for future applications in different sky regions and astrophysical environments. Future research should focus on expanding sample sizes and incorporating additional spectroscopic data to further refine classifications. Applying these methods to other sky regions or wavelengths could enhance our understanding of Hα-excess sources and their astrophysical contexts.

Acknowledgements

LAG-S acknowledges funding for this work from CONICET and FAPESP grants 2019/26412-0. RLO acknowledges financial support from the Brazilian institutions CNPq (PQ-312705/2020-4) and FAPESP (#2020/00457-4). DRG acknowledges grants

from FAPERJ (E-26/211.527/2023) and CNPq (315307/2023-4). LLN thanks Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) for granting the postdoctoral research fellowship E-40/2021(280692). PKH gratefully acknowledges the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) for the support grant 2023/14272-4. SP is supported by the international Gemini Observatory, a program of NSF NOIRLab, which is managed by the Association of Universities for Research in Astronomy (AURA) under a cooperative agreement with the U.S. National Science Foundation, on behalf of the Gemini partnership of Argentina, Brazil, Canada, Chile, the Republic of Korea, and the United States of America. This work is sponsored (in part) by the Chinese Academy of Sciences (CAS), through a grant to the CAS South America Center for Astronomy (CASSACA). We acknowledge the science research grants from the China Manned Space Project with NO. CMS-CSST-2021-A05. AAC acknowledges support from the State Agency for Research of the Spanish MCIU through the "Center of Excellence Severo Ochoa" award to the Instituto de Astrofísica de Andalucía (SEV-2017-0709). The authors would like to thank Amanda Reis Lopes for her useful suggestions and comments.

The S-PLUS project, including the T80-South robotic telescope and the S-PLUS scientific survey, was founded as a partnership between the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), the Observatório Nacional (ON), the Federal University of Sergipe (UFS), and the Federal University of Santa Catarina (UFSC), with important financial and practical contributions from other collaborating institutes in Brazil, Chile (Universidad de La Serena), and Spain (Centro de Estudios de Física del Cosmos de Aragón, CEFCA). We further acknowledge financial support from the São Paulo Research Foundation (FAPESP), the Brazilian National Research Council (CNPq), the Coordination for the Improvement of Higher Education Personnel (CAPES), the Carlos Chagas Filho Rio de Janeiro State Research Foundation (FAPERJ), and the Brazilian Innovation Agency (FINEP).

Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS Web Site is `http://www.sdss.org/`.

The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington.

Guoshoujing Telescope (the Large Sky Area Multi-Object Fiber Spectroscopic Telescope LAMOST) is a National Major Scientific Project built by the Chinese Academy of Sciences. Funding for the project has been provided by the National Development and Reform Commission. LAMOST is operated and managed by the National Astronomical Observatories, Chinese Academy of Sciences.

Scientific software and databases used in this work include TOPCAT[12] (Taylor 2005), simbad and vizier from Strasbourg Astronomical Data Center (CDS)[13] and the following python packages: `numpy`, `astropy`, `matplotlib`, `seaborn`, `pandas`, `scikit-learn`, `hdbscan`, `umap`.

References

Abril, J., Schmidtobreick, L., Ederoclite, A., & López-Sanjuan, C. 2020, MNRAS, 492, L40
Ahumada, R., Prieto, C. A., Almeida, A., et al. 2020, ApJS, 249, 3
Akras, S. 2023, MNRAS, 519, 6044
Akras, S., Gonçalves, D. R., Alvarez-Candal, A., & Pereira, C. B. 2021, MNRAS, 502, 2513
Akras, S., Guzman-Ramirez, L., & Gonçalves, D. R. 2019a, MNRAS, 488, 3238
Akras, S., Guzman-Ramirez, L., Leal-Ferreira, M. L., & Ramos-Larios, G. 2019b, ApJS, 240, 21
Akras, S., Leal-Ferreira, M. L., Guzman-Ramirez, L., & Ramos-Larios, G. 2019c, MNRAS, 483, 5077
Almeida-Fernandes, F., SamPedro, L., Herpich, F. R., et al. 2022, MNRAS, 511, 4590
Barentsen, G., Farnhill, H. J., Drew, J. E., et al. 2014, MNRAS, 444, 3230
Barentsen, G., Vink, J. S., Drew, J. E., et al. 2011, MNRAS, 415, 103
Becht, E., McInnes, L., Healy, J., et al. 2018, Nature biotechnology
Benitez, N., Dupke, R., Moles, M., et al. 2014, arXiv e-prints, arXiv:1403.5237
Bertin, E. 2011, in Astronomical Society of the Pacific Conference Series, Vol. 442, Astronomical Data Analysis Software and Systems XX, ed. I. N. Evans, A. Accomazzi, D. J. Mink, & A. H. Rots, 435
Bertin, E. & Arnouts, S. 1996, A&AS, 117, 393
Blair, W. P. & Long, K. S. 2004, ApJS, 155, 101
Bom, C. R., Cortesi, A., Lucatelli, G., et al. 2021, MNRAS, 507, 1937
Bonoli, S., Marín-Franch, A., Varela, J., et al. 2021, A&A, 653, A31
Breiman, L. 2001, Machine Learning, 45, 5
Campello, R. J. G. B., Moulavi, D., & Sander, J. 2013, in Advances in Knowledge Discovery and Data Mining, ed. J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (Berlin, Heidelberg: Springer Berlin Heidelberg), 160–172
Cenarro, A. J., Moles, M., Cristóbal-Hornillos, D., et al. 2019, A&A, 622, A176
Coelho, P. R. T. 2014, MNRAS, 440, 1027
Cook, D. O., Kasliwal, M. M., Van Sistine, A., et al. 2019, ApJ, 880, 7
Corradi, R. L. M. & Giammanco, C. 2010, A&A, 520, A99
Corradi, R. L. M., Rodríguez-Flores, E. R., Mampaso, A., et al. 2008, A&A, 480, 409
Corradi, R. L. M., Sabin, L., Munari, U., et al. 2011, A&A, 529, A56
Cutri, R. M., Wright, E. L., Conrow, T., et al. 2013, Explanatory Supplement to the AllWISE Data Release Products, Explanatory Supplement to the AllWISE Data Release Products, by R. M. Cutri et al.
Davies, D. L. & Bouldin, D. W. 1979, IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1, 224
Davies, R. D., Elliott, K. H., & Meaburn, J. 1976, MmRAS, 81, 89
Drew, J. E., Gonzalez-Solares, E., Greimel, R., et al. 2014, MNRAS, 440, 2036
Drew, J. E., Greimel, R., Irwin, M. J., et al. 2005, MNRAS, 362, 753
Drew, J. E., Greimel, R., Irwin, M. J., & Sale, S. E. 2008, MNRAS, 386, 1761
Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. 1996, in Proc. of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), 226–231
Fratta, M., Scaringi, S., Drew, J. E., et al. 2021, MNRAS, 505, 1135
Frew, D. J. 2008, PhD thesis, Department of Physics, Macquarie University, NSW 2109, Australia
Fukugita, M., Ichikawa, T., Gunn, J. E., et al. 1996, AJ, 111, 1748
González-López lira, R. A., Lomelí-Núñez, L., Álamo-Martínez, K., et al. 2017, ApJ, 835, 184
González-López lira, R. A., Lomelí-Núñez, L., Ordenes-Briceño, Y., et al. 2022, ApJ, 941, 53
Greer, P. A., Payne, S. G., Norton, A. J., et al. 2017, A&A, 607, A11
Gutiérrez-Soto, L. A., Gonçalves, D. R., Akras, S., et al. 2020, A&A, 633, A123
Gutiérrez-Soto, L. A., Mari, M. B., Weidmann, W. A., & Faifer, F. R. 2024, New A, 109, 102207
Herpich, F. R., Almeida-Fernandes, F., Oliveira Schwarz, G. B., et al. 2024, A&A, 689, A249
Jacoby, G. H., Kronberger, M., Patchick, D., et al. 2010, PASA, 27, 156
Jaiswal, S. & Omar, A. 2016, MNRAS, 462, 92
Kalari, V. M., Vink, J. S., Drew, J. E., et al. 2015, MNRAS, 453, 1026

[12] `http://www.star.bristol.ac.uk/~mbt/topcat/`
[13] `https://cds.u-strasbg.fr/`

Kron, R. G. 1980, ApJS, 43, 305

Lloyd, S. 1982, IEEE Transactions on Information Theory, 28, 129

Lomelí-Núñez, L., Mayya, Y. D., Rodríguez-Merino, L. H., Ovando, P. A., & Rosa-González, D. 2022, MNRAS, 509, 180

Marín-Franch, A., Chueca, S., Moles, M., et al. 2012, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 8450, Modern Technologies in Space- and Ground-based Telescopes and Instrumentation II, ed. R. Navarro, C. R. Cunningham, & E. Prieto, 84503S

McInnes, L., Healy, J., & Astels, S. 2017, The Journal of Open Source Software, 2

McInnes, L., Healy, J., & Melville, J. 2020, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction

Mendes de Oliveira, C., Ribeiro, T., Schoenell, W., et al. 2019, MNRAS, 489, 241

Merc, J., Gàlis, R., & Wolf, M. 2019, Eruptive Stars Information Letter, 41, 78

Merc, J., Gális, R., Wolf, M., et al. 2022, MNRAS, 510, 1404

Merc, J., Gális, R., Wolf, M., et al. 2021, MNRAS, 506, 4151

Merc, J., Mikołajewska, J., Gromadzki, M., et al. 2020, A&A, 644, A49

Mikołajewska, J., Caldwell, N., & Shara, M. M. 2014, MNRAS, 444, 586

Mikołajewska, J., Shara, M. M., Caldwell, N., Iłkiewicz, K., & Zurek, D. 2017, MNRAS, 465, 1699

Miszalski, B., Acker, A., Moffat, A. F. J., Parker, Q. A., & Udalski, A. 2009, A&A, 496, 813

Miszalski, B. & Mikołajewska, J. 2014, MNRAS, 440, 1410

Monguió, M., Greimel, R., Drew, J. E., et al. 2020, A&A, 638, A18

Munari, U., Alcalá, J. M., Frasca, A., et al. 2022, A&A, 661, A124

Munari, U., Traven, G., Masetti, N., et al. 2021, MNRAS, 505, 6121

Nakazono, L., Mendes de Oliveira, C., Hirata, N. S. T., et al. 2021, MNRAS, 507, 5847

Oke, J. B. & Gunn, J. E. 1983, ApJ, 266, 713

Parker, Q. A., Bojičić, I. S., & Frew, D. J. 2016, in Journal of Physics Conference Series, Vol. 728, Journal of Physics Conference Series, 032008

Parker, Q. A., Phillipps, S., Pierce, M. J., et al. 2005, MNRAS, 362, 689

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825

Peters, C. M., Richards, G. T., Myers, A. D., et al. 2015, ApJ, 811, 95

Pickles, A. J. 1998, PASP, 110, 863

Pollmann, E., Bennett, P. D., Vollmann, W., & Somogyi, P. 2018, Information Bulletin on Variable Stars, 6249, 1

Raddi, R., Drew, J. E., Steeghs, D., et al. 2015, MNRAS, 446, 274

Rousseeuw, P. J. 1987, Journal of Computational and Applied Mathematics, 20, 53

Sabin, L., Zijlstra, A. A., Wareing, C., et al. 2010, PASA, 27, 166

Scaringi, S., Groot, P. J., Verbeek, K., et al. 2013, MNRAS, 428, 2207

Sokolova, M. & Lapalme, G. 2009, Information Processing & Management, 45, 427

Taylor, M. B. 2005, in Astronomical Society of the Pacific Conference Series, Vol. 347, Astronomical Data Analysis Software and Systems XIV, ed. P. Shopbell, M. Britton, & R. Ebert, 29

Viironen, K., Mampaso, A., Corradi, R. L. M., et al. 2009, A&A, 502, 113

Vink, J. S., Drew, J. E., Steeghs, D., et al. 2008, MNRAS, 387, 308

Waskom, M. L. 2021, Journal of Open Source Software, 6, 3021

Wevers, T., Jonker, P. G., Nelemans, G., et al. 2017, MNRAS, 466, 163

Witham, A. R., Knigge, C., Aungwerojwit, A., et al. 2007, MNRAS, 382, 1158

Witham, A. R., Knigge, C., Drew, J. E., et al. 2008, MNRAS, 384, 1277

Witham, A. R., Knigge, C., Gänsicke, B. T., et al. 2006, MNRAS, 369, 581

Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, AJ, 140, 1868

Wu, Y., Luo, A. L., Li, H.-N., et al. 2011, Research in Astronomy and Astrophysics, 11, 924

Yang, L., Yuan, H., Xiang, M., et al. 2022, A&A, 659, A181

York, D. G., Adelman, J., Anderson, John E., J., et al. 2000, AJ, 120, 1579

[1] Instituto de Astrofísica de La Plata (CCT La Plata - CONICET - UNLP), B1900FWA, La Plata, Argentina
e-mail: gsotoangel@fcaglp.unlp.edu.ar
[2] Departamento de Astronomia, Instituto de Astronomia, Geofísica e Ciências Atmosféricas da USP, Cidade Universitária, 05508-900 São Paulo, SP, Brazil
[3] Departamento de Física, Universidade Federal de Sergipe, Av. Marechal Rondon, S/N, 49100-000, São Cristóvão, SE, Brazil
[4] Observatório Nacional, Rua Gal. José Cristino 77, 20921-400, Rio de Janeiro, RJ, Brazil
[5] Institute for Astronomy, Astrophysics, Space Applications and Remote Sensing, National Observatory of Athens, GR 15236 Penteli, Greece
[6] Universidade Federal do Rio de Janeiro, Observatório do Valongo, Ladeira do Pedro Antônio, 43, Saúde CEP 20080-090 Rio de Janeiro, RJ, Brazil
[7] Instituto de Astrofísica de Andalucía, CSIC, Apt 3004, E18080 Granada, Spain
[8] Instituto de Física Aplicada a las Ciencias y las Tecnologías, Universidad de Alicante, San Vicent del Raspeig, E03080, Alicante, Spain
[9] Instituto de Astronomía y Ciencias Planetarias, Universidad de Atacama, Copayapu 485, Copiapó, Chile
[10] Millennium Institute of Astrophysics, Nuncio Monseñor Sotero Sanz 100, Of. 104, Providencia, Santiago, Chile
[11] Instituto Multidisciplinario de Investigación y Postgrado, Universidad de La Serena, Raúl Bitrán 1305, 1700000 La Serena, Chile
[12] Departmento de Astronomía, Universidad de La Serena, Avenida Raúl Bitrán 1305, La Serena, Chile
[13] International Gemini Observatory/NSF NOIRLab, Casilla 603, La Serena, Chile
[14] Departamento de Física, Universidade Federal de Santa Catarina, Florianópolis, SC, 88040-900, Brazil
[15] NOAO, P.O. Box 26732, Tucson, AZ 85726
[16] GMTO Corporation 465 N. Halstead Street, Suite 250 Pasadena, CA 91107

Appendix A: RR Lyrae Stars in the $(r - J0660)$ versus $(r - i)$ Diagram

We cross-matched the RR Lyrae catalog from Greer et al. (2017), which contains 4 963 objects, and found 375 matches with S-PLUS data. Figure A.1 shows the distribution of these RR Lyrae stars in the $(r - J0660)$ versus $(r - i)$ diagram.

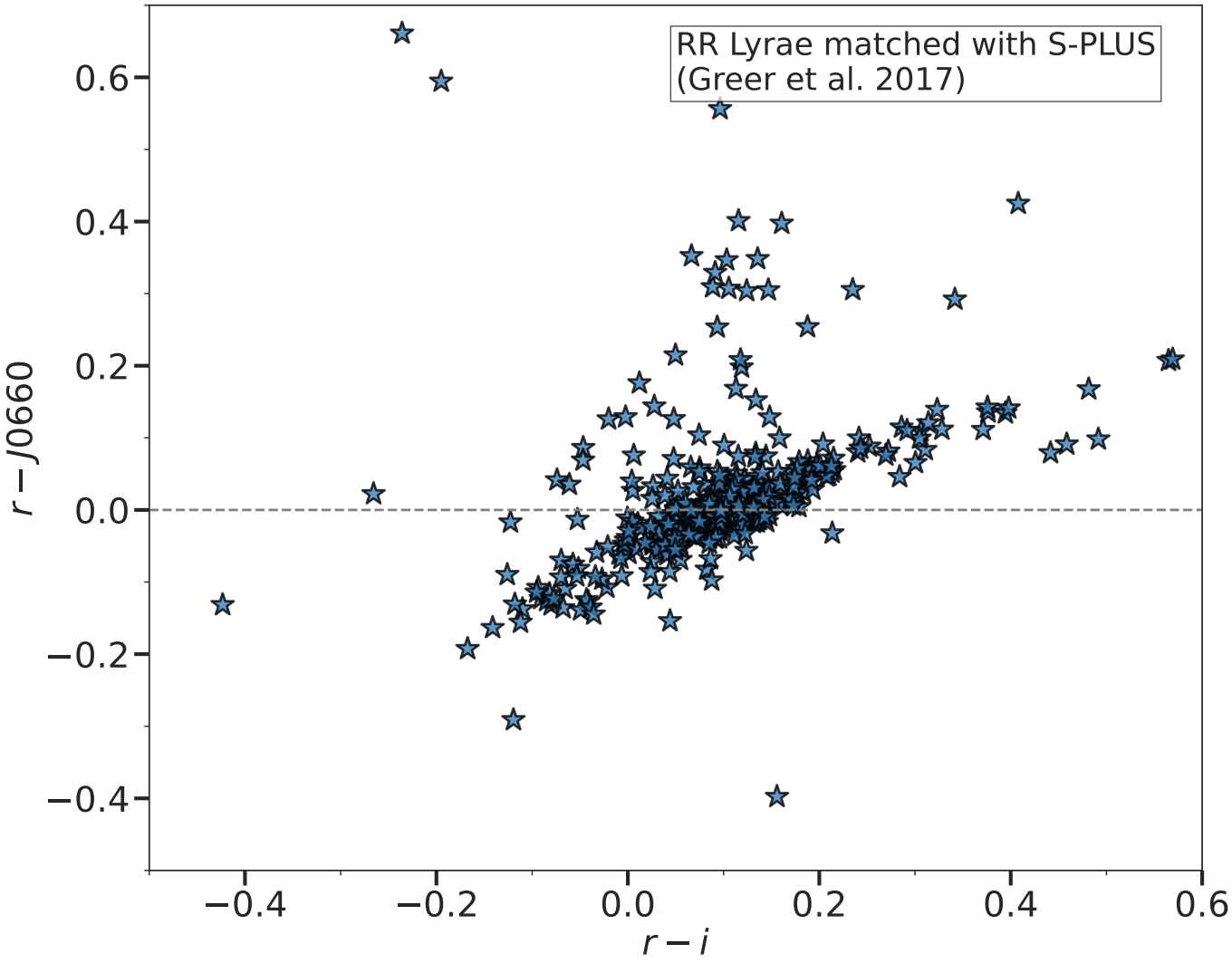

Fig. A.1. Distribution of RR Lyrae stars from the catalog of Greer et al. (2017) matched with S-PLUS data, shown in the ($r - J0660$) versus ($r - i$) diagram. The horizontal dashed line indicates the ($r - J0660$) = 0 value.