Astronomy & Astrophysics manuscript no. output
July 16, 2024

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], and et al.

[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, IAG, Universidade de São Paulo, Rua do Matão, 1226, 05509-900, São Paulo, 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

Received September 15, 1996; accepted March 16, 1997

This sentence does not fit here or
first say that Ha emission is also found in QSOs
and AGNs but it is detected in larger wavelengts due to
the redshift. Possible detection of an emission near to 6563A
corresponds to UV lines that shifted at the visible.

ABSTRACT

Context. We leverage the Southern Photometric Local Universe Survey (S-PLUS) to identify and classify Hα-excess sources in the Southern Sky. This approach combines extensive photometric data with advanced machine learning techniques to enhance source classification.
Aims. We aim to enhance the classification of Hα-excess sources by integrating multi-wavelength photometric data with advanced machine learning methods. Our goal is to accurately distinguish between extragalactic and Galactic sources and to address the challenge of incomplete photometric data.
Methods. We selected Hα-excess candidates using a $(r-J0660)$ versus $(r-i)$ color-color diagram from both the main and disk surveys of S-PLUS. Dimensionality reduction was performed using UMAP, followed by clustering with HDBSCAN. Initially, clustering was conducted using only the S-PLUS filter colors. Subsequently, we incorporated WISE data to evaluate improvements in classification performance. Finally, a Random Forest model was employed, utilizing the combined S-PLUS and WISE filter colors to identify key distinguishing features.
Results. Combining multi-wavelength photometric data with machine learning techniques has substantially enhanced the identification and classification of Hα-excess sources. We identified a total of 7,371 sources with excess in the $J0660$ filter, indicative of Hα-excess. Among these, we classified various object types, including cataclysmic variables, quasars, young stellar objects, and different types of stars and galaxies, in agreement with the SIMBAD database. Using only the 12 S-PLUS filters, UMAP and HDBSCAN effectively clustered the data, distinguishing between Hα-emission objects and those with Hα in absorption, such as RR Lyrae stars. Incorporating additional WISE filters further refined this clustering, enabling successful separation of extragalactic sources from Galactic ones and improving the differentiation between cataclysmic variables and QSOs. The Random Forest model, based on HDBSCAN results, identified key color features that effectively distinguished between the different classes of Hα-excess sources.

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

decretion? you mean accretion

1. Introduction

Atomic excitation followed by recombination in Balmer hydrogen emission lines may be ignited in different ways, thermal and nonthermal collisional excitation in shock-heated gas and energetic photons acting over a diffuse gas. As a practical result, and the Universe is 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. In particular, the Hα emission line – rest-frame wavelength of 6564.614 Å at vacuum – that corresponds to the electron transition from the $n = 3$ to the $n = 2$ energy levels, is the strongest one, in both emission or absorption, and the most widely used to identify various types of objects (e.g 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 stars (RGB), active late-type dwarfs). Amongst massive stars, emission lines are observed in Be stars with decretion disks, Wolf-Rayet (WR) stars, and interacting binary systems that are experiencing mass exchange, like symbiotic stars (SySt), cataclysmic variables (CVs), among others. In the case of high redshifted sources like starburst galaxies and quasi-stellar objects (QSOs), the detection of emission at 6563 Å is not associated with the recombination of Hα but with other UV emission lines.

Most of the aforementioned classes of objects are not homogeneous and far from complete even in the local universe, with some being highly populated while others being highly underrepresented. For example, there are ∼320 known SySts, with only ∼65 of those located in galaxies other than the Milky Way (Akras et al. 2019b; Merc et al. 2019). 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 carried out in the past in a variety of angular resolutions, sky coverage, and sensitivity. Some of

are you sure about that?

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

320 in the Milky way and only 65 in nearby galaxies.

there are also some ref from Miszalskit et al. 2009A&A...496..813M , 2014MNRAS.440.1410M
Mikołajewska et al. 2014MNRAS.444..586M, 2017MNRAS.465.1699M
what do you mean?

them, with modest spatial resolutions, have revealed themselves spatially resolved, extended nebular emission to study supernova remnants, galaxy groups, and star-forming regions (e.g. Davies et al. 1976). Others, with higher spatial resolution, disclosed 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; Akras et al. 2019c), early-type emission-line stars (Drew et al. 2008), and PNe (Viironen et al. 2009; Sabin et al. 2010; Akras et al. 2019a).

In general terms, Witham et al. (2008) developed a method to select Hα emission line sources in the IPHAS survey by implementing the aforementioned color-color diagram (r - Hα) versus ($r - i$). Hα excess line objects are identified by iteratively fitting the stellar locus and considering those objects as candidates that fall several sigma above this stellar locus in the r - Hα color. This conservative method leaves a total of 4 853 point sources 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) created the INT Galactic Plane Survey (IGAPS), which is a merger between the optical surveys, IPHAS and UVEX ((the UV-Excess survey of the Northern Galactic Plane; Groot et al. 2009). This catalog have a total of 295.4 million rows providing photometry in the filters, i, r, narrow-band Hα, g, and U_{RGO}. In this catalog, they find 8 292 candidate emission line stars and over 53 000 variables (both at $> 5\sigma$ confidence). The astrometry for all five photometric bands has been aligned with the Gaia DR2 reference frame. The g, r, and i magnitudes in IGAPS were calibrated using the 'Pan-STARRS photometric reference ladder' (Magnier et al. 2013), while Hα narrow-band calibration was based on methods by Glazebrook et al. (1994).

try to make these simple write only the basic info

More recently, Fratta et al. (2021) developed a new technique that leverages astrometric and photometric information from Gaia to select Hα-bright outliers in the IPHAS catalog, across the color-absolute magnitude diagram. To mitigate contamination due to selection biases for different stellar populations and extinction, the sources in the catalog were first partitioned based on their positions in both Gaia's color-absolute magnitude space and Galactic coordinate space, respectively. Subsequently, they applied the strategy by Witham et al. (2008) to the catalogs based on these partitions. The input catalog used in this work to identify Hα-excess sources is that of Scaringi et al. (2023), the Gaia/IPHAS catalog, which results from a positional sub-arcsecond cross-match between the sources in Gaia and IPHAS DR2 fields of view. *

is that necessary

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 one 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.

full name and refs

Here we present a census of Hα-excess point-like sources from the S-PLUS DR4, utilizing the (r - $J0660$) versus (r - i) color-color diagram. We leverage the S-PLUS DR4 dataset and employ advanced machine learning techniques to enhance the identification and classification of these sources. Specifically, we use UMAP for dimensionality reduction followed by HDBSCAN 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 WISE data and apply a Random Forest model to refine our classification and identify key features that distinguish different types of Hα-excess sources.

ref

Section 2 describes the observations related to the S-PLUS project, including important information on the fourth data release, photometry, and data. Section 3 presents the technique implemented to select the Hα-feature sources and includes the analysis of the results. In Section 4, we present the machine learning methods used to analyze and make a more accurate classification of the Hα sources. Finally, Section 5 discusses our main results and conclusions.

Is that info necessary?

2. Data and Observations

2.1. S-PLUS Survey Overview

This manuscript uses data from S-PLUS DR4 (Herpich et al., submitted to A&A). DR4 encompasses 171 fields at very low galactic latitudes ($|b| < 15°$), an additional 341 fields carried over from DR3 spanning the Main Survey footprint (with $|b| > 30°$), and 150 fields within the Magellanic Clouds 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. S-PLUS is conducted using a dedicated 0.83 m robotic telescope located at Cerro Tololo, Chile (Mendes de Oliveira et al. 2019).

SPLUS

The project surveys the southern sky using the 12 filters from the Javalambre filter system (Marín-Franch et al. 2012), spanning the wavelength range from 3 000Å to 10 000Å. This system comprises seven narrow-band filters ($J0378$, $J0395$, $J0410$, $J0430$, $J0515$, $J0660$, and five broad-band Sloan-like (Fukugita et al. 1996) filters (see Fig. 1). The narrow-band $J0660$ filter used in S-PLUS is centered at λ 6614Å and has a width of approximately 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 approximately 0.02.

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

you can also mention the spectroscopic survey of LAMOST, what about SDSS?

this figure can be larger

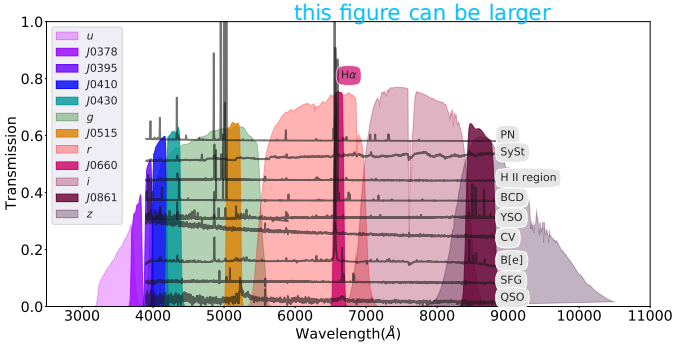


Fig. 1. Transmission curves of the S-PLUS filter set. The narrowband filter *J*0660 includes the Hα emission line. Over-plotted is spectra of different classes of emission line objects. From top to bottom: a PN, a symbiotic star, an extragalactic H II region, a blue compact/H II galaxy, a YSO, a CV star,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

2.2. Observational Data

The first goal of this work is to identify objects with Hα excess in S-PLUS DR4. For this, we applied an iterative and automatic technique to select objects with an excess in the *J*0660 band, which is consistent with the detection of the Hα line in emission. Next, the sample of Hα sources is divided into two subgroups: blue and red one. This classification was made by employing optical colors in combination with unsupervised machine learning/statistical tools. These procedures are described in the following subsections.

2.3. Main survey

what do you mean corrected add a footnote and explain it

Amongst the different aperture photometry available in the catalog, the `PStotal` photometry is used, which is a 3-arcsec aperture corrected magnitudes (Almeida-Fernandes et al. 2022). To acquire data with high-quality photometry and identify compact objects in the main survey, we initially apply several criteria: were applied

- Objects must possess an *r* magnitude within the range of $13 < r \leq 19.5$.
- For the filters, *J*0660 magnitude < 19.4 and *r* magnitude < 19.2 (see Table 4 Almeida-Fernandes et al. 2022).
- Objects must exhibit errors less than 0.2 in the *r*, *J*0660, and *i* filters.
- The signal-to-noise ratio (S/N) in their respective filter should be higher than 10.
- Objects should have `SEX_FLAGS_DET` < 4.
- Objects must satisfy `CLASS_STAR_`*r* > 0.5 and `CLASS_STAR_`*i* > 0.5.

you have to explain what are these flags

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 or irregular shapes and are therefore excluded from consideration.
- 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.

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

from total XXXXXX sources

These constraints led to the selection of 6,655,139 stars. The data were obtained by querying the project's database using the `splusdata` Python package, accessible via `S-PLUS Cloud`[3].

2.4. Disk

We created a photometric point spread function (PSF) specifically optimized for point sources within the disk... **LUIS LOMETÍ VA COLOCAR LA DESCRIPSÓN DE LA PSF FOTOMETRI AQUI!**

The same constraints as those explained in Section 2.4 were implemented for the disk to ensure high-quality data.

3. Identification of Hα Feature Sources

3.1. Photometry

SPLUS DR4

i have no idea what median-filter version is

We used a combination of SExtractor[4] (Bertin & Arnouts 1996) and PSFex[5] (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-López Lira et al. 2017). All median images were produced with a 11×11 pix^2 median filter.

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 similar to: $12 \lesssim$ MAG_AUTO $\lesssim 21.5$ and $1 \lesssim$ FLUX_RADIUS $\lesssim 3.5$. Since we are observing towards the Galactic disk, the number of sources for creation of each PSF can reach

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

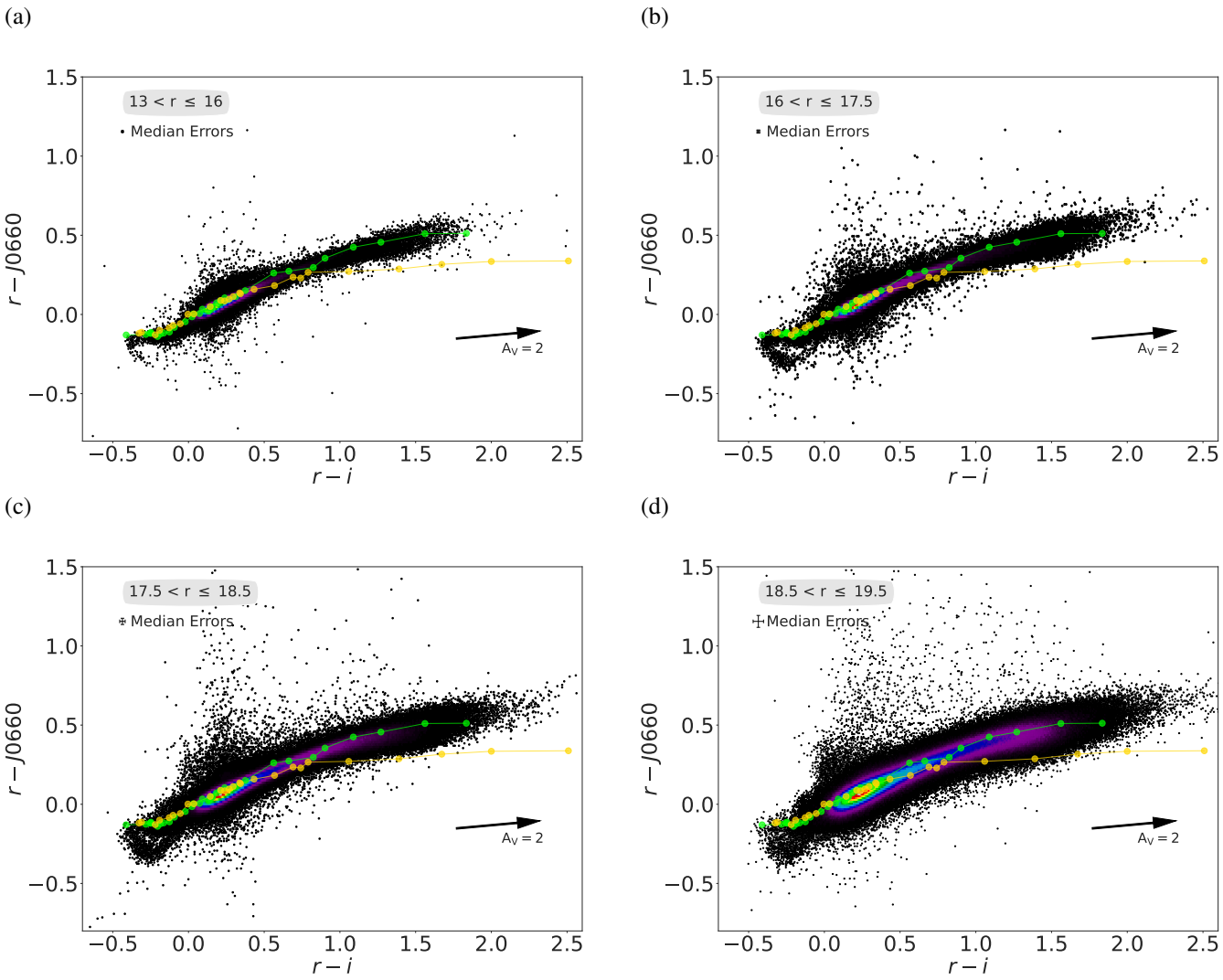


Fig. 2. The $r - J0660$ versus $r - i$ color-color plots used to select objects with Hα excess. These plots display data for all stars from the S-PLUS DR4 main survey, representing the `PStotal` photometry in these colors. The data are divided into four magnitude bins: *(a)* $13 < r \leq 16$, *(b)* $16 < r \leq 17.5$, *(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).

do you mean studied in SPLUS or somthing else

~20000 sources, which was not possible in the previous works since they were focused on extragalactic sources fara way from the Galactic disk. b) *PSF creation:* we used PSFex for PSF creation using the point sources selected in the last step. The spatial variations of the PSF were modeled with polynomials of a degree of 3. 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, by simplicity only the name of each band is written.

using the name of each band for simplicity (e.g. ?????)

To verify the plausibility of our magnitude measurements, we compared with (Aqui hay que poner los catalogos con los que comparaste las magnitudes. Y si quieres tambien poner una grafica. Veo que de esto hay algo en la figura 5).

3.2. Selection of Hα Excess Sources

Before searching for potential sources of Hα excess 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) $18.5 \leq r < 19.5$. In 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 entire sample from the main survey and the disk sources used in this study, respectively. They are presented in the color-color diagram (r - $J0660$) versus ($r - i$) across the magnitude bins. 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 main survey correspond to `PStotal`, while for the disk sources they correspond to PSF photometry.

Figures 2 and 3 display the r-J660 vs r-i color-color diagrams for the sources from the main survey of SPLUS and the sub-survey of the galactic disk, respectively.

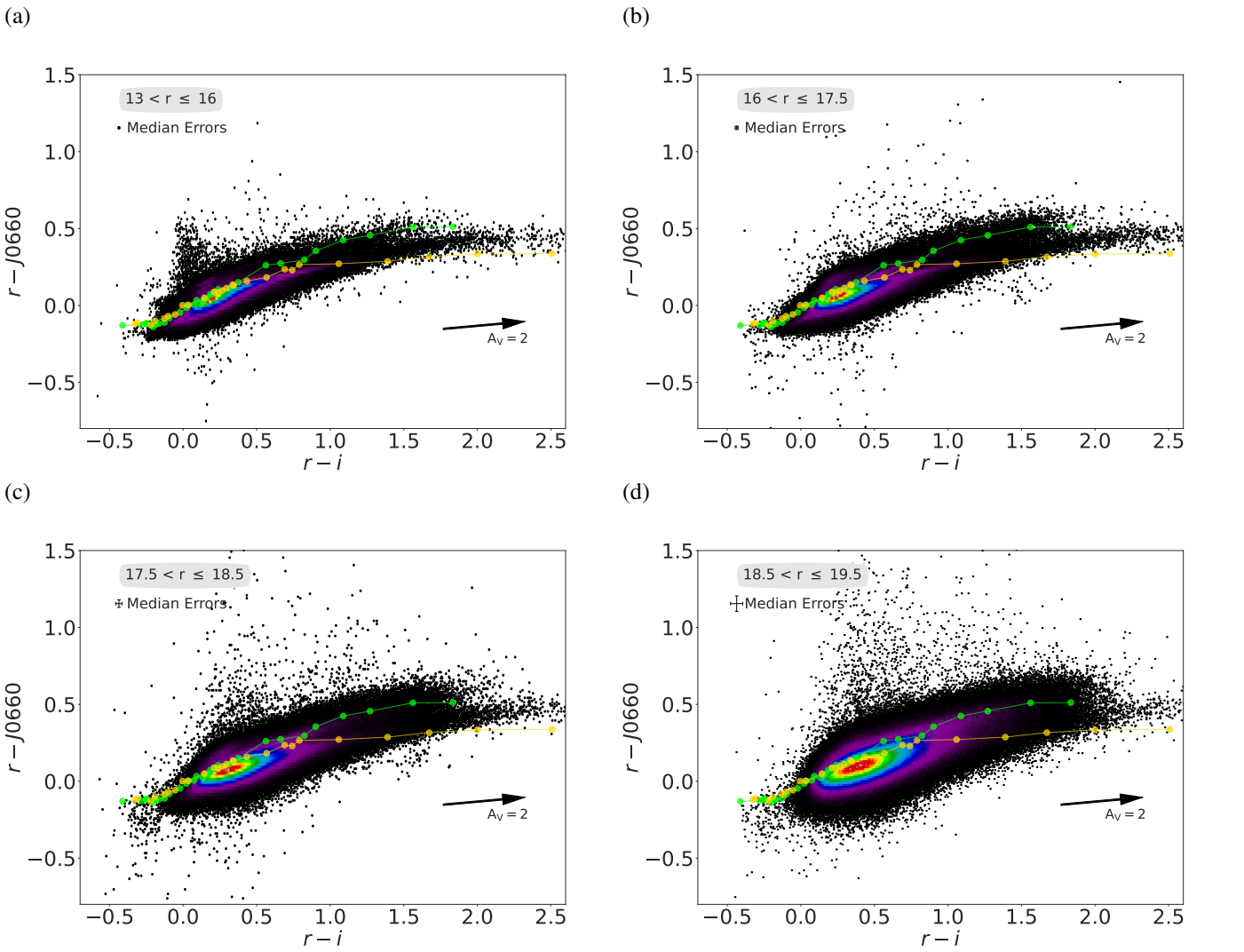


Fig. 3. Same as Fig. 2, but for the disk region and using PSF photometry.

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*, *J*0660, 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* - *J*0660) versus (*r* - *i*) diagrams for each subset of data and attempted to fit the regions predominantly occupied by main-sequence and giant stars with a linear regression model. Subsequently, we applied an iterative σ-clipping technique to the data. This method ensures that objects exhibiting excess in Hα emission should adhere to the specified criterion:

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

$(r - J0660)_{\rm obs}$ denotes the observed color difference between the *r* and *J*0660 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:

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

where $\sigma_{\rm s}$ 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 *J*0660 bands, and *m* represents the slope of the linear regression fit. The fits were performed using the `astropy.modeling` library [6].

Figure 4 provides an illustration of the procedure applied for one field corresponding to the main survey (STRIPE82-0142). The iterative approach was applied for each individual field, with solid red lines indicating the initial fit and dashed lines showing the 4-σ clipping fit. Sources with an excess in the *J*0660 filter or outliers from the stellar locus, defined as those deviating more than 5σ from these fitted lines were considered. The selection of these sources consisted of applying Eq.1 to the preselected data, for which we estimated σ using Eq.2. The large orange star in panel c of Figure 4 represents a known Hα emitter (CV) located significantly above the stellar locus, with $(r - J0660) > 0.5$.

add the name of the CV and the reference

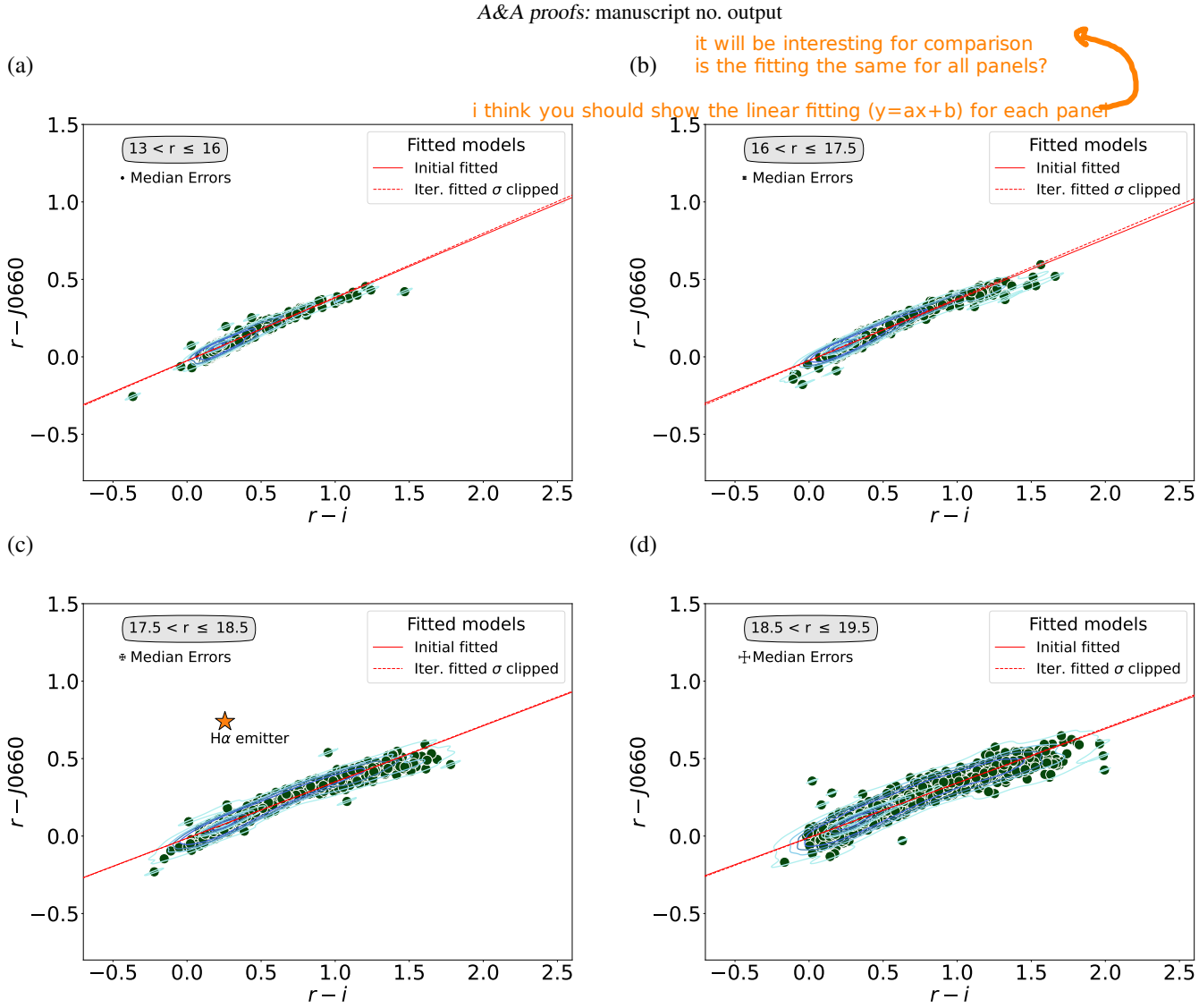


Fig. 4. An illustration of the selection criteria used to identify strong emission-line objects via colour-colour plots. The data shown here are all from the S-PLUS field STRIPE82-0142. The data are split into four magnitude bins, as shown in the four panels. Objects with Hα excess should be located near the top of the color-color diagrams. The thin red continuous lines illustrate the original linear fit to all data (green points). The dashed red lines represent the final fits the stellar locus of points which were obtained by applying an iterative σ-clipping technique to the initial fit. Objects selected as Hα emitters must be located above the dashed line. The big orange star in the c plot is CV FASTT 1560 and S-PLUS ID DR4_3_STRIPE82-0142_0021237.

3.3. Results and Analysis

Our strategy focuses on identifying 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 main survey and 3 734 for the disk. The distribution of sources with excess Hα emission in the $(r - J0600)$ versus $(r - i)$ color-color plane is depicted in Fig. 5. Square-shaded orange symbols represent objects with Hα excess identified in the main survey, while green circle symbols denote those found in the disk. All sources situated above the locus of the main and giant stars exhibit an excess in the $J0600$ filter, attributed to the Hα line. The broad distribution of sources on the color-color diagram of $(r - J0660)$ and $(r - i)$ indicates the selection of several types of Hα emitters. These sources are likely associated with PNe, CVs, SySt, YSOs, Be stars, and extragalactic compact objects like QSOs and galaxies, among others (see Fig. 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 objects' list with the SIMBAD database[7]. 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, we assumed the angular distance on the sky-plane between sources considering positive matches those of mutually closest sources to each other 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.

3.3.1. Matches with SIMBAD sources

We identified a total of 1 311 positive matches between our catalogs of Hα compact excess sources and the SIMBAD database,

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

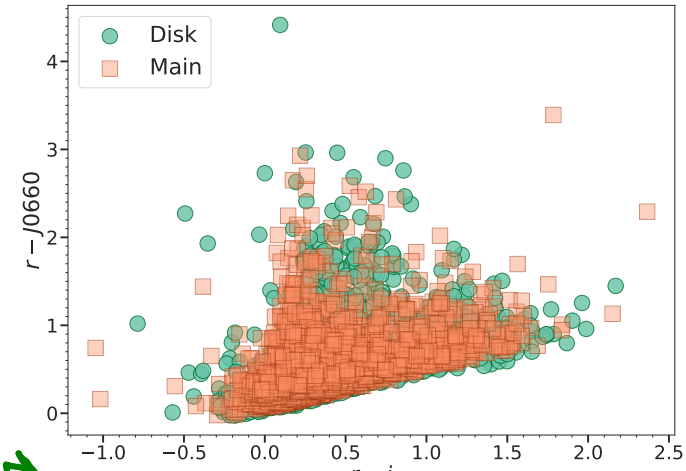
[7] http://simbad.u-strasbg.fr/simbad/

from my point of view, your selection criteria have excluded most of the extended sources. you have only 23 AGN kai 9 galaxies. I think you have to mention that. Moreover, you have only 143 QSOs.
I think you have to discuss in th text the precentage of these classes becuase your results are really nice!

the Ha

Table 2. Summary of the positional cross-match results between the S-PLUS list of emission line objects and the SIMBAD database. A search radius of 2 arcsec was used for the main survey, while 1 arcsec was used for the disk. 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*	156
Young stellar object	YSO, YSO_Candidate, Orion_V*, TTau*_Candidate, Ae*_Candidate	149
Stellar Binary System	CataclyV*, CV*_Candidate, RSCVn, EB*, EB*_Candidate, SB*	86
Variable star	PulsV*delSct, PulsV*, LPV*, LP*_Candidate, Mira, RRLyr, V*, V*?_Candidate, BYDra	56
Star	Star, **, RGB*, C*, WD*_Candidate	120
Nebula	PN?_Candidate, RfNeb, DkNeb, Nova	4
Other	EmObj, Hsd_Candidate, deltaCep, Cepheid_Candidate, Transient, X	9
Total		**580**



i suggest to use the term "Galactic disk" everywhere in the text

Fig. 5. The color-color diagram illustrates the distribution of Hα emitters in the $(r - i)$ vs $(r - J0660)$ color space. The data consists of two populations: "Disk" and "Main" representing distinct galactic components. The "Disk" population, depicted by filled circles (sky blue), corresponds to Hα emitters associated with disk structures. Meanwhile, the "Main" population, represented by open circles (salmon), signifies Hα emitters primarily located in the main body of galaxies. The markers' sizes reflect the relative abundance of emitters in each population. The plot enables the visual discrimination of these populations based on their color characteristics, aiding in the analysis of galactic structures and star-forming regions.

assuming

utilizing a search radius of $d_{\mathrm{max,proj}} = 2$ arcsec for the main survey and 1 arcsec for the disk. In the main survey, the identified objects primarily fall into categories such as variable stars, predominantly cataclysmic variables and/or candidates (CataclyV*), eclipsing binaries and/or candidates (EB*), 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 that coincide with the $J0660$ filter, simulating the Hα emission line, are also present, encompassing AGN, Seyfert galaxies, QSOs, and various types of other objects (see Table 2 for details).

For the disk, 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* category, which is expected. The second highest number belongs to the category of young stellar objects, reflecting the active star formation processes present in the Galactic disk.

An important consideration regarding the SIMBAD matches is that in the main survey, numerous extragalactic sources with emission lines are selected due to the mapping of high latitudes in the southern sky. Conversely, for the disk, no extragalactic sources have been selected. While the main survey 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 eclipsing binaries (EB*) and RR Lyrae variables (RRLyr), among others, are also present. The results are described below and listed in Table 2.

what do you mean "we include"

In our classification of Hα-excess sources, we include eclipsing binaries and variable stars such as RR Lyrae stars, due to their ability to exhibit significant Hα features. It is important to note that RR Lyrae stars, known for their distinctive spectral features, can exhibit Hα absorption lines. Consequently, they are sometimes identified as outliers in our analysis. This inclusion is a natural outcome of our selection criterion, which detects any significant deviation from the expected stellar colors, encompassing both emission and absorption features. Eclipsing binaries often show Hα emission due to complex interactions between their components and surrounding material, as demonstrated in studies of systems like eclipsing binary VV Cephei, where periodic variations in Hα emission have been observed throughout eclipse phases (Pollmann et al. 2018).

I can not undestand why do you need this paragraph about RR lyrae? it is not clear to me its importance!!!

A more detailed spectroscopic characteization of these sources is out the scope of this paper.

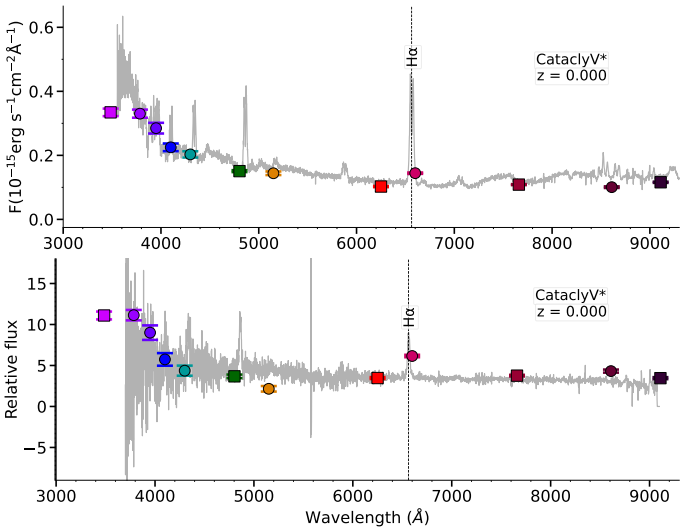


Fig. 6. SDSS (*upper*) and LAMOST (*bottom*) spectra of two objects that were selected as Hα excess sources by the approach implemented in this work. Coloured symbols represent the S-PLUS photometry in flux units; from left to right they are: *u*, *J*0378, *J*0395, *J*0410, *J*0430, *g*, *J*0515, *r*, *J*0660, *i*, *J*0861, and *z*. The SDSS and LAMOST IDs of both objects correspond to Xxxxx and xxxxx, respectively, and based on SIMBAD, both are cataclysmic variables with main IDs xxxxx and xxxxx. The dashed line indicates the Hα wavelength.

implementing our methodology in this work

do you need a new figure for that you can added in fig6, no?

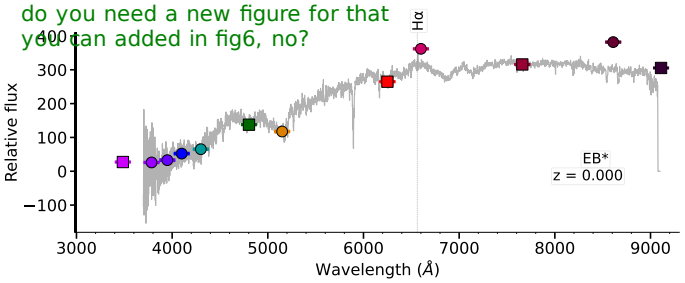


Fig. 7. The LAMOST spectrum and S-PLUS photometry of an eclipsing binary star show a weak Hα emission line.

it is not well explained.

Redshifted Lines Mimicking the Hα Emission. This particular population of apparent Hα emitters includes AGNs, Seyfert 1 galaxies, and other emission-line galaxies. Specifically, within the redshift range $0.306 < z < 0.376$—highlighted by the filled area in the figure—lines such as Hβ and [O III] 4959, 5007 Å are redshifted into the *J*0660 filter.

??

According to the classification in the literature, approximately 13% of the Hα excess sources in our sample are identified as QSOs. It is important to note that the excess observed in the *J*0660 filter for QSOs is due to redshifted emission lines that fall within the wavelength range of this filter, depending on the redshift of the QSOs. 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 Nakazono et al. 2021).

3.3.2. Spectral Analysis

We cross-matched our sample of Hα excess sources identified in the main S-PLUS survey with SDSS DR18 (Ahumada et al. 2020) and LAMOST DR9, using a 2 arcsecond cross-matching radius, identifying 212 common sources (138 from SDSS and 74 from LAMOST). Consequently, some transient Hα-excess sources identified by our algorithm may not exhibit clear Hα emission, and vice versa. Upon spectroscopic examination, approximately 60% of these sources showed emission lines, which could include redshifted lines other than Hα, while about 30% exhibited Hα absorption.

clearly

what do you mean 'transient"? it is not clear.

I understand now that you have also found the Ha line in absorption! but you have not discussed where these sources are in the color-clour diagrams !!! This part is not well presented.

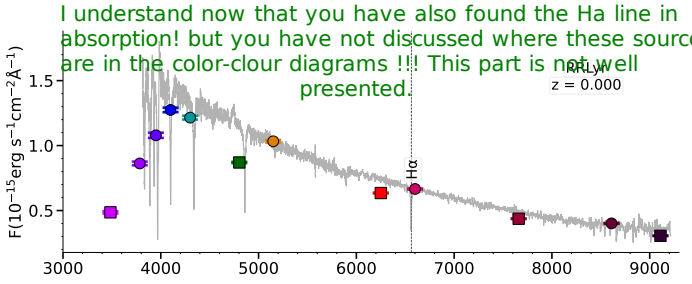


Fig. 8. SDSS spectrum and S-PLUS photometry of an RR Lyrae star show an Hα absorption line.

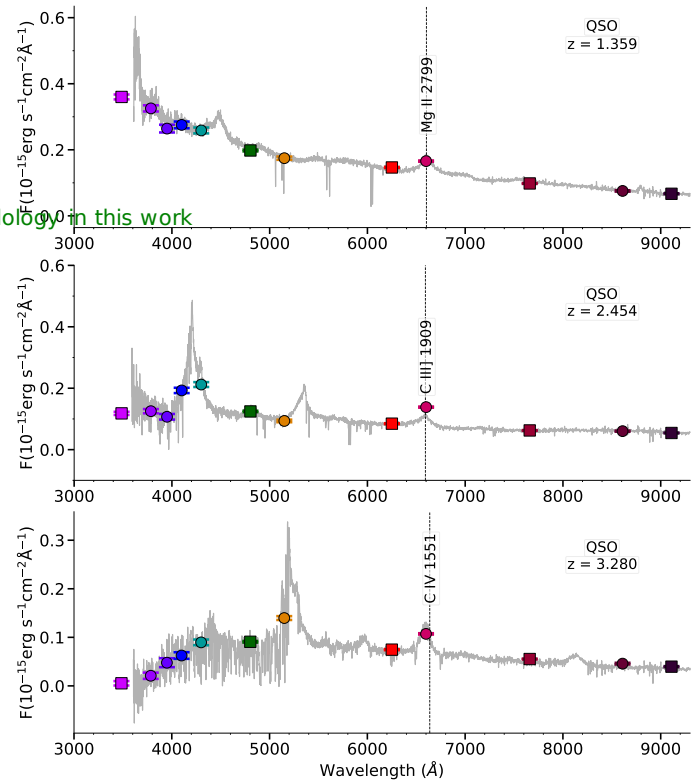


Fig. 9. S-PLUS photometry and SDSS spectra of three QSOs with redshifts of 1.359, 2.454 and 3.280 (top to bottom) selected as Hα excess sources. At these redshifts, the emission lines Mg II λ2799, C III] λ1909 and C IV λ1551 are detected in the *J*0660 filter. The SDSS IDs of the sources are: xxxx, xxxx,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

what?

Most of the objects with available spectroscopic information in SDSS and LAMOST correspond to CVs, Seyfert 1 and other AGN, and QSOs. However, we emphasize that a more detailed analysis is necessary to check which other types of objects are included in these samples of spectra – which is not in the scope of this paper. Also, it is worth noticing that part of the objects does not have a conclusive classification.

there is a number of objects without a

Figure 6 presents the SDSS (upper) and LAMOST (lower) spectra, along with the corresponding S-PLUS photometry (colored symbols) for two known cataclysmic variables (CVs). The excess in the *J*0660 filter is evidently produced by the Hα line. Note that the bluer emission tends to be more intense, which is consistent with the expected behavior of CVs showing strong Balmer series emission. Figure 7 displays the LAMOST spec-

Our list of Ha sources identified in the main SPLUS survey was cross-matched with the DR18 SDSS catalog (ref) and the DS9 Lamost catalog (ref), considering a 2 arcsec radius. We concluded to the identification of 212 commons sources (138……).

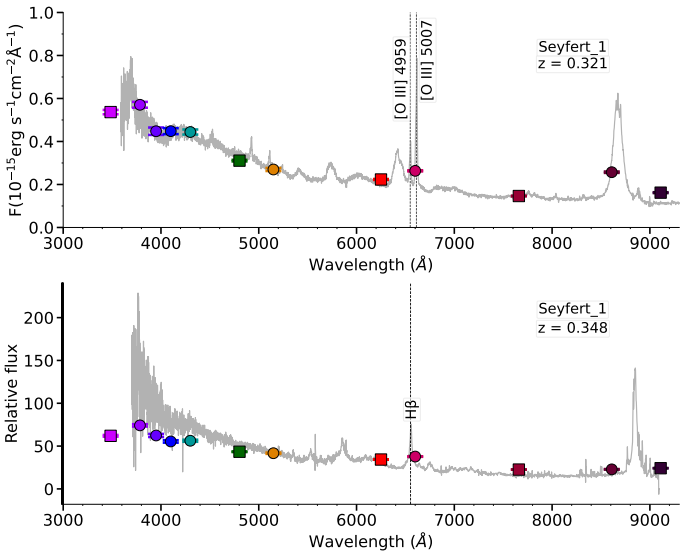


Fig. 10. As Figure 9 but for sources with redshifts of $z = 0.321$ and $z = 0.348$. At these redshifts, the lines [O III] $\lambda\lambda 4959, 5007$ doublet and Hβ line are detected in the narrow band filter, generating an Hα excess. The spectra of the sources are from SDSS (upper) and LAMOST (bottom)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

trum 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 r-band includes the Hα wavelength within its range, so the absorption feature reduces the overall flux detected, making the r band appear fainter. In contrast, the $J0660$ filter, being narrowly focused on the Hα line, 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).

Figure 9 shows examples of SDSS spectra for three objects classified as QSOs. In these cases, other bluer lines resemble the Hα emission line due to redshift, producing an excess in the $J0600$ filter. The upper panel of the figure shows a QSO with a redshift of approximately 1.36. At this redshift, the Mg II 2798 Å line is detected in the $J0600$ filter, as is perceptible in the figure. The middle panel displays the spectrum of a QSO with a redshift around 2.45, meaning that the excess in the narrow-band filter is produced by the C III] 1909 Å line. The bottom panel shows the spectrum of an object with an estimated redshift of around 3.28. In this case, the excess is generated because the $J0660$ filter is detecting the C IV 1550 Å emission line. The plots effectively show that the detected emission in the $J0600$ filter corresponds to the aforementioned redshifted emission lines.

Other extragalactic objects for which we found spectra in SDSS and LAMOST are Seyfert galaxies. Then, Figure 10 exhibits the spectra of two Seyfert 1 galaxies, one with $z \simeq 0.35$ (upper), indicating that the Hβ emission line falls within our narrow band filter. For the other one, the redshift is around $z \simeq 0.32$ (bottom). In this last case, the [O III] 4959, 5007 Å doublet

emission lines are detected in the $J0660$ filter, generating the observed excess.

The analysis of individual spectra reveals distinct features indicative of Hα lines, including both prominent emission and absorption at the expected wavelengths. These spectral properties provide valuable insights into the physical characteristics and evolutionary stages of the objects. We note that the spectral confirmation rates presented constitute a lower limit for the purity of our selection. This is because our algorithm targets Hα-excess sources rather than exclusively identifying Hα emitters. As a result, objects that exhibit an excess of Hα flux, even if they do not display a prominent Hα emission line, are selected as outliers. This approach allows us to identify a broader range of objects with Hα excess, including those with weak or subtle Hα features.

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

We performed a comparative analysis of PSF photometric colors between the S-PLUS disk data and those provided by VPHAS+[8]. By cross-matching our Hα excess sources selected in the disk with VPHAS+ DR2 using a search radius of 1 arcsecond, we obtained 998 matches. We computed the differences in two key color indices: $r - i$ and $r - H\alpha$. Specifically, we investigated the mean difference and standard deviation of these color differences to assess the consistency and agreement between the two surveys. It is noteworthy that VPHAS+, like S-PLUS, uses the r, i, and a narrowband filter designed to detect the Hα line, facilitating a meaningful comparison of Hα emission.

The comparison of photometric colors reveals important insights into the consistency and reliability of S-PLUS photometry (see Fig. 11). The mean difference in the $r - i$ color between S-PLUS and VPHAS+ was -0.24, with a standard deviation of 0.19. For the $r - H\alpha$ color, the mean difference was -0.01 with a standard deviation of 0.43. 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.

The observed consistency between the S-PLUS and VPHAS+ surveys supports the validity of the S-PLUS photometry and the precision of the Hα excess selection process. Despite the observed systematic differences, the standard deviations suggest that the photometric measurements from both surveys exhibit relatively good agreement. This consistency is crucial for 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 between the two surveys. Further investigations are warranted to better understand the contributions of these factors to the observed discrepancies.

3.3.4. 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 main survey. The normalized density facilitates comparison between different subsets. The blue curve represents Hα excess objects, while the red curve depicts all main stars. The magnitude distribution for Hα excess sources shows a higher concentration at in-

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

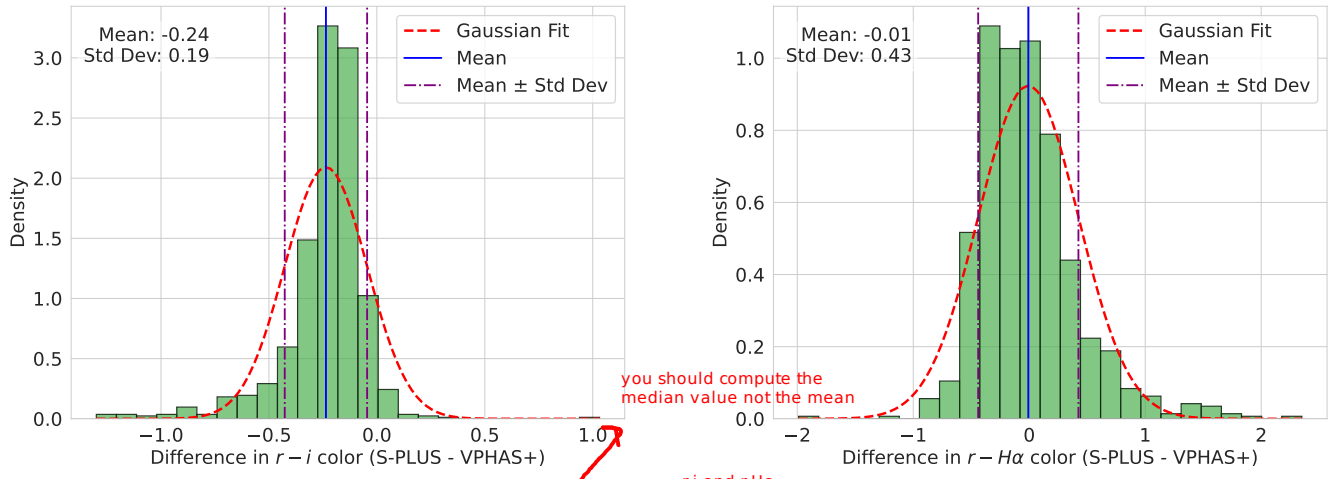


Fig. 11. Histograms illustrating the discrepancies in photometric colors between the S-PLUS and VPHAS+ surveys. The left panel depicts the differences in the $r - i$ color, while the right panel shows the differences in the $r - H\alpha$ color. The distributions represent variations observed in common stellar objects detected by both surveys. The mean difference and standard deviation for each color discrepancy are provided, offering insights into the agreement between the two datasets. These findings highlight both the consistency and systematic differences in photometric colors, which are crucial for evaluating the reliability and comparability of these surveys in characterizing stellar populations and astrophysical phenomena.

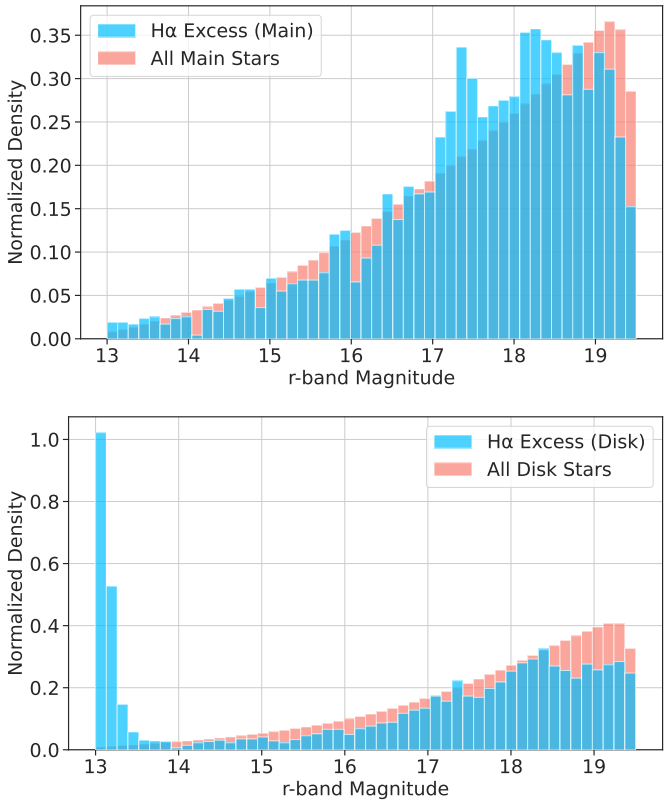


Fig. 12. *Upper panel*: Distribution of r-band magnitudes for Hα excess sources (blue curve) compared to all main stars (red curve) in the main survey. *Lower panel*: Distribution of r-band magnitudes for Hα excess sources (blue curve) in the disk region compared to all disk stars (red curve). The histogram heights represent density normalization scales.

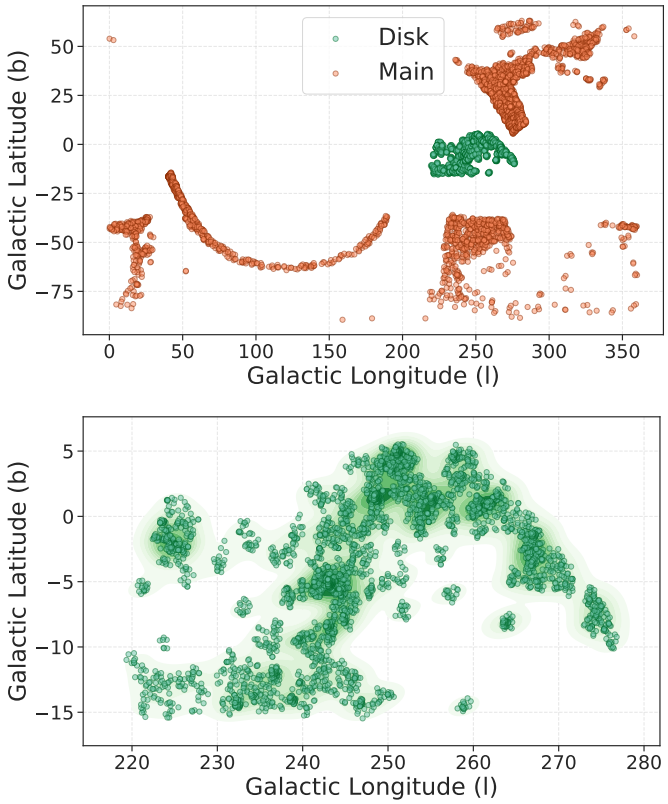
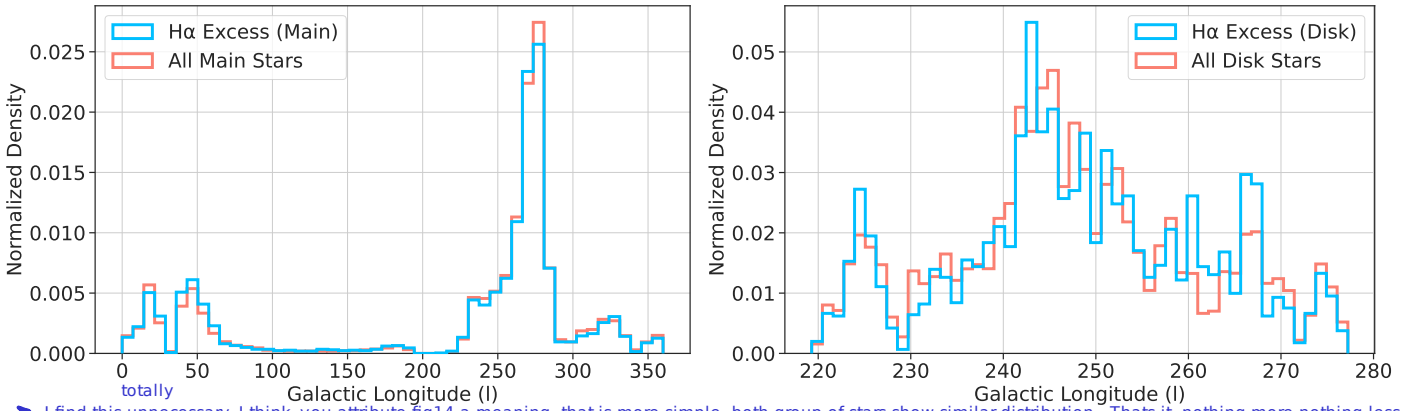


Fig. 13. Distribution of emission-line objects in Galactic longitude and latitude coordinates. The *upper panel* shows all the Hα sources selected, and the *lower panel* is a zoomed-in view of the disk region, highlighting the Hα excess sources located in this area.

termediate magnitudes. The lower panel of Figure 12 focuses on the r-band magnitude distribution for the subset of Hα excess objects in the disk. The blue curve highlights these objects, showing they tend to have brighter magnitudes, with a peak around 13.25 in the r band. This suggests that Hα excess objects could be intrinsically more luminous or closer to us than the general population of all stars.



totally

I find this unnecessary. I think you attribute fig14 a meaning that is more simple. both group of stars show similar distribution . Thats it, nothing more nothing less.

Fig. 14. Distribution of the objects in galactic longitude for Hα excess sources (blue bars) and all stars (pink bars) for the main survey (*left panel*) and the disk ~~area~~ (*right panel*).

galactic

be consistent! sometime you write, disk, some disk region some disk area. I suggest "Galactic disk"

the

Figure 13 shows the distribution of all Hα excess sources in Galactic latitude and longitude, along with a zoomed-in view of the disk region in the bottom panel. The distribution of objects in Galactic longitude for the main survey (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 main stars (pink bars). This suggests that Hα excess sources are distributed throughout the galaxy within the area observed by DR4 S-PLUS. Two relatively small peaks are observed around Galactic longitudes of 15° and 50°, with a larger peak around 270°. Notably, these peaks are also present in the general star population of the main survey. The peaks suggest that Hα excess sources are influenced by specific structural features of the Milky Way or are linked to certain types of objects, such as stellar binary systems, variable stars, and quasars, as indicated by the SIMBAD cross-match results. The peak around 270°, present in both Hα excess sources and the general star population, points to a significant concentration of these objects in that region, possibly reflecting the distribution of evolved stars or stellar remnants.

On the other hand, the bottom panel of Figure 13 and the right panel of Figure 14 show the distribution of objects in Galactic longitude specifically within the disk region. Unlike the main survey, there is a noticeable concentration of Hα excess sources at specific longitudes, particularly around certain regions of the Galactic plane, with a larger peak around 243°. This concentration may indicate regions with higher star formation rates. Additionally, there are two small peaks around 225° and 268° in Galactic longitude. These small peaks likely signify localized concentrations of Hα excess sources within the Galactic disk, suggesting regions where there is an elevated presence of ionized hydrogen emission, indicative of active star formation processes. Such features are typically associated with dense molecular clouds or areas of enhanced interstellar gas density, where conditions are favorable for the formation of new stars. Despite Hα excess sources following a distribution similar to that of all stars, the peaks tend to be higher for Hα excess sources.

My opinion is that you exaggerate the figures 13 and 14.

4. 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 main survey 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 was is used to reduce the dimensions of our data and perform a feature extraction, while HDBSCAN classified 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 the other adding filters from the Wide-Field Infrared Survey Explorer (Wright et al. 2010, WISE). 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. These approaches were applied in this work to the Hα excess sources from the main survey only.

This methodology is applied to the list of Ha excess sources obtained from the main survey of SPLUS.

4.1. Dimensionality Reduction and Clustering

4.1.1. UMAP

what do you mean "by learning"?

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. By learning a low-dimensional representation, UMAP aims to preserve intricate nonlinear relationships present in the original high-dimensional features. 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`[9]. UMAP has three key hyperparameters: `n_neighbors`, `n_components`, and `min_dist`.

The `n_neighbors` parameter balances local versus global structure 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.

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

If i were you i would simply described the plots without make any assumption.
in this case, you will avoid problem with the referee (why this, why that ,explain this explain that.)

i have not worked with these codes. They also look a bit complicated. You exaplain lot of thing and the text seems ok.

A grid of tests was constructued over a range of .....

The n_components parameter, similar to the parameter used in standard dimension reduction algorithms in the scikit-learn package, allows us to set the number of dimensions in the reduced space into which we will embed the data.

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.

4.1.2. 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[10] (McInnes et al. 2017). The two most critical parameters are "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 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.

Honestly, I had not comments on section 4.

4.2. Classification Results

Our unsupervised UMAP model projects the data, and HDBSCAN subsequently identifies the clusters. To ensure high-quality photometry, we set the criteria for the error to be less than 0.2 in all filters. This results in a list of 2181 objects for the main survey. This reduction in sample size is crucial to mitigate the influence of noisy or unreliable photometric measurements, which can significantly affect the performance and outcomes of both the UMAP and HDBSCAN algorithms. By focusing on high-quality photometric data, we aimed to enhance the accuracy and robustness of the clustering results, thereby providing more reliable classifications of the Hα excess sources.

To perform cross-validation for choosing 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 quantitative metrics, including the Silhouette Score and Davies-Bouldin Index, to objectively evaluate the performance of different parameter combinations. Silhouette Score measures how well-defined the clusters are in the reduced space, with higher values indicating better separation between clusters, while Davies-Bouldin Index evaluates the average similarity between each cluster and its most similar cluster, with lower values indicating better-defined clusters. A systematic grid search was performed 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, and the metrics were computed to determine the optimal parameter set.

For the min_dist we used the defult value=0.1

4.2.1. Using just the S-PLUS photometry

For the first experiment, where we used the 66 S-PLUS colors as input parameters, we applied the metric evaluation method described above. After evaluation, we identified that setting n_neighbors = 50 and n_components = 2 yielded the highest silhouette score and lowest Davies-Bouldin Index, indicating optimal performance[11]. Subsequently, we adopted these values for these hyperparameters. The min_dist parameter was maintained at its default value of 0.1. Following dimensionality reduction with UMAP, the resultant variables were utilized to construct HDBSCAN models. We experimented with varying "minimum cluster size" and "minimum number of samples", ultimately selecting min_cluster_size = 5 and min_samples = 50. Euclidean metric was employed for distance calculations throughout.

The left panel of Figure 15 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 main survey. 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.

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 suggests that Group 0 primarily comprises extragalactic sources, with a peak in the redshift distribution around 2.45, likely indicating active galactic nuclei or similar objects.

Group 1 contains 166 objects, 149 of which have matches in SIMBAD. 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 a mix of variable stars.

Group 2 includes 1539 objects with a diverse range of variable stars. The majority are eclipsing binaries (275), followed by RR Lyrae stars, cataclysmic variables, and quasars. This group

we epxermented various compination for the ????? and ???? parameters. We ended up with the optimal value of 5 and 50, respectively.

[10] https://hdbscan.readthedocs.io/en/latest/

[11] To estimate the silhouette score and Davies-Bouldin Index, we used the Python package scikit-learn

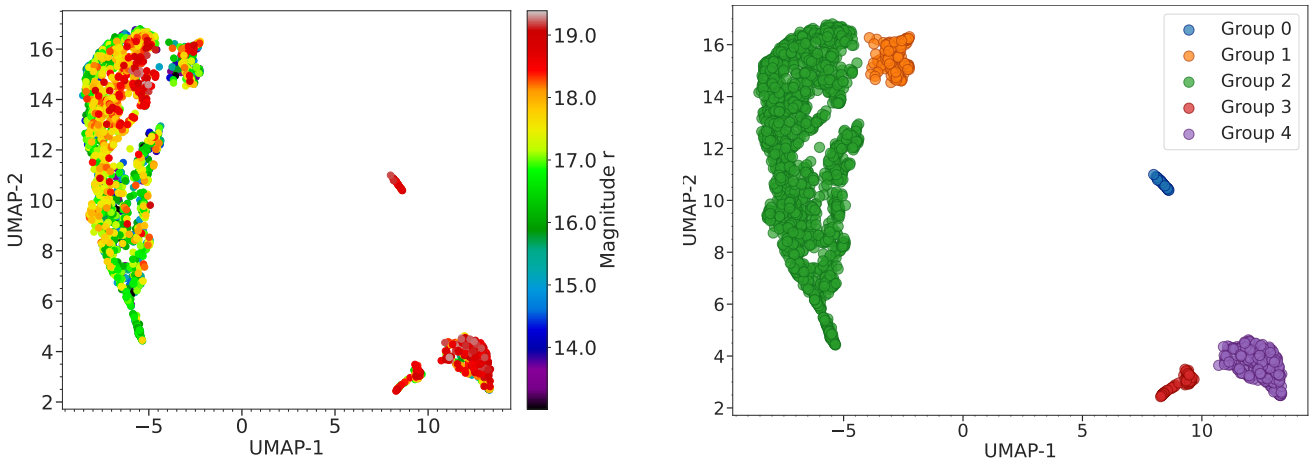


Fig. 15. After dimension reduction by UMAP to two dimensions. The left panel indicates the UMAP result using only the S-PLUS colors as input parameters, while the right panel shows the result after adding other colors created using W1 and W2 bands of WISE to generate additional features.

I do not undestand why these two figures are different. you say nothing in the captions.!!!

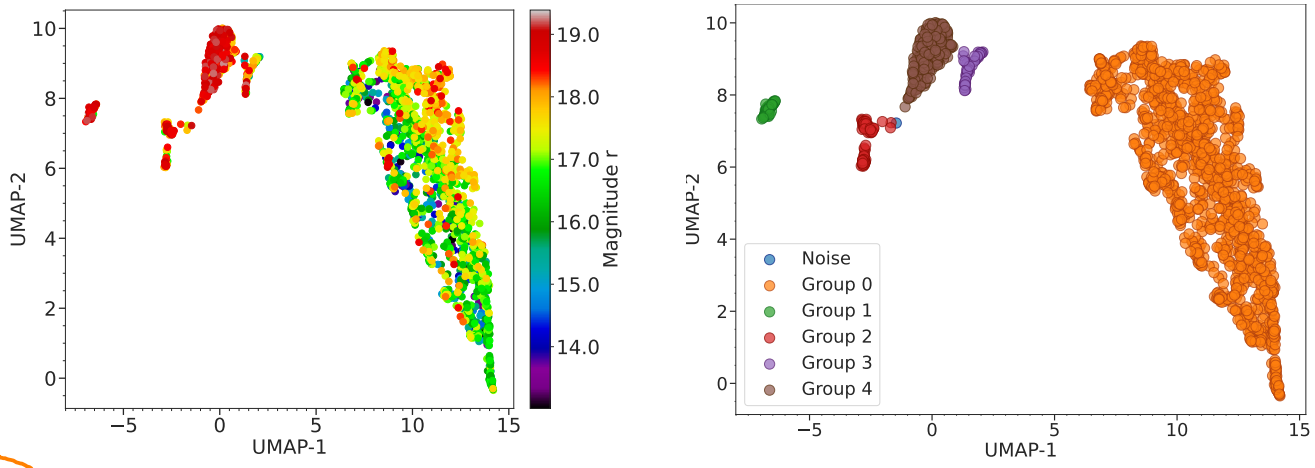


Fig. 16. After dimension reduction by UMAP to two dimensions. The left panel indicates the UMAP result using only the S-PLUS colors as input parameters, while the right panel shows the result after adding other colors created using the W1 and W2 bands of WISE to generate additional features.

indicates a significant presence of binary star systems and other variable types.

Group 3 consists of 93 objects, predominantly QSOs (17) and Seyfert 1 galaxies (10). Other identifications include AGN candidates, radio sources, and a few galaxies. The redshift distribution for extragalactic objects in this group ranges approximately from 0.31 to 0.37.

Group 4 includes 325 objects with a high concentration of QSOs (78) and cataclysmic variables (25). Additionally, this group features a mix of blue stars, AGNs, radio sources, and white dwarf candidates. The extragalactic objects in this group have a peak in the redshift distribution around 1.35.

In summary, our application of UMAP and HDBSCAN to the Hα excess sources has effectively identified distinct groups with varying astrophysical characteristics. The classification successfully separates 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. This separation enhances our understanding of the objects in our dataset and provides a foundation for further detailed analysis.

4.2.2. Using the S-PLUS photometry PLUS WISE

The second experiment included the W1 and W2 WISE filters, adding new colors to the original variable set used for the machine learning models. To do this, we first crossmatched the Hα sources of the main survey with the ALLWISE catalog using a search radius of 2 arcsec, obtaining 3,173 matches. This number was then reduced to 1,910 after applying error criteria to the S-PLUS filters and filtering for objects with errors less than 0.5 in the W1 and W2 filters. The additional colors combined the WISE bands (W1 and W2) with the S-PLUS broadband filters. For instance, we calculated colors such as W1 - W2, W1 - *u*, W2 - *u*, W1 - *g*, W2 - *g*, W1 - *r*, W2 - *r*, and so on. This resulted in 11 new colors being added to the original 66 S-PLUS colors, generating a dataset with 77 variables in total.

Figure 16 shows the results of the reduction in dimensionality and the groups identified by applying UMAP followed by

In general, I would say that the sources with high UMAP1 parameters seems to be extragalactic while those with low UMAP1 stellar.

General question. why did you use W1 and W2 bands
but not 2MASS bands?
Can we reply that to the referee? is there any
specific reason?

HDBSCAN, using the input parameters described in the previous paragraph. On this occasion, HDBSCAN found five groups and five objects that were classified as noise. Table 3 summarizes these results:

Group 0 ontains 1,437 objects, 424 of which match with SIMBAD. 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 have matches in 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.

Group 2 consists of 93 objects, 43 of which have SIMBAD matches. 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.

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 have SIMBAD matches. 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. The redshift distribution has a peak around 1.35.

In summary, including the WISE filters in our analysis has further refined the clustering of Hα excess sources. The integration of WISE data has allowed us to more effectively distinguish between galactic and extragalactic sources, providing a richer understanding of the objects in our dataset. Notably, it is possible to separate the CVs from the QSOs that have a redshift around 1.35. Revisit Section 3.3 to see the redshifted line of the extragalactic objects detected here in the *J*0660 filter.

groups 1,2, and 4 are domianted by extrgalactic sources

4.3. Extracting Main Features: Color Analysis

Based on the important features extracted from the Random Forest model, we focused on the colors derived from the S-PLUS and WISE filters. These colors effectively distinguish between the groups identified by UMAP + HDBSCAN among Hα-excess objects in S-PLUS data.

i do not understand that!

In the main survey of the Hα-excess list, we identified extragalactic emitters with red-shifted bluer emission lines resembling the Hα emission line, with sources having a redshift of less than 0.02. 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. 4.2 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 best contributed to the separation or classification of the classes of objects. We implemented the `scikit-learn` package for the 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

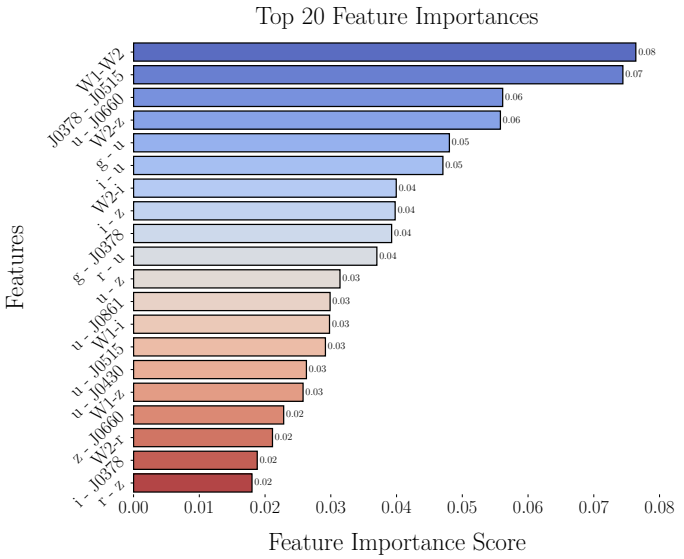


Fig. 17. Top 20 feature importances identified by the Random Forest model, showing the colors that contributed most significantly to the clustering of Hα-excess objects using UMAP + HDBSCAN. The importance scores indicate the relative impact of each color on the classification of different object classes.

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). Despite this imbalance, we opted not to apply additional techniques to manage it because the Random Forest classifier achieved an F1 Macro Average of 0.96 ($\pm$0.13) during 5-fold cross-validation. This high score, coupled with low variability, indicates that the model effectively handles the imbalance and consistently classifies different clusters.

After performing the model, we accessed the feature importances using `feature_importances` from the Random Forest package. Figure 17 shows the top 20 feature importances and their respective scores, indicating the colors that contributed 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 to generate all possible color-color diagrams using the top 20 features. This allowed us to identify the color-color diagrams that best separate the different classes of objects and choose the most effective ones. Figure 18 shows six examples of color-color diagrams that we selected for their ability to best separate the groups found in our Hα-excess sources list. These diagrams are constructed based on the main features of importance.

better

In Appendix A.1, we present additional color-color diagrams that also effectively separate the groups or classify the object classes identified by HDBSCAN. These diagrams are likewise based on the top 20 features.

This exercise demonstrates that using specific color-color diagrams with selected filters can effectively classify objects clustered using machine learning techniques. 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 4.2, 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 classifi-

more

It is know that CVs can be separared from QSOs in the W1-W2 colour (e.g. Scaringi et al. 2011?)

These sources are characterized by lower UMAP1 values
in comparison with the majority of the stellar sources.
Group 3 with low UPAM1 values also lists stellar sources (e.g. CVs, WDs etc.)
suggeating similar photometric characteristics with the extragalactic group.

Table 3. Summary of clustering outcomes achieved using the UMAP and HDBSCAN unsupervised machine learning methods applied to Hα excess sources. Clustering is performed using S-PLUS and S-PLUS + WISE filter combinations for both the main survey. 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**	

cation of more objects, as it circumvents the need for complete data across all filters.

the SPLUS

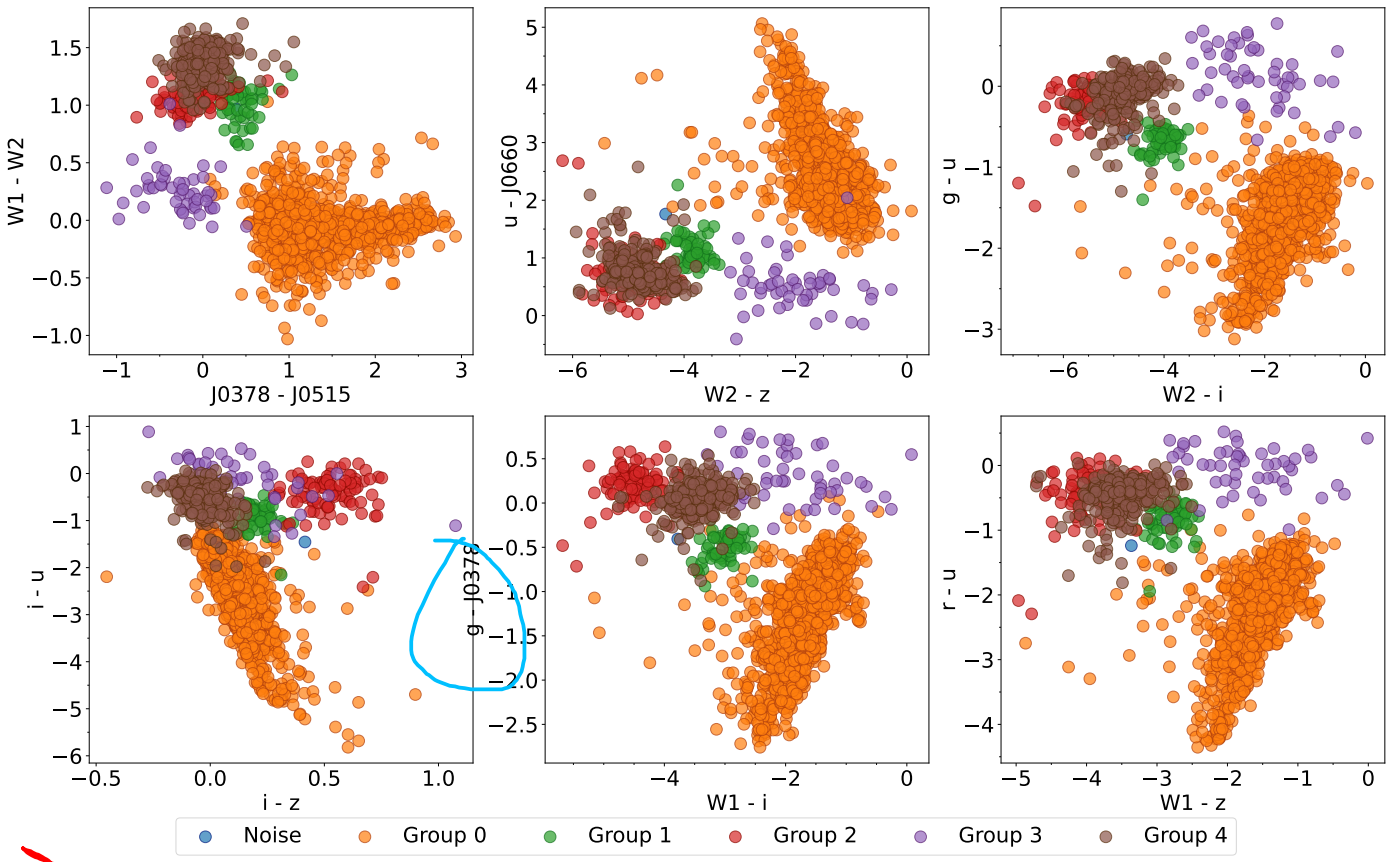
5. Conclusions

In this study, we leveraged the extensive capabilities of the S-PLUS project to explore and classify Hα-excess sources in the Southern Sky. Our approach was systematic and multifaceted, comprising three key steps: the identification of potential Hα-excess sources, the differentiation of these sources based on their color properties, and the validation of our findings through SIMBAD counterpart and various spectroscopic measurements.

this sections has to be rewritten and better describe the results and conclusions.
it is very confusing.

We identified Hα-excess sources using the narrow $J0660$ filter along with the broad r and i filters from S-PLUS. By analyzing a $(r-J0660)$ versus $(r-i)$ color-color diagram, we were able to differentiate between main-sequence and giant stars and identify potential Hα-excess objects as outliers. From this analysis, we selected 3,637 sources from the high-latitude main survey and 3,734 from the Galactic disk, resulting in a combined total of 7,371 Hα-excess candidates from both areas.

Among these candidates, we found matches with the SIMBAD database, identifying various classes of emission line objects, including YSOs, Be stars, CVs, PNs, and EB stars. Additionally, we identified quasars (QSOs), non-local galaxies, and objects with Hα in absorption, such as RR Lyrae stars. This un-



This has to be rephrased

Fig. 18. Examples of color-color diagrams using the top 20 features identified by the Random Forest model. These diagrams illustrate the separation of different classes of objects found in the Hα-excess sources list. The selected diagrams demonstrate effective clustering achieved through UMAP + HDBSCAN, highlighting the key colors that contribute to the classification.

what do you mean?
confirm the accuracy? does make any sense!

derscores the importance of meticulous classification to accurately distinguish between different Hα-related phenomena. We further validated our findings by comparing them with spectroscopic data from LAMOST and SDSS. Our analysis revealed that approximately 60% of the spectra from LAMOST and SDSS exhibit Hα emission lines, while around 30% show Hα in absorption. This comparison confirmed the accuracy and limitations of our classifications and validated the reliability of our Hα-excess source identifications. Additionally, for the disk, VPHAS+ data corroborated our results.

To enhance the robustness of our classifications, we employed machine learning techniques, specifically UMAP for dimensionality reduction and HDBSCAN for clustering, to effectively analyze the Hα-excess sources. Using the 12 S-PLUS filters, we distinguished between Hα-emission objects and those with Hα in absorption, such as RR Lyrae stars. Our methods allowed us to separate compact extragalactic objects from stars, although differentiating cataclysmic variables from QSOs or AGN with redshifts around 1.35 remained a challenge. Incorporating additional colors from WISE filters further refined the clustering process, enabling more precise separation of extragalactic sources from Galactic ones and improving differentiation between cataclysmic variables and QSOs. The inclusion of WISE data in a Random Forest model was instrumental in identifying the most significant features for classification. This approach generated effective color-color diagrams, facilitating the separation of different object classes. Including diverse objects with different Hα characteristics, such as Hα emission lines, compact extragalactic objects (with redshifted lines), variable stars, and objects with absorption lines, improved the robustness of our study. Our comprehensive analysis, utilizing UMAP, HDBSCAN, and Random Forest models, effectively managed this variability and ensured accurate classification, providing valuable insights into Hα-excess phenomena.

Future work could focus on expanding the sample size and exploring additional spectroscopic data to further refine our classifications. Moreover, applying these methods to other regions of the sky or different wavelengths could enhance our understanding of Hα-excess sources and their astrophysical contexts. Our study provides a solid foundation for ongoing research and highlights the potential of combining advanced observational and analytical techniques to address complex astronomical questions.

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). DGR acknowledges the CNPq (428330/2018-5; 313016/2020-8) and FAPERJ (269312) grants. F. A. -F. acknowledges funding for this work from FAPESP grants 2018/20977-2 and 2021/09468-1. FRH ackowledges funding from FAPESP through the project 2018/21661-9. C. C. is supported by the National Natural Science Foundation of China, No. 11803044, 11933003, 12173045. 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

it is confusing !!!
it is not easy for a reder to understand it

We did not that.
emission and absorpion line sources were never separated?
I am probably confused.

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

Ahumada, R., Prieto, C. A., Almeida, A., et al. 2020, ApJS, 249, 3
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
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
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
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
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
Groot, P. J., Verbeek, K., Greimel, R., et al. 2009, MNRAS, 399, 323
Gutiérrez-Soto, L. A., Gonçalves, D. R., Akras, S., et al. 2020, A&A, 633, A123
Jacoby, G. H., Kronberger, M., Patchick, D., et al. 2010, PASA, 27, 156
Kron, R. G. 1980, ApJS, 43, 305
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
Monguió, M., Greimel, R., Drew, J. E., et al. 2020, A&A, 638, A18
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
Pickles, A. J. 1998, PASP, 110, 863
Pollmann, E., Bennett, P. D., Vollmann, W., & Somogyi, P. 2018, Information Bulletin on Variable Stars, 6249, 1
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
Scaringi, S., Monguió, M., Knigge, C., et al. 2023, MNRAS, 518, 3137
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
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
York, D. G., Adelman, J., Anderson, John E., J., et al. 2000, AJ, 120, 1579

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

Appendix A: More color-color diagrams

This section presents additional color-color diagrams using the colors identified by the Random Forest model as most contributing to the classification. These diagrams further demonstrate the separation and clustering of different classes of objects in the Hα-excess sources list, as achieved through the UMAP + HDBSCAN approach. The extended set of diagrams, shown in Figure A.1, highlights the key colors that enhance the effectiveness of object classification.

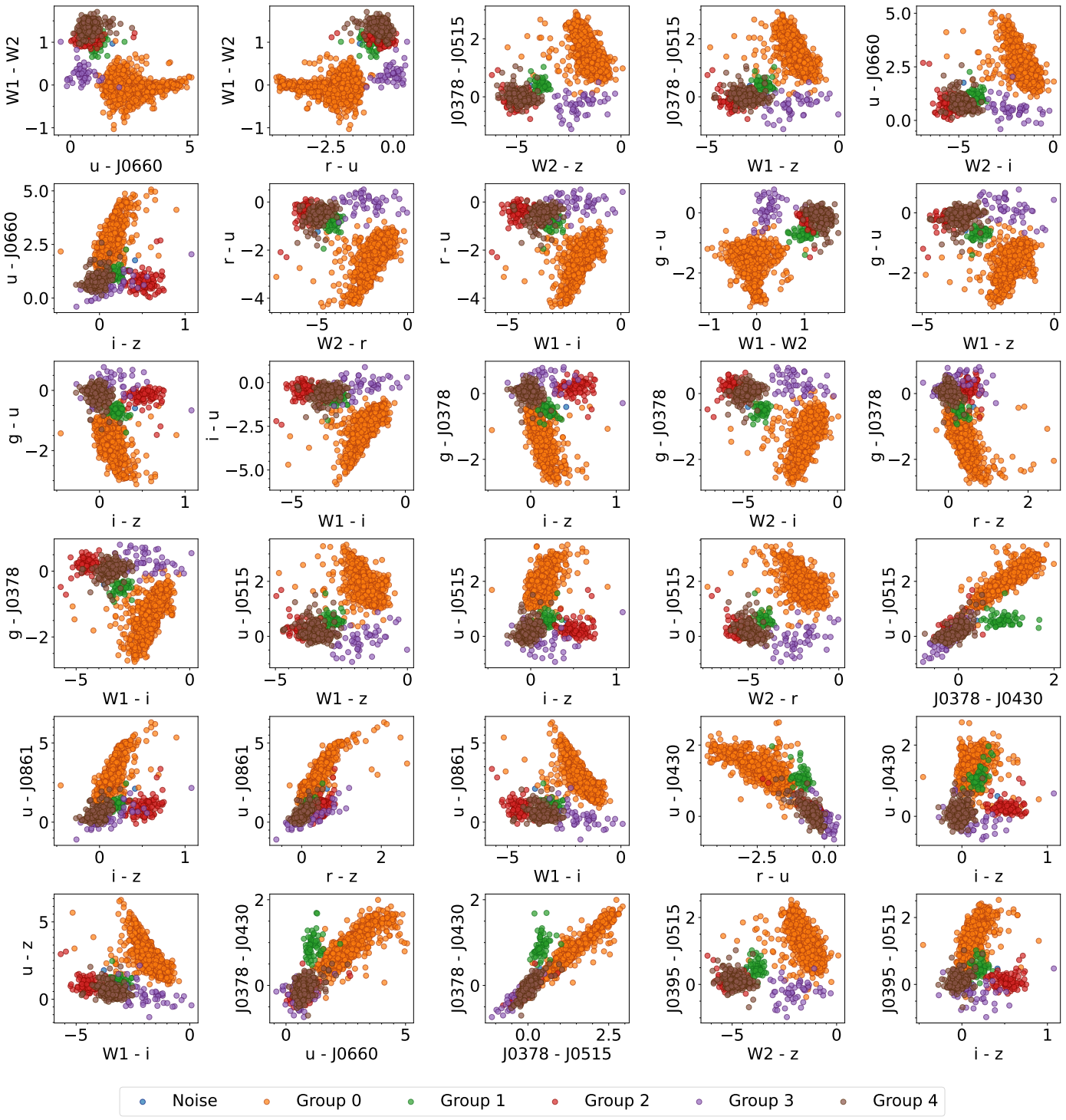


Fig. A.1. Extended set of 30 color-color diagrams using the top 20 features identified by the Random Forest model. These diagrams further illustrate the separation of different classes of objects found in the Hα-excess sources list. The diagrams demonstrate effective clustering achieved through UMAP + HDBSCAN, showcasing the key colors that contribute to the classification.