MONTHLY NOTICES of the Royal Astronomical Society

Hα emitters from the Southern Photometric Local Universe Survey (S-PLUS)

Journal:	*Monthly Notices of the Royal Astronomical Society*
Manuscript ID	MN-22-3687-MJ
Manuscript type:	Main Journal
Date Submitted by the Author:	21-Sep-2022
Complete List of Authors:	GUTIERREZ SOTO, LUIS ANGEL; Universidade de São Paulo Instituto de Astronomia Geofísica e Ciências Atmosféricas Lopes de Oliveira, Raimundo Akras, Stavros; National Observatory of Athens Institute for Astronomy Astrophysics Space Applications and Remote Sensing, Gonçalves, Denise; UFRJ, Observatorio do Valongo Mendes de Oliveira, Claudia Almeida Fernandes, Felipe; Universidade de São Paulo Instituto de Astronomia Geofísica e Ciências Atmosféricas, Astronomia Herpich, Fábio; Universidade de Sao Paulo, Instituto de Astronomia, Geofísica e Ciências Atmosféricas; Universidade Federal de Santa Catarina, Physics Cheng, Cheng; National Astronomical Observatories,Chinese Academy of Sciences Goncalves, Thiago; Universidade Federal do Rio de Janeiro, Observatorio do Valongo Nakazono, Lilianne; Universidade de São Paulo Instituto de Astronomia Geofísica e Ciências Atmosféricas, Telles, Eduardo; Observatorio Nacional, Astronomia Alvarez-Candal, Alvaro; Observatorio Nacional, Kanaan, Antonio; Universidade Federal de Santa Catarina, Departamento de Física Ribeiro, Tiago; National Optical Astronomy Observatory; Universidade Federal de Sergipe, Departamento de Fisica Schoenell, William; Instituto de Astrofísica de Andalucía (CSIC)
Keywords:	surveys < Astronomical Data bases, techniques: photometric < Astronomical instrumentation, methods, and techniques, (stars:) novae, cataclysmic variables < Stars, galaxies: dwarf < Galaxies, (galaxies:) quasars: emission lines < Galaxies

MNRAS 000, 1–19 (2021) Preprint 21 September 2022 Compiled using MNRAS LATEX style file v3.0

Hα emitters from the Southern Photometric Local Universe Survey (S-PLUS)

L. A. Gutiérrez-Soto[1]★, R. Lopes de Oliveira[1,2,3], S. Akras[4], D. R. Gonçalves[5], C. Mendes de Oliveira[1], F. Almeida-Fernandes[1,6], F. R. Herpich[1], C. Cheng[7], T. S. Gonçalves[5], L. Nakazono[1], E. Telles[3], A. Alvarez-Candal[3,8,9], A. Kanaan[10], T. Ribeiro[11], W. Schoenell[12]

[1]*Departamento de Astronomia, IAG, Universidade de São Paulo, Rua do Matão, 1226, 05509-900, São Paulo, Brazil*
[2]*Departamento de Física, Universidade Federal de Sergipe, Av. Marechal Rondon, S/N, 49100-000, São Cristóvão, SE, Brazil*
[3]*Observatório Nacional, Rua Gal. José Cristino 77, 20921-400, Rio de Janeiro, RJ, Brazil*
[4]*Institute for Astronomy, Astrophysics, Space Application & Remote Sensing, National Observatory Athens, GR-15236, Athens, Greece*
[5] *Observatório do Valongo, Universidade Federal do Rio de Janeiro, Ladeira Pedro Antonio 43, 20080-090, Rio de Janeiro, Brazil*
[6] *Community Science and Data Center/NSFś NOIRLab, 950 N. Cherry Ave., Tucson, AZ 85719, USA*
[7] *Chinese Academy of Sciences South America Center for Astronomy, National Astronomical Observatories, CAS, Beijing 100101, China*
[8] *Instituto de Astrofísica de Andalucía, CSIC, Apt 3004, E18080 Granada, Spain*
[9] *Instituto de Física Aplicada a las Ciencias y las Tecnologías, Universidad de Alicante, San Vicent del Raspeig, E03080, Alicante, Spain*
[10] *Departamento de Física, Universidade Federal de Santa Catarina, Florianópolis, SC, 88040-900, Brazil*
[11] *NOAO, P.O. Box 26732, Tucson, AZ 85726*
[12] *GMTO Corporation 465 N. Halstead Street, Suite 250 Pasadena, CA 91107*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

The ongoing multi-band survey performed by the S-PLUS project will have covered 9300 deg^2 of the Southern skies by the time it is completed. S-PLUS has a crucial feature: images over the whole field taken in the Hα narrow-band. The Hα transition provides a superb tool for the study of a number of important astrophysical processes and, in particular, it allows the classification of different types of astrophysical sources. Here we explore the S-PLUS data release 3, which covers 2000 deg^2, including the Stripe-82 area, to highlight the potential of the survey for finding compact Hα emitters using the (r - $J0660$) versus (r - i) colour-colour diagram and in distinguishing the red from the blue sources based on the (r - z) versus (g - r) diagram. Our Hα-emitter catalog contains 2,187 objects that exhibit excess in the narrow $J0660$ band. For 1,869 of them (or 86%), the excess is thought to be due to the Hα emission line, while for the remaining, the excess may be due to redshifted lines. Unsupervised clustering machine learning approach reveals two distinct populations: one with an intense blue continuum and another one with an intense red one. The hierarchical agglomerative clustering algorithm (`HAC`) was compared with the hierarchical density-based cluster selection (`HDBSCAN`) in order to reinforce the robustness of the red and blue populations' classification. By adopting a so-called "soft" clustering approach, we assigned the probability of each emitter belonging to a given population, blue or red. Around 87% of emitters were successfully classified as blue or red sources. We use synthetic and observed spectra to emphasize the potential of colour-colour diagrams in distinguishing several classes of Hα emission-line emitters that include planetary nebulae, H II regions, young stellar objects, symbiotic stellar systems, cataclysmic variables, blue compact galaxies, star-forming galaxies, and quasars. In summary, the method described in detail in this paper is shown to be an efficient tool to find new emitters and to classify them, using multi-colour data.

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

1 INTRODUCTION

Atomic excitation followed by recombination in Balmer hydrogen emission lines may be ignited in different ways, thermal and non-thermal collisional excitation in shock-heated gas and energetic photons acting over a diffuse gas. As a practical result, and the Universe being hydrogen abundant, the observation of those electronic transitions offer an important window into the study of astrophysical objects. Among all the possible electronic transitions, the Balmer series represent extremely useful tools in Astronomy. Particularly, 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 $n = 2$

★ E-mail: gsoto.angel@gmail.com

energy level, 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, interacting binary systems that are experiencing mass exchange, like symbiotic stars (SySt), cataclysmic Variables (CVs), among others.

At much larger scales, Hα emission is also detected in extended no-point sources like PNe, H II regions, supernova remnants, as well as star-forming regions in galaxies, among others. In the case of high redshifted sources like starburst galaxies and quasi-stellar object (QSOs), the detection of an 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 SySt, with only ~ 65 of those located in galaxies other than the Milky Way (Akras et al. 2019a; Merc et al. 2019). The number of known PNe in our Galaxy is of 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 in a variety of angular resolutions, sky coverage, and sensitivity were carried out in the past. Some of 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. 2019b), early-type emission-line stars (Drew et al. 2008), and PNe (Viironen et al. 2009; Sabin et al. 2010; Akras et al. 2019c).

There are two ongoing multi-band surveys 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.

Here we present a census of Hα emitters from the S-PLUS DR3 by employing the (r - $J0660$) versus ($r - i$) colour-colour diagram and unsupervised machine learning techniques to classify them as blue or red sources. Section 2 describes the observations related to the S-PLUS project, as well as important information on the third data release. It also presents the technique implemented to select the Hα emitters and machine learning approaches used to divide the sample into two populations based on their colours. In Section 3 our findings are described, and finally Section 4 discusses our main results and conclusions.

2 METHODOLOGY

2.1 Observations: the S-PLUS project

This manuscript uses data from the S-PLUS DR3 (Buzzo et al., in prep), which covers 2,000 square degrees of the southern sky. The S-PLUS DR3 can be accessed in the database of the project, `S-PLUS Cloud`[3]. S-PLUS is being carried out by a dedicated 0.83m robotic telescope located at Cerro Tololo, Chile (Mendes de Oliveira et al. 2019). The project is surveying the southern sky using the 12 filters from the so-called Javalambre filter system (Marín-Franch et al. 2012), that spans the wavelength range from 3000Å to 10000Å. The system includes seven narrow-band filters ($J0378$, $J0395$, $J0410$, $J0430$, $J0515$, $J0660$, and $J0861$) and five broad-band Sloan-like (Fukugita et al. 1996) filters (see Fig. 1). The narrow-band $J0660$ filter used in S-PLUS is centred at lambda 6614 Å and has a width of about 147 Å (Table 2 of Mendes de Oliveira et al. 2019), and therefore 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.

The data set used for this study, DR3, includes about 60 million objects distributed over ~2,000 deg^2 (of the total of ~8,000 deg^2 of high Galactic latitudes fields with b > 30° planned to be covered when the survey is complete). The galactic disk and bulge are not included in DR3 despite S-PLUS plans to cover an area ~ 1,300 deg^2 of them and will be available in DR4. 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). In order to ensure that high-quality data are used in the present analysis, only objects detected simultaneously in at least the r, i and $J0660$ bands, with errors less than 0.2 mag, are considered. We also selected objects which are probably point sources by setting `CLASS_STAR` > 0.5. Following Almeida-Fernandes et al. (2022), we implemented `PhotoFlag` = 0 in the filters r, $J0660$ and i for the selection of targets with good photometry in these three filters.

The first goal of this work is the identification of Hα emitters in the S-PLUS DR3. For this, we applied an iterative and automatic technique to select objects with an excess in the $J0660$ band, which is consistent with the detection of the Hα line in emission. Next, the sample of Hα sources is divided into two subgroups: the blue and red one. This classification was made by employing optical colours in combination with unsupervised machine learning/statistical tools. These procedures are described in the following subsections.

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

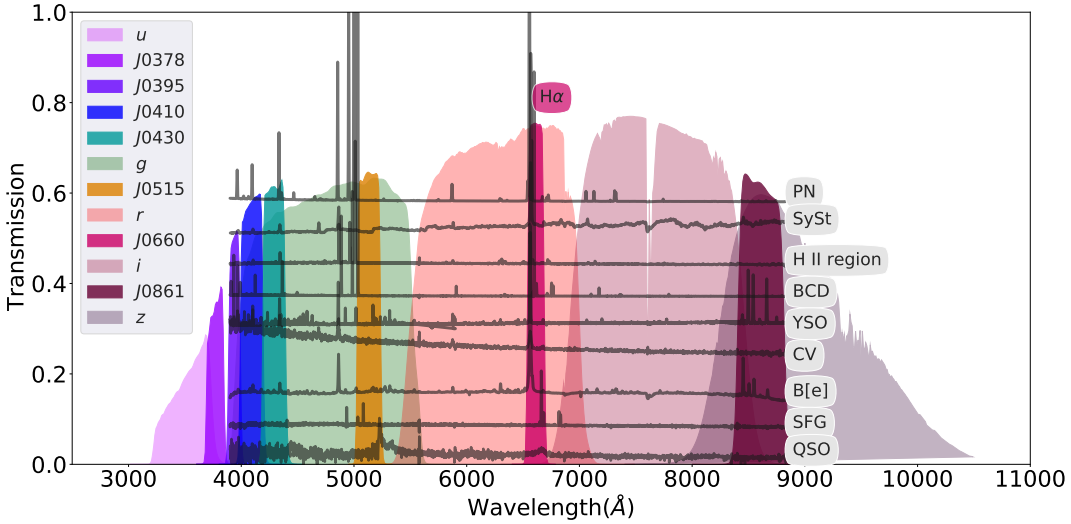


Figure 1. Transmission curves of the S-PLUS filter set. The narrow-band filter $J0660$ includes the Hα emission line. Over-plotted are spectra of different classes of emission line objects. From top to bottom: a PN, a symbiotic star, an extragalactic H II region, a blue compact/H II galaxy, a YSO, a CV star, a B[e] star, a star forming galaxy and a QSO at a redshift of ~3.31.

2.2 Selection of Hα emitters

Before searching for the potential Hα emitters hidden in the S-PLUS DR3 footprint, we first divided our sample into four sub-samples based on their magnitudes in the r-band: (i) $13 \leq r < 16$, (ii) $16 \leq r < 18$, (iii) $18 \leq r < 20$, and (iv) $20 \leq r < 21$. 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.

The identification of Hα emitters is based on the method successfully applied by Witham et al. (2008) to the IPHAS project, given that similar filters are also available in S-PLUS: r, $J0660$, and i. The same technique was also used by Scaringi et al. (2013) and Wevers et al. (2017) to reveal Hα emitters.

We first generated the (r - $J0660$) versus (r - i) diagram for each sub-sample and attempted to fit the loci mainly occupied by main-sequence and giant stars with a linear regression. We then implemented an iterative σ-clipping technique so that, by construction, Hα emitters should satisfy the condition:

$$(r - J0660)_{\rm obs} - (r - J0660)_{\rm fit} \geq C \times \sqrt{\sigma_{\rm s}^2 + \sigma_{\rm phot}^2} \qquad (1)$$

where $\sigma_{\rm s}$ is the root mean squared value of the residuals around the fit and $\sigma_{\rm phot}$ is the error on the observed $(r - J0660)$ colour index. C is a constant parameter with a value of 4, following Wevers et al. (2017). The fits were made by employing `astropy.modeling` [4].

Fig. 2 illustrates the procedure applied. The solid black lines indicate the initial fit and the dashed lines show the 4-σ clipping fit. The dotted lines correspond to the selection criteria for the Hα emitters, 4-σ above of the final fit. It is worth be noted that these cut-off lines are just approximations, as they only represent the residual around the fit. The photometric uncertainty of the (r - $J0660$) colour index for each individual point is also taken into account (see Equation 1).

Once the list of Hα emitters was obtained, we proceeded with a visual inspection of their false-colour images and their spectral energy distributions, constructed with 12 points corresponding to the 12 S-PLUS filter mean magnitudes for each source, hereafter called the S-spectra, to remove artefacts from the list. We repeat the methodology explained above without considering `PhotoFlag` = 0, to recover objects with good photometry labeled with a flag other than 0.

The upper panel of Fig. 3 shows an example of what the S-spectrum of an Hα emitter looks like, while the bottom panel presents the SDSS spectrum of the same source. It is evident from the comparison of the two observable that the excess in the $J0660$ band is linked with the Hα emission line. The spectroscopic redshift of this object is 0.009, which puts the Hα emission line within the detectability range of the $J0660$ band, thus allowing us to assume that the excess is due to this emission.

The distribution of the Hα emitters in the $(r - J0600)$ versus $(r - i)$ colour-colour plane is shown in Fig. 4. The loci of the main-sequence and giant stars derived from synthetic spectra (Pickles 1998) convolved with the S-PLUS transmission curves in the AB magnitude system (Oke & Gunn 1983) are also plotted. All sources located above the locus of the main and giant stars exhibit an excess in $J0600$ filter, which it is attributed to Hα line. The wide distribution of sources across the $(r - J0660)$ and $(r - i)$ colour-colour diagram indicates that several types of Hα emitters are selected. Sources with high $(r - J0660)$ colour index are likely associated with PNe, H II regions, SySt or blue compact galaxies. On the other hand, the (r - i) colour index indicates redder sources such as SySt and YSO, while sources with strong blue continuum such as CVs and QSOs exhibit lower (r - i) values (see Fig. 2 of Gutiérrez-Soto et al. 2020).

Fig. 5 displays the distribution of all Hα emitters in Galactic latitude and longitude. The density map regions represent the spatial positions of the objects on the sky. The surface density of $J0660$-excess objects is highest near the Galactic plane. In fact, the distribution in function of the latitude present two peaks, one at -16°, which corresponds to the Stripe-82 region and another at -41° (see inset figure).

Our list of Hα emitters includes 2,187 sources. We now proceed to their classification into blue and red populations.

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

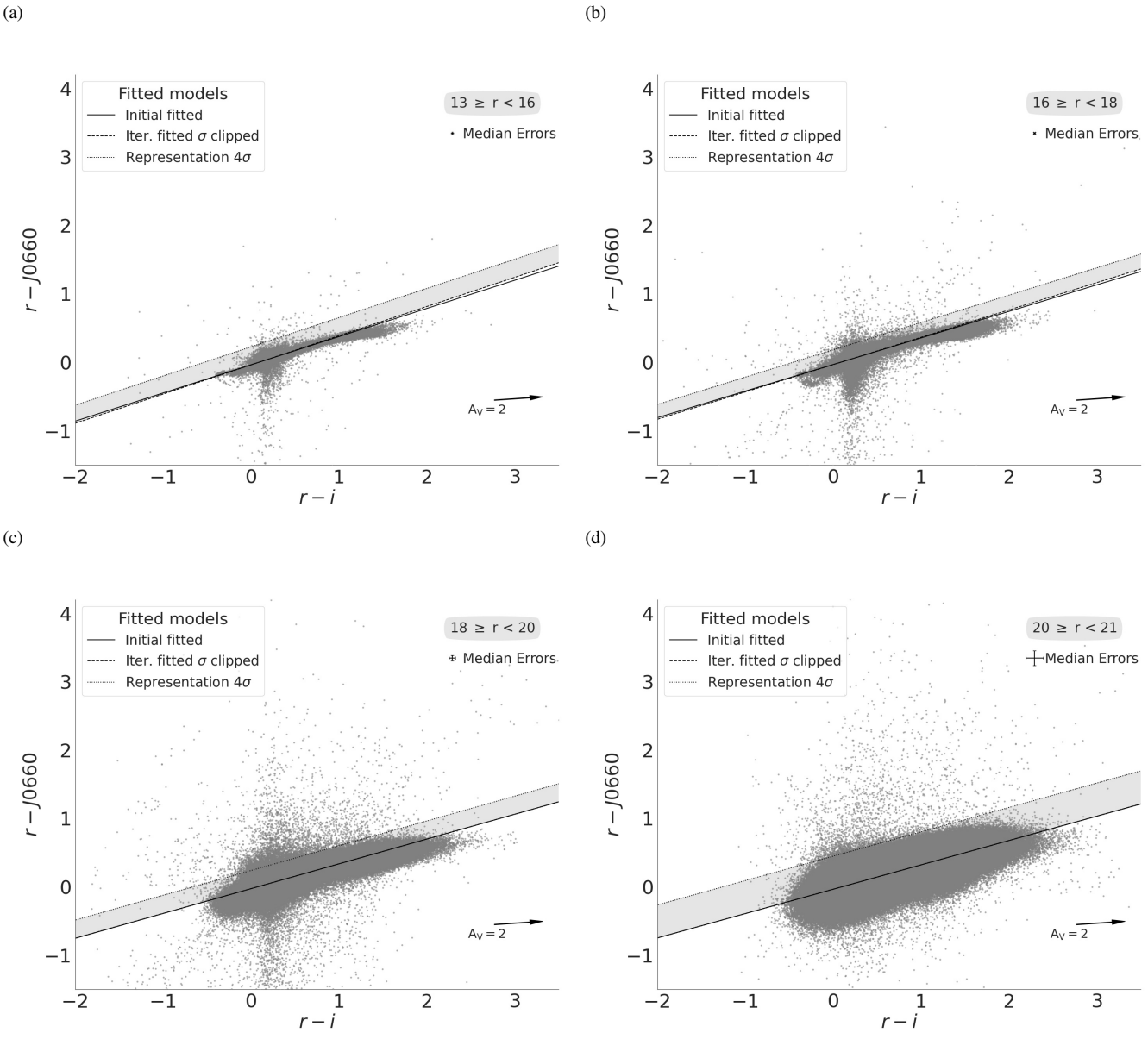


Figure 2. An illustration of the selection criteria used to identify strong emission-line objects via colour-colour plots. The data shown here are all from the S-PLUS DR3. The data are split up into four magnitude bins, as shown in the four panels. Objects with Hα excess should be located near the top of the colour-colour diagrams. The thin continuous lines illustrate the original linear fit to all the data (grey points). The dashed lines represent the final fits to the stellar locus of points which were obtained by applying an iterative σ-clipping technique to the initial fit. The actual cuts used to select Hα emitters are shown by the dotted lines. These correspond to 4-σ above of the final fit. Objects selected as Hα emitters must be located above the dotted line. Note that the position of these lines (selection criteria) shown in the figure are approximated, given that the actual selection criterion also considers the errors on each source.

2.3 Unsupervised machine learning clustering approach

For the split of the sample of Hα emitters into two classes, the blue and the red populations, we follow an unsupervised machine learning approach implementing two clustering techniques: hierarchical agglomerative clustering and hierarchical density-based cluster selection, both based on the $(r - z)$ and $(g - r)$ colours, whose results are mutually compared.

2.3.1 Hierarchical agglomerative clustering

Hierarchical clustering (HC) belongs to the family of clustering algorithms of which clusters are constructed by recursively grouping and splitting the sources. Being an unsupervised algorithm, HC does not require a training sample or pre-conceived hypotheses. Data elements are grouped based on patterns in a given space of parameters and on the levels of similarity at which the groupings change (Jain et al. 1999). In the end, HC returns a diagrammatic representation of the groups as a tree – a dendrogram that follows an hierarchical structure.

There are two types of hierarchical clustering: the *hierarchical*

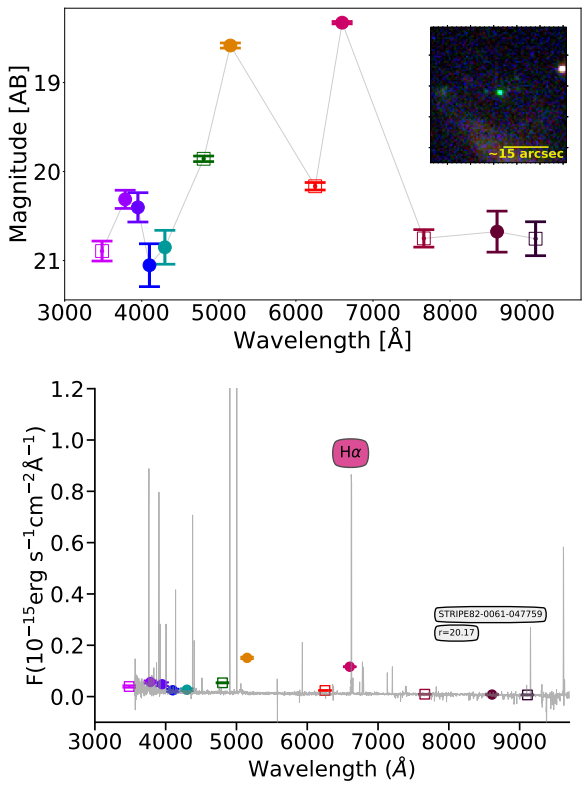


Figure 3. *Top panel:* S-spectrum of a random emitting object found by the method explained in Section 2.2. Open squares represent the SDSS-like broad-band filters. From left to right: u, g, r, i and z magnitudes. Circles represent the narrow-band filters, which from left to right correspond to $J0378$, $J0395$, $J0410$, $J0515$, $J0660$ and $J0861$. The inset figure is the coloured image of the object which was produced by combining all twelves bands. *Bottom panel:* SDSS spectrum and S-PLUS photometry in flux unity of the object - the Hα line is marked. Th spectroscopic redshift of this object is $z = 0.009$.

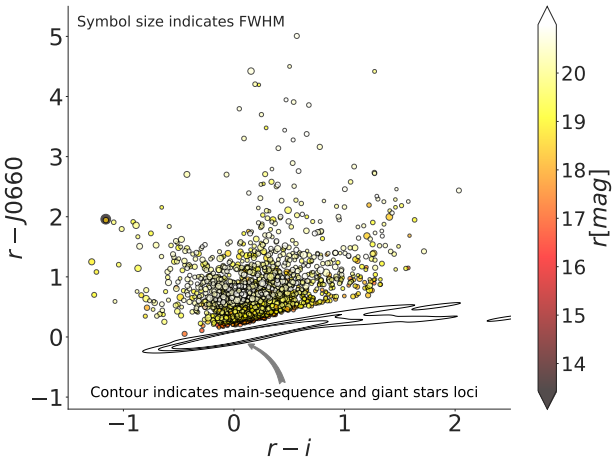


Figure 4. Colour-colour diagram with all the emission-line objects selected from the S-PLUS DR3. The size of the symbols represent the measured FWHM assuming a Gaussian core (for more detail see Almeida-Fernandes et al. 2022). Coloured bar indicates the magnitude values of the r-band. The contours represent the S-PLUS synthetic photometry of main-sequence and giant stars loci from the library of stellar spectral energy distributions of Pickles (1998).

agglomerative clustering (HAC; the one used in this work), which follows "bottom-up" approach, and the *hierarchical divisive clustering* that follows "top-down" approach. The HAC consists of building a binary merge tree, starting from each data element stored at the leaves (interpreted as individual clusters) and proceeds by merging two by two the "closest" sub-sets (stored at nodes) until the root – unique cluster – of the tree that contains all the elements of the data set is reached. The term "agglomerative" is used to point out that data elements are successively agglomerated into higher-levels. In each iteration, two "nearby" clusters are

collapsed into a new, more populated group (Mann & Kaur 2013; Aggarwal 2015). Hence, each step reduces the number of clusters. The procedure may be summarized in three steps:

(i) Initially, each data element represents one cluster, i.e. "leaves of the tree". This means that at the beginning, the total number of clusters/leaves is equal to the number of the elements in the data set.
(ii) Through a looping process, the clusters are merged into new ones that are described by the maximum similarity among them.
(iii) Finally, all the clusters belong to an unique cluster, "the root of the tree" structure.

On the other hand, the *hierarchical divisive clustering* algorithm follows a "top-down" approach. This means that the clustering starts from data element from only one cluster and then moves down recursively in the hierarchy to smaller groups. In simple words, hierarchical (agglomerative and divisive) clustering algorithms intend to gather similar objects into groups called clusters in the space of parameters which is investigated.

2.3.2 *Hierarchical density-based cluster selection*

Hierarchical density-based cluster selection (hereafter HDBSCAN; Campello et al. 2013) is another unsupervised machine learning algorithm that relies on clustering. It is based on a slightly modified version of density-based spatial clustering of applications (DBSCAN; Ester et al. 1996) which declares data points as noise. It assumes that clusters are characterized by "islands" of high density in the sea of the parameter space. HDBSCAN takes the DBSCAN concept forward by introducing a hierarchy to the clustering, with "persistent" clusters finally extracted from the hierarchical tree. The main advantage of HDBSCAN in comparison with its predecessor consists in the possibility of finding clusters of variable densities and different shapes. Following Malzer & Baum (2021) and Ntwaetsile & Geach (2021) it works as follows:

(i) HDBSCAN defines the "core" distance for a data point x, $\mathrm{core}_k(x)$, as the distance of an object to its kth nearest neighbour. This mean that lower values of $\mathrm{core}_k(x)$ represent higher densities and vice-versa.
(ii) The "mutual readability distance" between two points a and b is defined as $d_m(a,b) = \min\{\mathrm{core}_k(a), \mathrm{core}_k(b), d(a,b)\}$, where $d(a,b)$ is the distance between a and b according, for instance, to Euclidean metric. The mutual readability distance allows data points in dense regions to stay close together and those that are in less dense regions to move away.
(iii) The mutual readability plot is used to construct the minimum spanning tree, and sorting its edges by the mutual readability distance resulting in a hierarchical tree structure. The hierarchy of connected components is defined by sorting the edges of the tree by distance

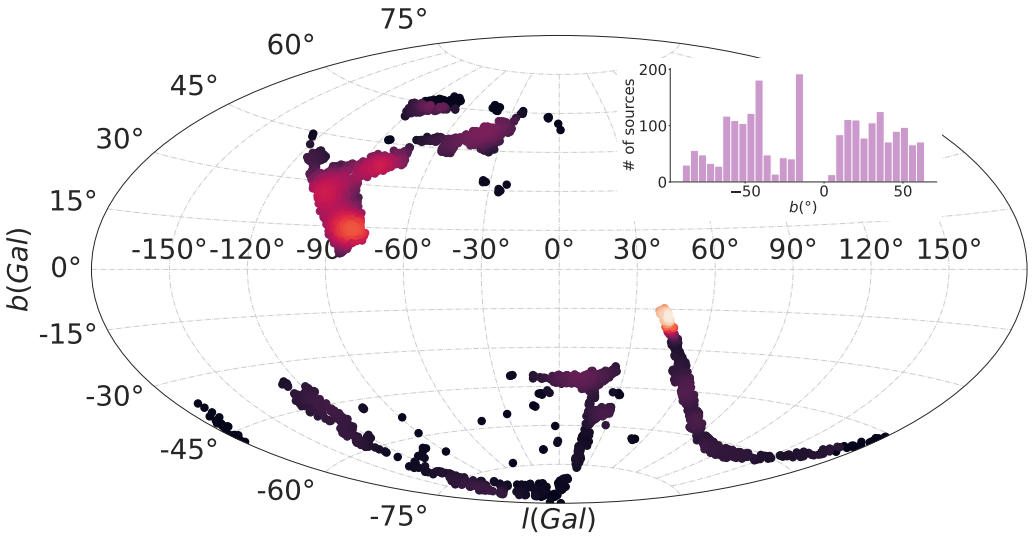


Figure 5. Distribution of the emission-line objects in galactic longitude and latitude coordinates. The inset figure represents the distribution of the objects in galactic latitude.

in reverse order, describing a dendrogram (the diagram explained in 2.3.1). This is the structure from which the cluster will be identified.

(iv) `HDBSCAN` allows extracting clusters of variable density by cutting the dendrogram at different levels of grouping.

(v) The cluster tree is condensed into a simpler structure (see, for instance, Figure A1 of Appendix A). Considering the single main trunk which contains all the data points, the tree splits into branches. A condensed cluster hierarchy can be described by considering the number of points that are kept in each branch as it splits. It is important to mention that there is a key parameter called minimum cluster size. If a given branch splits into two, with one branch containing fewer points than the minimum cluster size, the larger branch "persists" and the smaller split branch "falls out" of the cluster. If a branch splits into two with both branches exceeding the minimum cluster size, both new branches are preserved.

(vi) The clusters are extracted on the notion of persistence in the hierarchy. The parameter $\lambda = d_{\rm m}^{-1}$ is defined, and each cluster has a $\lambda_{\rm birth}$ (the point at which the cluster split off) and $\lambda_{\rm death}$ (the point when the cluster split into other clusters). In each cluster, we have λ_p describing when each point fell out of the cluster (or was split off into a new cluster), so that $\lambda_{\rm birth} \leq \lambda_p \leq \lambda_{\rm death}$. Cluster stability S is defined as the sum of $\lambda_p - \lambda_{\rm birth}$ for all points in the cluster. To extract clusters, the following procedure is implemented. First, each leaf constitutes a cluster. Then, moving through the hierarchy, it is considered the stability of a parent cluster S_p and its n descendants $S_d^{0,1,2,\ldots,n}$: if $S_p > \sum_{i=0}^{n} S_d^i$, we unselect all the descendants; otherwise, the cluster stability is set as $S_p = \sum_{i=0}^{n} S_d^i$. At the root node, we have our set of the selected clusters. Any data point in the sample that does not fall into one of the selected clusters is defined as noise.

(vii) By adopting the **soft clustering** or **the fuzzy clustering** technique it is possible to mitigate the need to establish or define the cluster membership limit. In fact, each source has a finite probability of belonging to every selected cluster. In this approach, all points (including noises) are not assigned to a cluster label, but are instead assigned to a vector of probabilities whose length is equal to the number of clusters found. Such an approach aims to solve the problem of noise classification.

The HDBSCAN algorithm starts off much the same as DBSCAN: transforming the space according to density, and perform single linkage clustering on the transformed space. Instead of taking an epsilon[5] value as a cut level for the dendrogram, a different approach is followed: the dendrogram is condensed by viewing splits that result in a small number of points splitting off as points "falling out of a cluster". This results in a smaller tree with fewer clusters that "lose points". That tree can then be used to select the most stable or persistent clusters. This process allows the tree to be cut at varying height, picking our varying density clusters based on cluster stability. The immediate advantage of this is that we can have varying density clusters; the second benefit is that we have eliminated the epsilon parameter as we no longer need it to choose a cut of the dendrogram. Instead we have a new parameter `min_cluster_size` which is used to determine whether points are "falling out of a cluster"or splitting to form two new clusters.

Over the last few years, `HDBSCAN` has been used for different tasks in Astronomy. `HDBSCAN` was used to identify IR bubbles from Spitzer images (Jayasinghe et al. 2019). Webb et al. (2020) implemented `HDBSCAN` for discovering transients. Logan & Fotopoulou (2020) presented `HDBSCAN` as a viable tool to separate stars, galaxies and QSOs using photometric data. Recently, Ntwaetsile & Geach (2021) employed `HDBSCAN` to group radio sources into a sequence of morphological classes, illustrating a simple methodology to classify and label new, unseen galaxies in large samples. This approach was also implemented to identify stellar groups in Canis Major OB1 (Santos-Silva et al. 2021).

[5] Epsilon parameter in DBSCAN represents the maximum distance between two samples for one to be considered as in the neighborhood of the other. This is the most important DBSCAN parameter to choose appropriately for the data set and distance function.

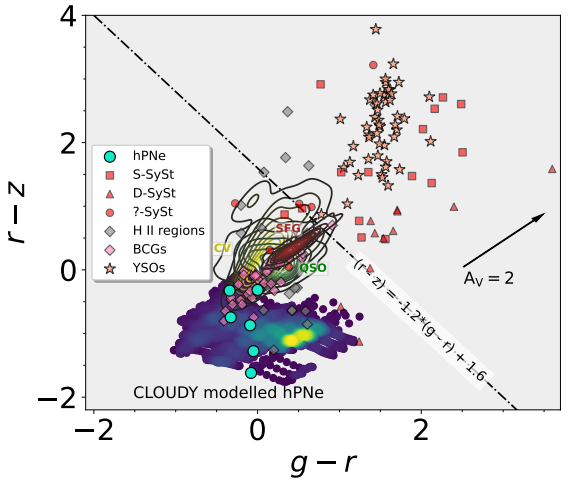


Figure 6. The $(r-z)$ versus $(g-r)$ synthetic colour-colour diagram of several classes of emission lines objects. Included in the diagrams, there are families of CLOUDY modelled halo PNe spanning a range of properties (density map region). Cyan circles represent S-PLUS photometry from observed spectra of PNe. Grey diamonds represent H II regions in NGC 55. Red boxes and triangles display S- and D-type symbiotic stars, respectively. Red circles are SySt with no associated type. This group includes Galactic and external SySt from NGC 205 IC 10 and NGC 185. Yellow contours correspond to CVs from SDSS. Pink diamonds indicate blue compact galaxies (BCGs) from SDSS. Brown contours refer to SDSS star-forming galaxies (SDSS SFGs). SDSS QSOs at different redshift ranges are shown as green contours, and YSOs from Lupus and Sigma Orionis are represented by salmon stars. The diagonal dashed line represents a subjective criterion to separate the objects into two colour types. The arrow indicates the reddening vector with $A_V \simeq 2$ mag.

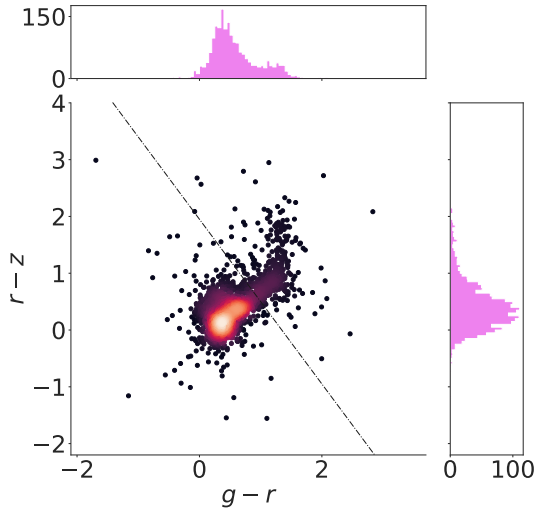


Figure 7. The $(r-z)$ versus $(g-r)$ colour-colour diagram with all the emission line objects selected in S-PLUS. The side and upper figures represent the $(r-z)$ and $(g-r)$ colour distributions, respectively.

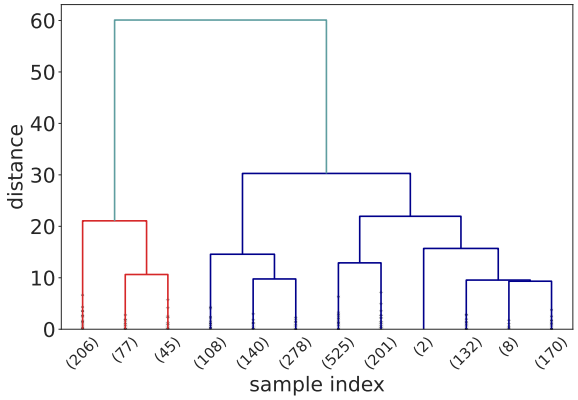


Figure 8. Truncated dendrogram of complete-linkage hierarchical clustered based on $(r-z)$ and $(g-r)$ colours. The cluster sizes are exposed in the brackets for the 12 truncated clusters.

2.4 Splitting the Hα emitters into blue and red populations

To select the blue and red populations in the sample of Hα emitters, we first looked for the best colour-colour diagram by using the S-PLUS synthetic photometry of several classes of emission line objects. The $(r-z)$ versus $(g-r)$ diagram is displayed in Fig. 6. SySt and YSOs span a range on $(g-r)$, from 0.4 to 3.6 at the upper right. On the other hand, PNe, H II regions, CVs, QSOs, and emission line galaxies are located on the lower-left region of the diagram. The dashed line in Fig. 6 separates the blue and red zones.

Fig. 7 displays the $(r-z)$ versus $(g-r)$ diagram from the list of Hα emitters in S-PLUS. Obviously only such emitters with detection in the g and z filters are considered for this colour classification by making a cut in the magnitude errors at 0.2, totalizing 2,892 objects. A bi-modal distribution is found for both colour indices (see side and top plots of the Fig. 7.

The two peaks on the $(g-r)$ and $(r-z)$ distributions have immediate correspondence with the blue and red zones pointed out from the synthetic diagram (Fig. 6). We can also see that the fraction of blue objects is considerably larger than that of the red ones.

2.4.1 Application of hierarchical agglomerative cluster

The ideal way to choose the number of clusters is by displaying the dendrogram diagram. Firstly, the hierarchical cluster output dendrogram can be used to obtain the desired clustering. Secondly, the dendrogram allows a convenient way to establish the entity-relationship at all levels of granularity. Fig. 8 illustrates the dendrogram based on the $(g-r)$ and $(r-z)$ colours of Hα emitters, and it highlights the order and distances of the groups in the hierarchical clustering, stopping at 12 nodes:

- The x-axis specifies the population in the nodes in a given level of grouping – that summed up correspond to the total number of elements under investigation.
- The y-axis represents the "distance", which is a measurement of the closeness of the clusters or data points in different levels of clustering.

Reading the diagram from the top to the bottom, we see that all systems are divided after the very first level from the top already into (only) two groups: as expected, they correspond to the red and blue populations of Hα emitters presented in Fig. 7 as well as shown in Fig. 9. From that point on, the groups were subdivided without evident distinction, and truncation was thus assumed when the 12-node level was reached. The truncation is an usual procedure when dealing with big data.

In this work, `HAC` was employed by using the library

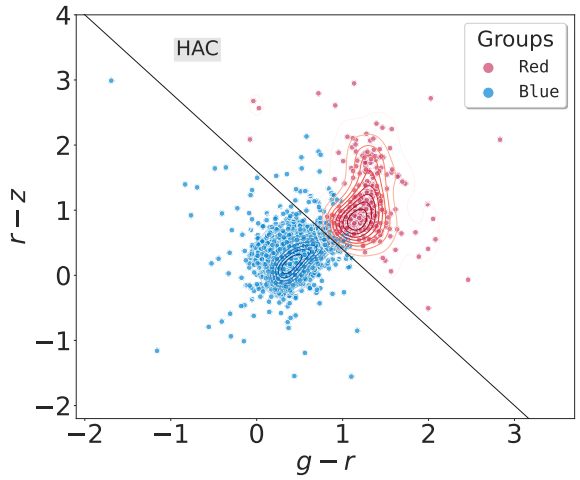


Figure 9. The $(r-z)$ versus $(g-r)$ colour-colour diagram with the two population found by implementing HAC algorithm. The blue and red symbols represent the sources with intense blue continuum and those with intense red continuum, respectively. The straight line is the same line in Fig. 6.

`Scikit-learn`[6] (Pedregosa et al. 2011). Then, the DENDROGRAM() function, which is included in the package `scipy`[7], was used in conjunctions with the task *Dendrogram Truncation* to generate the (truncated) dendrogram. The following parameters must be taken into account when the algorithm is applied to data: `n_clusters`, `Affinity`, and `Linkage`. The parameter `n_clusters` defines the number of clusters expected by the user. Given that our goal is to divide our sample into two groups, `n_clusters` is set to "2". `Affinity` determines the "metric "that compute the linkage. We have found that a simplistic metric, the "Euclidean", is effective for our purpose.

`Linkage` determines which distance to use between sets of observations. Linkage defines how the similarity between two clusters is calculated, by determining the distance between sets of observations as a function of the pairwise distances between elements. The algorithm merges the pairs of clusters that minimize this criterion. `Ward's method` minimizes the variance of the clusters that are being merged (Ward 1963). To implement this method, find the pair of clusters that leads to a minimum increase in total within-cluster variance after merging. Ward procedure uses the error sum of squares to measure this variance. The two clusters with the smallest error sum of squares will eventually form a new cluster.

At this point, our list of Hα emitters is divided into two populations based on their integrated flux, with a blue population (1,564 objects) larger than the red one (328).

2.4.2 *Application of HDBSCAN*

For the sake of comparison with the results from HAC, we also used HDBSCAN to distinguish the blue and the red sources.

The main difference between these two algorithms is that HDBSCAN is more conservative in the sense that several data points are classified as noise. For this task, the Python implementation of HDBSCAN[8] (McInnes et al. 2017) was adopted.

Similarly to HAC, there are some key parameters that should be considered when the algorithm is applied. Regarding the metric, the "Euclidian" one is assumed.

The two most critical parameters are the "minimum cluster size" and "minimum number of samples". The former refers to the smallest size of a group that it is considered as a cluster. The value of 60 has been adopted for the "minimum cluster size". The "minimum number of samples" provides a measure of how conservative our clustering method will be, expressed as the fraction of data classified as noise, and the value of 15 was adopted. With this model configuration, two clusters were identified.

Left panel of Fig. 10 shows the two clusters found with HDBSCAN. One cluster contains 1,413 blue sources and the other 131 red sources. The number of objects classified as noise is 348. This result is overall consistent with those obtained with HAC, although more restrictive when classifying members of the red group. The two main clusters obtained with HDBSCAN are located in the same region in the $(r-z)$ versus $(g-r)$ diagram as those groups found using the HAC. About 98% of the blue sources selected by HDBSCAN are in the list of blue objects identified by HAC. All the red sources selected by HDBSCAN were also classified by HAC as red objects. In fact, by applying the `condensed_tree_` to the data colours two clusters, are selected. The `condensed_tree_` attribute is the equivalent dendrogram plot for HDBSCAN which displays the cluster tree mentioned in the section 2.3.2. (see Appendix A for more details about `condensed_tree_` attribute).

2.4.3 *Soft clustering for HDBSCAN*

The main disadvantage of HDBSCAN is that several sources are labelled as "noise", so that they are not assigned to any cluster. As mentioned earlier, this comes from the conservative nature of HDBSCAN and the fact that these data points (data noise) are located far away of the clusters' cores. An alternative way to avoid outliers (data noise) classifications is the implementation of the "soft clustering" (see Section 2.3.2). Soft clustering from HDBSCAN[9] was used to assign every object to a cluster that they most likely belong to. According to this approach, data points are not assigned in a deterministic way to a cluster but to a vector of probabilities as a measure of belonging to different clusters: the probability value at the ith entry of the vector is the probability that a data point is a member of the ith cluster. We can, then, simply assign cluster labels for every data point by taking the most likely cluster it belongs to, using probability thresholds. Therefore, soft clustering for HDBSCAN is achieved through an outlier score modification to consider how distant an outlier is from each cluster, which is based on the Global-Local Outlier Score from Hierarchies (GLOSH) algorithm (Campello et al. 2015). This is combined with a measurement of distance from a given cluster to estimate the probability that a given data point belongs to any of the fixed groups drawn from the condensed tree.

The right panel of Fig. 10 shows which cluster the data points classified as the noise by HDBSCAN belong to. Blue and red points indicate those with the highest probability of being in the blue and red groups, respectively. This procedure fills out the clusters nicely. There were many noise points that most likely belong to the expected clusters in very good agreement with the results obtained from HAC. Indeed, our separation of the Hα emitters into blue and red sources

[6] https://scikit-learn.org/stable/
[7] https://www.scipy.org/
[8] https://hdbscan.readthedocs.io/en/latest/

[9] https://hdbscan.readthedocs.io/en/latest/soft_clustering_explanation.html

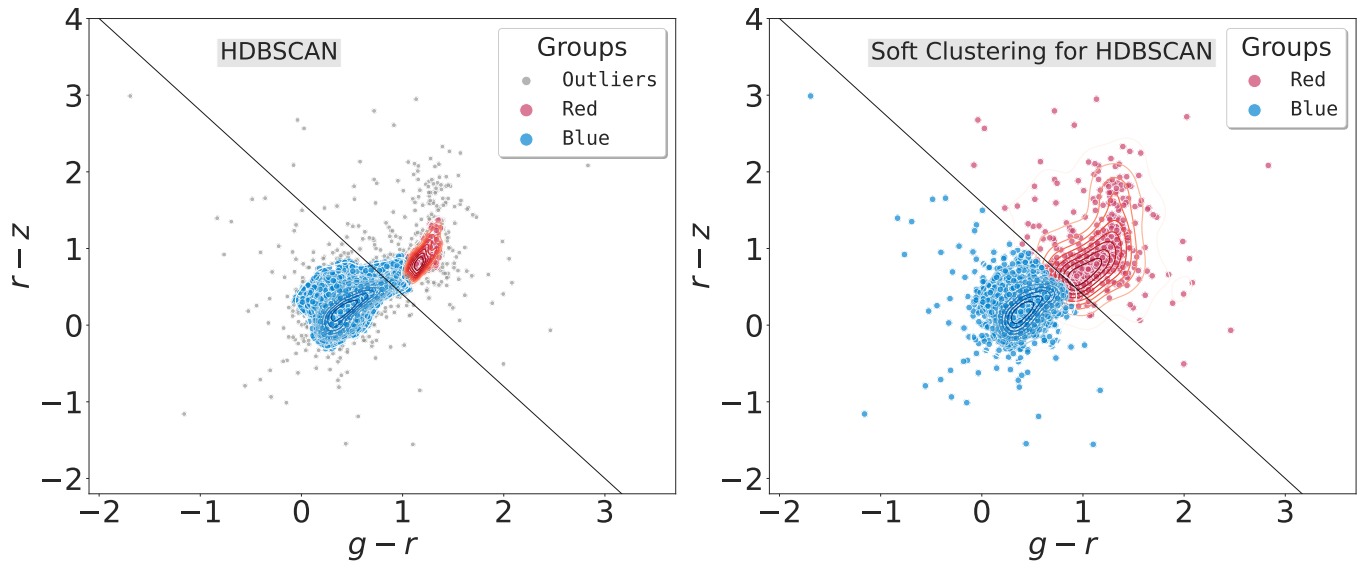


Figure 10. *Left panel*: as Figure 9 but with the results after to apply HDBSCAN to the sample of emission-line sources. Blue and red symbols correspond to blue and red sources, respectively, with gray symbols representing sources classified as noise by HDBSCAN. The straight line is the same line as Figures 6 and 9. *Right panel*: results after apply a soft clustering to the HDBSCAN results.

has been improved. Instead of forcing the algorithm to make a decision to which group a data point belongs to, as HAC does, we have quantified the likelihood of a given observation to belong to any of the two clusters found in our data set (see, for instance, the two last columns of Table B1).

3 RESULTS AND DISCUSSION

Our strategy is focused on the identification of Hα emitters in the S-PLUS footprint, exploiting the unique filter system of the survey returned 2,187 objects with excess in the $J0660$ band. The fractional contribution of different classes of Hα emitters to the overall sample was evaluated by cross-matching the objects' list with the SIMBAD database[10]. Optical spectra available in the SDSS DR16 (Ahumada et al. 2020) and in the Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST DR7; 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 ($d_{max,proj}$).

3.1 Matches with SIMBAD sources

We found 393 positive matches between our catalog of Hα emitters and SIMBAD database considering a radius of $d_{max,proj} = 2$ arcsec. The results are described below and are listed in Table 1.

3.1.1 Ionized nebulae

As it was mentioned, several classes of objects with diffuse appearance and/or nebular lines in our Galaxy as well as in nearby galaxies are listed in our sample, most being PNe.

Planetary nebulae represent the final stages of low- and intermediate-mass stars from which the material has been previously ejected in the phases of AGB and post-AGB and is ionized by the high energetic radiation from a hot stellar remnant core. In our list of Hα emitters one compact PN is catalogued in SIMBAD. This object was classified as a blue source. The S-spectrum and spectrum from Parker et al. (2016) of the PN G006.0-41.9 are displayed on panel (*a*) of Fig 12. Emission lines like Hα and [N II] are clearly visible in its spectra. PN G006.0-41.9 belongs to the group of Galactic halo PNe. This rare group of PNe are of particular interest because they are characterized by low metallicity and present large velocities.

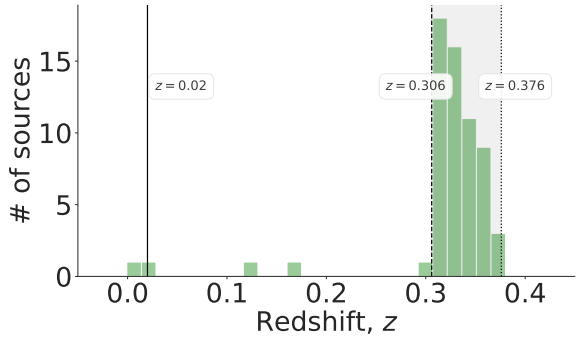


Figure 11. Histogram of redshift for the galaxies with SIMBAD coincidence. The black vertical continuous line indicates the threshold value ($z \sim 0.02$) on which the Hα emission-line is detected in $J0660$ filter. The black vertical dashed and dotted lines represent the accumulative redshift range where the Hβ and [O III] 4959, 5007 Å emission lines are detected on the $J0660$-band.

3.1.2 Supernovae

Supernovae also display emission line spectra from evolved stars through multiple channels. However, the energy-input mechanism is quite different in each case. One confirmed SN is in our list. This SN is classified as a blue source by HAC but the probability of being blue and red of 0.16 and 0.29, respectively, according to HDBSCAN.

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

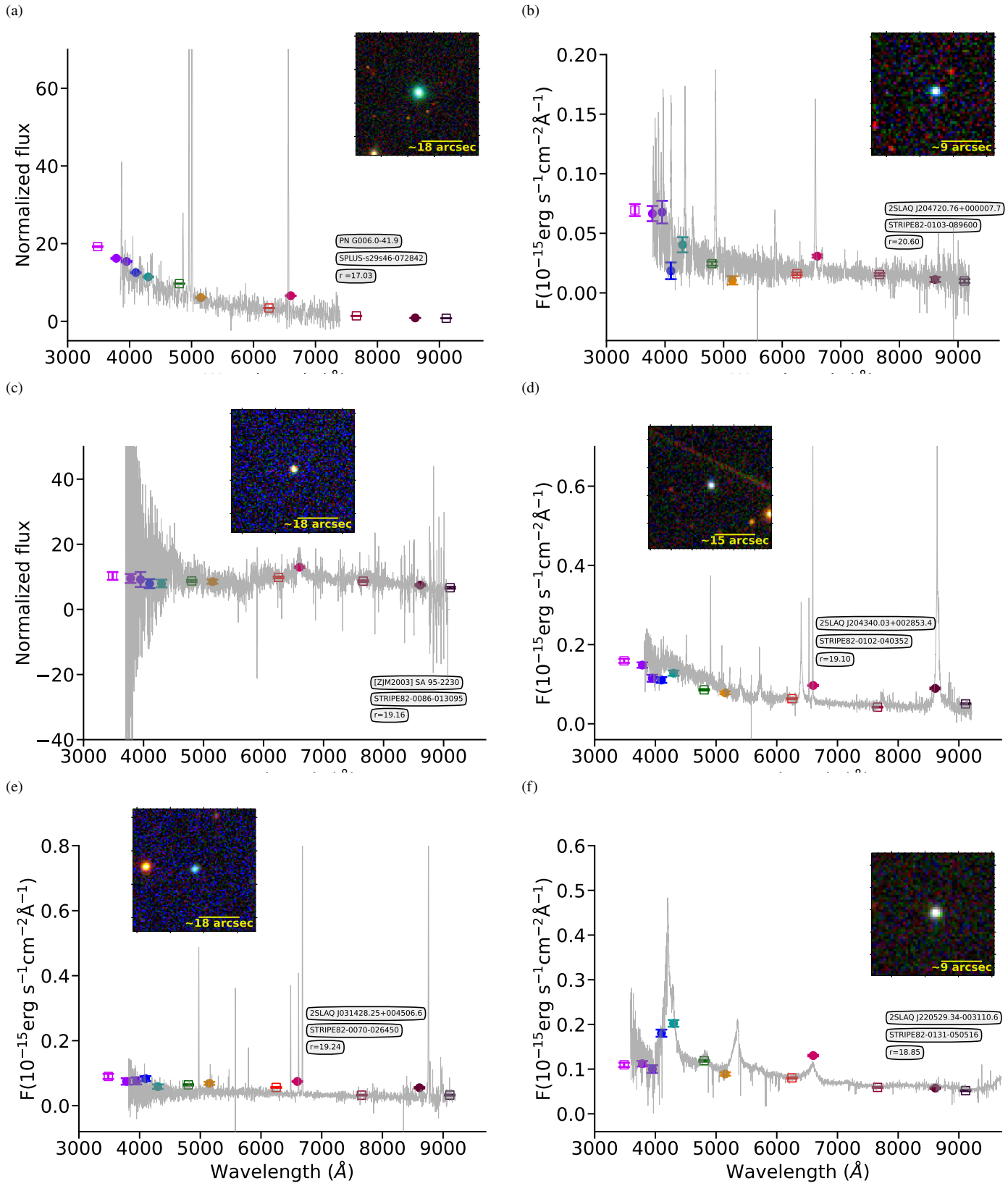


Figure 12. Summary of our selection results showing the spectrum (gray line en each panel) of different classes of emission line sources identified in our target list. A spectrum of a PN (*a*) from Parker et al. (2016). The SDSS spectra of a nova (*b*) and a cataclysmic variable star (*c*). SDSS spectra of a Seyfert 1 (*d*) and an emission line galaxy (*e*) with $z = 0.317$ and $z = 0.334$, respectively. For this two objects, the excess on the $J0660$ is due to the [O III] 4959, 5007 Å emission lines, and a QSO (*f*) with redshift of ~ 2.45, meaning that excess on the narrow-band filter is produced by the C III] emission line. As in Figure 3, coloured square and circles symbols represent the S-PLUS photometry. All these objects show a significance excesses on the $J0660$ filter in comparison with the broad-bands.

Table 1. A summary of the results obtained of the positional cross-match between between the S-PLUS list of emission line objects and the SIMBAD database. We used a search radius of 2 arcsec. SIMBAD categories of objects are listed in the first column. The numbers of objects of each SIMBAD category are exposed in the third column.

Main type	Associated SIMBAD types	Number of S-PLUS objects with SIMBAD match
Nebulae	PN	1
Spernovae	SN	1
Stellar binary system	CataclyV*, CV*_Candidate, EB*, EB*_Candidate	40
Star	star, WD*, WD*_Candidate, Blue, PM*, low-mass*	32
Variable star	RRLyr, Candidate_RRLyr, V*	7
Galaxy	EmG, Galaxy	42
QSO	QSO, QSO_Candidate	229
AGN	AGN, AGN_Candidate, Seyfert_1	31
Other type	UV, X, Unknown_Candidate, Radio(cm), Radio, EmObj, GlCl	10
Total		393

3.1.3 *Interacting binary systems*

Following the classification available in SIMBAD, 27 known and 6 CV candidates were found by our analysis. CVs are interacting binary systems of very short orbital period, in which a low-mass, early-type star fills its Roche lobe and transfers mass to a white dwarf companion (Patterson 1984). For the sake of illustration, Fig. 12 (panel *b*) shows the S-PLUS photometry overlapped to the SDSS spectrum of 2SLAQ J204720.76+000007.7, a CV, which we correctly classified as a blue source. In fact, 31 of theses CV are classified as blue sources by HAC. Eclipsing binaries are also listed on our catalog.

3.1.4 *Stars*

Table 1 shows that 32 objects in our sample of Hα emitters are categorized, by SIMBAD, as: normal stars, white dwarf (WD*), white dwarf candidates (Candidate_WD*), blue stars, high proper-motion stars (PM*), variable stars of RR Lyr type and, low-mass star (low-mass*; M<1M$_\odot$). Panel *c* of Fig. 12 shows the LAMOST and S-spectrum, as well as the coloured image of the source [ZJM2003] SA 95-2230 catalogued as star in SIMBAD. Its spectrum exhibits an emission line at the wavelength range covered by the *J*0660 filter. The spectroscopic redshift of this object is ∼1.36, indicating that the excess on the narrow-band filter is caused by the Mg II 2798 Å emission line. This object could be a QSO misclassified in SIMBAD, which has been classified as a blue source for both HAC and HDBSCAN.

3.1.5 *Galaxies*

Galaxies are also included in our catalog: emission-line galaxies (EmG), Seyfert types-1 and other type of AGNs.

Fig. 11 shows the redshift distribution of the galaxies in our sample that have SIMBAD correspondents. Of the 73 galaxies with SIMBAD counterparts, 65 objects have redshift measurements of which only 2 objects of them have small redshift values ($z < 0.02$) showing that the emission detected in the *J*0660 filters is associated with the real Hα λ6563 emission line. On the other hand, Fig. 11 also shows that there is an increment of galaxies with redshift between ∼0.31 and ∼0.38. This particular population of seemingly Hα emitters is represented by AGN, Seyfert 1 galaxies and other emission line galaxies. In fact, at the accumulative redshift range, $0.306 < z < 0.376$, represented by the filled area in the figure, with Hβ and [O III] 4959, 5007Å lines redshifted into the *J*0660 filter.

Panel *d* of Fig. 12 shows the S-PLUS photometry and the SDSS spectrum of the Seyfert type-1 galaxy 2SLAQ J204340.03+002853.4. The spectrum clearly exhibits strong emission lines. The redshift of this object is around 0.317 indicating that the excess on the *J*0660 is due to the [O III] 4959, 5007 Å emission lines. Panel (*e*) exhibits the S-PLUS photometry and the SDSS spectrum of the galaxy 2SLAQ J031428.25+004506.6. This source has $z \sim 0.334$ indicating, as the Seyfert galaxy above, that the excess in the *J*0660 filter is not due to the Hα emission line but to another redshifted line. Note that 88% of these galaxies are classified as blue objects by HAC and HDBSCAN, with 5% of the galaxies being found in the population of red sources. It is important to mention that this sample consists of the compact objects only, and the fraction of extended emission line galaxies that could be found in this survey with similar methodology is much larger than presented here.

3.1.6 *QSOs: redshifted lines mimicking the Hα emission*

Following the classification available in the literature, about 13% of blue Hα emitters sources in our sample are found to be QSOs. We have to point out here that the excess in the *J*0660 filter for QSOs is attributed to redshifted lines that fall in the wavelength range covered by that filter depending on the redshift of the QSOs – e.g., Hβ, Mg II 2798 Å, C III] 1909 Å and C IV 1550 Å (see Gutiérrez-Soto et al. 2020 and Nakazono et al. 2021). QSO 2SLAQ J220529.34-003110.6, shown in Panel *f* of Fig. 12, is an example of a QSO at redshift ∼ 2.45, for which the C III] line falls at the range covered by the *J*0660 filter.

3.1.7 *Other classifications of objects*

As it can be seen in Table 1, our sample also gathers a variety of objects without any previous classification. They may also be X- and UV-sources, radio, globular cluster (GLCL), among others, indicating the richness of the sample in nature and in physical properties.

3.2 SDSS and LAMOST: a spectroscopic validation

Finally, we also cross-matched out a sample of Hα emitters in the S-PLUS with the SDSS DR16 (Ahumada et al. 2020). For doing this, we adopted a 2 arcsec as the cross-matching radius. In the case of the cross-match with LAMOST (Wu et al. 2011), the same radius was considered, and we ended up with 218 sources belonging to both

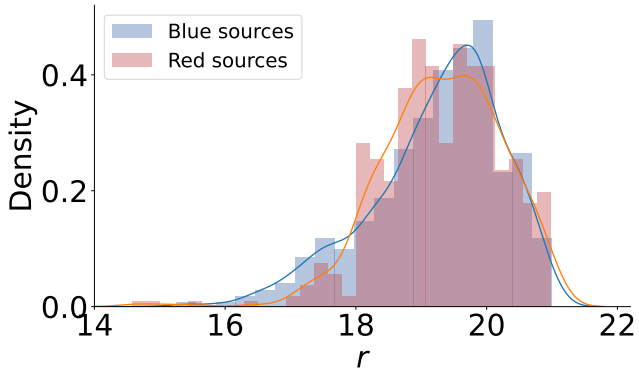


Figure 13. Distribution of r magnitude for the blue and red sources of the sample of Hα emitters. The histogram heights show density normalization scales. The smooth curves represent a Kernel density estimation for both samples. The classification of the two group was performed by implemented the soft clustering technique.

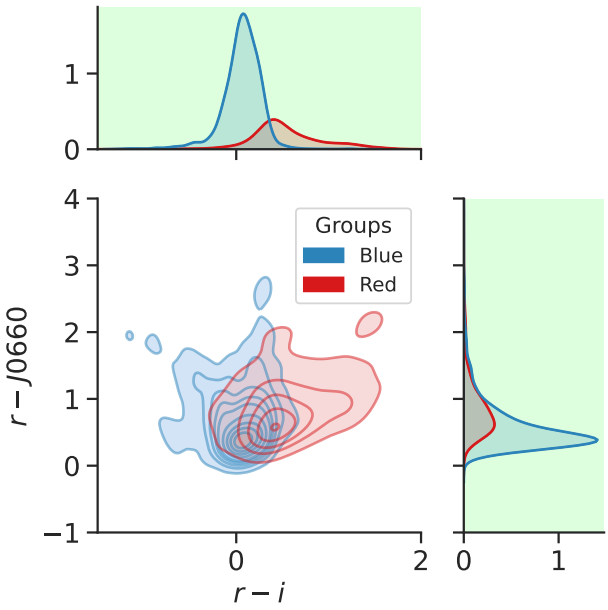


Figure 14. $(r - J0660)$ versus $(r - i)$ colour-colour diagram for the blue (blue contours) and red sources (red contours) of the sample of Hα emitters. The side and upper panels show the respective colours distributions. The two groups are the results of applying the soft clustering approach.

catalogues and approximately 95% of them display emission lines spectra.

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

The spectra from both SDSS and LAMOST provide a good validation to our approach, clearly showing that the methodology is actually effective for selecting sources with emission lines. Figs. C1 and C2 of Appendix C show examples of SDSS and LAMOST spectra, respectively, of some objects classified as emitters in S-PLUS. The analysis of individual sources will be studied in the future.

3.3 Magnitudes and colour distributions

In Fig. 13, we demonstrate the distribution of the blue and red population of S-PLUS Hα emitters in terms of their r-magnitude. Both, blue and red sources can be as bright as $r = 16$ mag, while they show a peak at ∼19.7 mag. The red group exhibits another peak at ∼19.1 mag. Both groups present other peaks at ∼17.5.

Fig. 14 displays the $(r - J0660)$ versus $(r - i)$ diagram with the position of the two populations and the distributions of the blue and red Hα sources in the $(r - i)$ (top histogram) and $(r - J0660)$ (side) colours. The $(r - i)$ colour distribution of the blue and red sources have peaks at distinct values of 0.08 and 0.42, respectively. This result is consistent with that obtained from Wevers et al. (2017) who also used the $(r - i)$ colour index to select blue outliers from the Galactic Bulge Survey (GBS; Jonker et al. 2011). Finally, the $(r - J0660)$ colour index distribution of the blue and red objects peaks at 0.37 and 0.60, respectively. This result implies a stronger Hα emission in the red sources compared to the blue ones.

4 CONCLUSIONS

Here we exploited the capability of the S-PLUS project (Mendes de Oliveira et al. 2019) to survey Hα emitters in the Southern Sky following a three-steps approach: identify Hα emitters, distinguish the blue and red populations as a first diagnostic about the nature of the sources, and validate the results through spectroscopic measurements.

The Hα emitters were identified by employing the $J0660$ narrow-filter and r and i broad-filters available in the S-PLUS project. The $(r - J0660)$ versus $(r - i)$ colour-colour diagram was used to define the loci of the main-sequence and giant stars and disentangle objects with probability to being point sources in the local Universe with an Hα-excess $(r - J0660 > 0)$ (see Fig. 4). 2,187 sources matched this criterion, with 292 of them claimed in the literature (`SIMBAD`) as QSOs and non-local galaxies, and therefore being false positive identifications of Hα emission (see sections 3.1.5 and 3.1.6).

The $(r - z)$ and $(g - r)$ colour distributions of the Hα emitters were found to be bimodal, indicating the presence of two distinct populations of bluer and redder sources with a narrow overlapping zone (Fig. 7). Two algorithms of unsupervised machine learning classification were used to distinguish the two populations: the `HAC` and the `HDBSCAN` clustering algorithms. Both algorithms ended up to very similar clusters, on the $(r - z)$ and $(g - r)$ colour indices space.

Given that `HDBSCAN` is considered as a conservative algorithm, many objects were labeled as noise data points while they did not by the `HAC` algorithm (Section 2.4.3). To overcome this problem, a so-called "soft" clustering approach for `HDBSCAN` was employed and the probabilities of each data point to belong to the "blue" and "red" subgroup were computed. The results from the `HAC` and `HDBSCAN` algorithms are mutually consistent and, therefore, we reckon that the $(r - z)$ and $(g - r)$ colours are ideal for separating objects into the bluer and redder populations and correlate their colour to the nature of the sources. In particular, the bluer objects were found to be mainly CVs, PNe, dwarf compact galaxies, galaxies with redshifted spectra and QSOs, among others, while the redder sources are early type galaxies with emission lines, probably young/active late-type stars

or even symbiotic stars (in fact, evolved binary systems hosting a red giant star).

We also cross-matched our catalogue of Hα emitters with spectroscopic databases (SDSS and LAMOST; see Section 3.2). This exercise demonstrated that at least 95% of the objects with available spectroscopic information are genuine emission line sources, validating our approach to identify emission line objects in the S-PLUS project. The spectroscopic sample of Hα emitters lists 21 sources of the local Universe (with $z < 0.02$) indicating that the emission on the *J*0660 filter corresponds to the Hα line, 197 sources with redshift larger than 0.02, indicating that they are very likely QSOs and AGN and/or non-local galaxies on which the excess of the *J*0660 filter is due to Hβ, Mg II 2798 Å, C III] 1909 Åand C IV 1550 Å emission lines for the case of QSOs and Hβ and [O III] 4959, 5007 Å emission lines for galaxies, those depending on their redshift.

Finally, we make our catalogue of Hα-emitter candidates selected using the S-PLUS 12-band photometric system for it to be explored by the community. The sources presented here are good targets for follow-up observations to expand and to further explore their nature and physical properties.

ACKNOWLEDGEMENTS

LAG-S acknowledges funding for this work from 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 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 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[11] (Taylor 2005), simbad and vizier from Strasbourg Astronomical Data Center (CDS)[12] and the following python packages: numpy, astropy, matplotlib, seaborn, scikit-learn.

DATA AVAILABILITY

The S-PLUS DR3 data are publicly available through the S-PLUS service (`https://splus.cloud/`). The `SIMBAD` data can be obtained via `https://simbad.u-strasbg.fr/simbad/`. The SDSS DR16 catalogue can be obtained at `https://www.sdss.org/dr16`. The LAMOST DR7 catalogue can be obtained at `https://dr7.lamost.org/catalogue`. The catalogue of Hα-emitter candidates will be made available at Vizier.

REFERENCES

Aggarwal C. C., 2015, Data Mining: The Textbook. Springer, Cham, doi:10.1007/978-3-319-14142-8
Ahumada R., et al., 2020, ApJS, 249, 3
Akras S., Guzman-Ramirez L., Leal-Ferreira M. L., Ramos-Larios G., 2019a, ApJS, 240, 21
Akras S., Leal-Ferreira M. L., Guzman-Ramirez L., Ramos-Larios G., 2019b, MNRAS, 483, 5077
Akras S., Guzman-Ramirez L., Gonçalves D. R., 2019c, MNRAS, 488, 3238
Almeida-Fernandes F., et al., 2022, MNRAS, 511, 4590
Barentsen G., et al., 2014, MNRAS, 444, 3230
Benitez N., et al., 2014, arXiv e-prints, p. arXiv:1403.5237
Bonoli S., et al., 2021, A&A, 653, A31
Campello R. J. G. B., Moulavi D., Sander J., 2013, in Pei J., Tseng V. S., Cao L., Motoda H., Xu G., eds, Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 160–172
Campello R., Moulavi D., Zimek A., Sander J., 2015, A C M Transactions on Knowledge Discovery from Data, 10, 1
Cenarro A. J., et al., 2019, A&A, 622, A176
Corradi R. L. M., Giammanco C., 2010, A&A, 520, A99
Corradi R. L. M., et al., 2008, A&A, 480, 409

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

Corradi R. L. M., Sabin L., Munari U., Cetrulo G., Englaro A., Angeloni R., Greimel R., Mampaso A., 2011, A&A, 529, A56
Davies R. D., Elliott K. H., Meaburn J., 1976, Mem. RAS, 81, 89
Drew J. E., et al., 2005, MNRAS, 362, 753
Drew J. E., Greimel R., Irwin M. J., Sale S. E., 2008, MNRAS, 386, 1761
Drew J. E., et al., 2014, MNRAS, 440, 2036
Ester M., Kriegel H.-P., Sander J., Xu X., 1996, in Proc. of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96). pp 226–231
Frew D. J., 2008, PhD thesis, Department of Physics, Macquarie University, NSW 2109, Australia
Fukugita M., Ichikawa T., Gunn J. E., Doi M., Shimasaku K., Schneider D. P., 1996, AJ, 111, 1748
Gutiérrez-Soto L. A., et al., 2020, A&A, 633, A123
Jacoby G. H., et al., 2010, Publ. Astron. Soc. Australia, 27, 156
Jain A. K., Murty M. N., Flynn P. J., 1999, ACM Comput. Surv., 31, 264
Jayasinghe T., et al., 2019, MNRAS, 488, 1141
Jonker P. G., et al., 2011, ApJS, 194, 18
Logan C. H. A., Fotopoulou S., 2020, A&A, 633, A154
Malzer C., Baum M., 2021, Sensors, 21
Mann A., Kaur N., 2013.
Marín-Franch A., et al., 2012, in Navarro R., Cunningham C. R., Prieto E., eds, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 8450, Modern Technologies in Space- and Ground-based Telescopes and Instrumentation II. p. 84503S, doi:10.1117/12.925430
McInnes L., Healy J., Astels S., 2017, The Journal of Open Source Software, 2
Mendes de Oliveira C., et al., 2019, MNRAS, 489, 241
Merc J., Gàlis R., Wolf M., 2019, Eruptive Stars Information Letter, 41, 78
Nakazono L., et al., 2021, MNRAS, 507, 5847
Ntwaetsile K., Geach J. E., 2021, MNRAS, 502, 3417
Oke J. B., Gunn J. E., 1983, ApJ, 266, 713
Parker Q. A., et al., 2005, MNRAS, 362, 689
Parker Q. A., Bojičić I. S., Frew D. J., 2016, in Journal of Physics Conference Series. p. 032008 (arXiv:1603.07042), doi:10.1088/1742-6596/728/3/032008
Patterson J., 1984, ApJS, 54, 443
Pedregosa F., et al., 2011, Journal of Machine Learning Research, 12, 2825
Pickles A. J., 1998, PASP, 110, 863
Sabin L., Zijlstra A. A., Wareing C., Corradi R. L. M., Mampaso A., Viironen K., Wright N. J., Parker Q. A., 2010, Publ. Astron. Soc. Australia, 27, 166
Santos-Silva T., et al., 2021, arXiv e-prints, p. arXiv:2108.06234
Scaringi S., Groot P. J., Verbeek K., Greiss S., Knigge C., Körding E., 2013, MNRAS, 428, 2207
Taylor M. B., 2005, in Shopbell P., Britton M., Ebert R., eds, Astronomical Society of the Pacific Conference Series Vol. 347, Astronomical Data Analysis Software and Systems XIV. p. 29
Viironen K., et al., 2009, A&A, 502, 113
Vink J. S., Drew J. E., Steeghs D., Wright N. J., Martin E. L., Gänsicke B. T., Greimel R., Drake J., 2008, MNRAS, 387, 308
Ward J. H., 1963, Journal of the American Statistical Association, 58, 236
Webb S., et al., 2020, MNRAS, 498, 3077
Wevers T., et al., 2017, MNRAS, 466, 163
Witham A. R., et al., 2006, MNRAS, 369, 581
Witham A. R., et al., 2007, MNRAS, 382, 1158
Witham A. R., Knigge C., Drew J. E., Greimel R., Steeghs D., Gänsicke B. T., Groot P. J., Mampaso A., 2008, MNRAS, 384, 1277
Wu Y., et al., 2011, Research in Astronomy and Astrophysics, 11, 924

APPENDIX A: CONDENSED TREES

The condensed Trees is a diagram for `HDBSCAN` that allows to see the cluster hierarchy as a dendrogram. It can be displayed via the `condensed_tree_` attribute of the `HDBSCAN` package. Figure A1 shows the condensed trees which was obtained by using the $(r - z)$ and $(g - r)$ colours as the the input parameters. It is possible to see that `HDBSCAN` has found two clusters in agreement with previous results. This means that they represent the blue and red sources.

APPENDIX B: SIMBAD OBJECTS

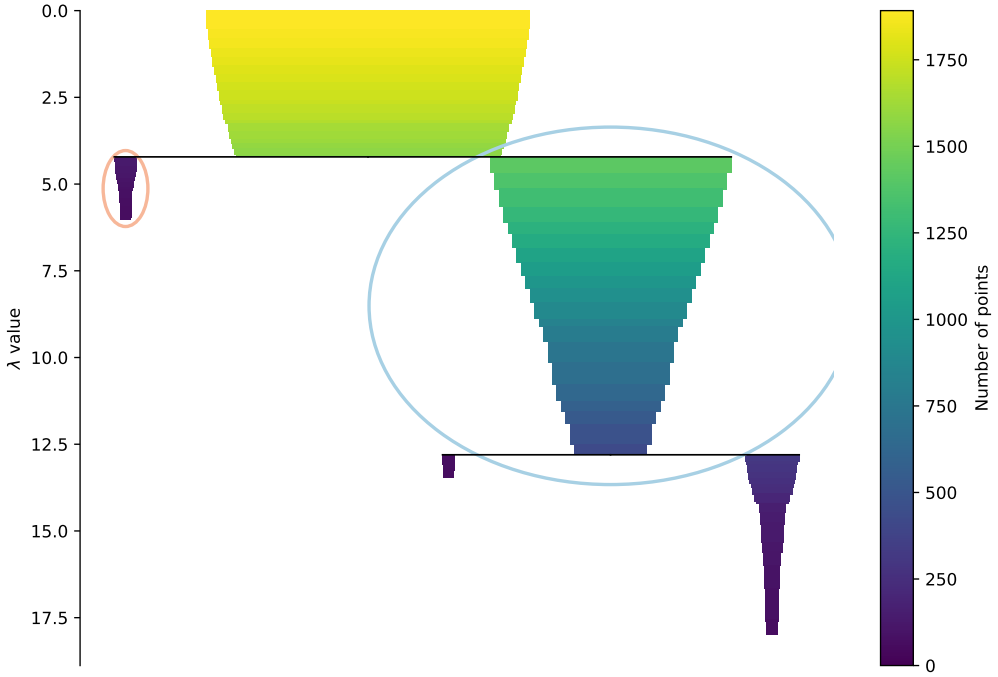


Figure A1. The condensed Trees for our sample of Hα emitters. The width and colour of each branch represent the number of points in the cluster at that level. The orange and blue ellipses represent the branches selected by the `HDBSCAN` algorithm.

Table B1: Examples of objects from the SIMBAD data base. The first column presents the ID SIMBAD of the source in question. Right ascension and declination are shown in the second and third columns, respectively. The type is given in the fourth column. The redshift, if exist in SIMBAD, is displayed in the fifth column. The colour-type classification performed with HAC algorithm is presented in the sixth column. The seventh and eighth columns show the probability estimated from HDBSCAN approach to being a blue and red source, respectively. The entire table will be available online as electronic format.

Id Object	RA	Dec	Type	Redshift	Group	P(Blue)	P(Red)
					HAC	HDBSCAN	HDBSCAN
SDSS J203521.96-011413.5	20:35:21.96	−01:14:13.4	low-mass*	–	Red	0.04	0.78
2SLAQ J204340.03+002853.4	20:43:40.04	00:28:53.6	Seyfert 1	0.317	Blue	0.44	0.16
SDSS J204643.27-000630.2	20:46:43.28	−00:06:30.1	CataclyV*	–	Blue	0.25	0.07
2SLAQ J204720.76+000007.7	20:47:20.76	00:00:07.7	CataclyV*	0.001	Blue	0.39	0.09
2SLAQ J204910.96+001557.2	20:49:10.95	00:15:57.5	Seyfert 1	0.363	Blue	0.28	0.13
[VV2006] J204956.6-001201	20:49:56.62	−00:12:01.7	QSO	0.369	Blue	0.40	0.18
SDSS J205352.03-001601.5	20:53:52.04	−00:16:01.5	QSO	0.363	Blue	0.37	0.14
2SLAQ J205712.69+001211.3	20:57:12.69	00:12:11.4	QSO	0.335	Blue	0.40	0.19
Gaia EDR3 6794425304909258752	20:58:06.45	−30:08:18.1	WD* Candidate	–	Blue	0.36	0.05
6dFGS gJ205957.5-213935	20:59:57.53	−21:39:34.9	Galaxy	-0.001	Blue	0.87	0.02
SDSS J210014.12+004446.0	21:00:14.11	00:44:45.9	CataclyV*	0.000	Blue	0.33	0.07
QSO B2059-330	21:02:41.71	−32:52:44.1	QSO	3.280	Blue	0.84	0.05
Gaia EDR3 6808104805812408064	21:03:56.66	−21:47:27.1	Star	–	Blue	0.66	0.06
[VV2006] J210514.1-004326	21:05:14.04	−00:43:26.4	QSO	3.250	Blue	0.92	0.08
PN G006.0-41.9	21:05:53.57	−37:08:40.4	PN	–	Blue	0.06	0.03
EC 21035-4032	21:06:48.02	−40:20:03.7	Star	–	Blue	0.28	0.07
2MASS J21122459-4128534	21:12:24.59	−41:28:53.3	AGN	0.349	Blue	0.40	0.10
1RXS J211805.2-341343	21:18:04.28	−34:13:43.3	CataclyV*	–	Blue	1.00	0.00
AT20G J212302-291504	21:23:02.82	−29:15:04.0	Radio(cm)	–	Blue	0.52	0.09
LBQS 2128-4555	21:31:29.53	−45:41:50.5	QSO	0.623	Blue	0.71	0.03
2MASS J21333817+0126291	21:33:38.14	01:26:29.0	QSO	1.010	Blue	0.99	0.01
SDSS J213455.08+001056.9	21:34:55.09	00:10:56.8	QSO	3.289	Blue	0.99	0.00
WISEA J213649.75-012852.2	21:36:49.75	−01:28:52.2	QSO	3.280	Blue	0.82	0.07
2MASS J21381896+0112224	21:38:18.96	01:12:22.5	Seyfert 1	0.344	Blue	0.23	0.14
LAMOST J213925.57+012345.6	21:39:25.57	01:23:45.5	QSO	1.389	Blue	1.00	0.00
CRTS J213937.6-023913	21:39:37.58	−02:39:13.0	CV* Candidate	–	Blue	0.25	0.14
SN 2017hxv	21:44:22.94	−29:54:59.0	SN	0.019	Blue	0.16	0.29
6dFGS gJ214540.0-291937	21:45:40.01	−29:19:36.9	Galaxy	0.341	Blue	0.14	0.31
SDSS J215002.70+011343.8	21:50:02.70	01:13:43.8	QSO	3.267	Blue	0.60	0.07
2MASS J21501054-0010002	21:50:10.53	−00:10:00.6	QSO	0.335	Blue	0.37	0.18
SDSS J220242.61-012528.0	22:02:42.61	−01:25:28.1	QSO	1.376	Blue	0.94	0.02
PB 5049	22:03:15.14	01:17:21.0	Star	–	Blue	0.08	0.04
SDSS J220529.34-003110.6	22:05:29.34	−00:31:10.7	QSO	2.454	Blue	0.44	0.10
[VV2006] J220601.0-304958	22:06:01.01	−30:49:57.5	QSO	1.330	Blue	0.87	0.03
2MASS J22085196-0106038	22:08:51.97	−01:06:03.7	QSO	0.351	Blue	0.43	0.17
2dFGRS TGS061Z180	22:09:19.05	−24:07:12.4	QSO	0.320	Blue	0.14	0.26
2QZ J220948.6-301357	22:09:48.63	−30:13:55.8	WD*	–	Blue	0.16	0.06
SDSS J220954.57-012717.6	22:09:54.57	−01:27:17.6	QSO	3.296	Blue	0.78	0.06
2QZ J221000.7-311400	22:10:00.75	−31:14:00.0	EmG	0.328	Blue	0.40	0.22
2QZ J221005.7-275439	22:10:05.76	−27:54:38.7	Galaxy	0.330	Blue	0.81	0.07
[VV2006] J221046.6-282345	22:10:46.69	−28:23:44.8	QSO	2.450	Blue	0.50	0.08
2QZ J221058.3-273930	22:10:58.33	−27:39:29.4	Galaxy	0.313	Blue	0.69	0.09
[VV2006] J221335.7-282542	22:13:35.65	−28:25:41.7	QSO	2.469	Blue	0.35	0.10
[VV2006] J221532.6-281805	22:15:32.58	−28:18:03.9	QSO	1.330	Blue	1.00	0.00
SDSS J221546.92-015906.6	22:15:46.92	−01:59:06.6	QSO	1.361	Blue	0.70	0.06
SDSS J221722.45+010436.3	22:17:22.44	01:04:36.3	QSO	1.403	Blue	0.13	0.08
SDSS J221811.46-005631.2	22:18:11.46	−00:56:31.3	QSO	1.351	Blue	0.88	0.01
2QZ J221819.4-271544	22:18:19.39	−27:15:44.2	Seyfert 1	0.355	Blue	0.33	0.16
2QZ J221945.1-293414	22:19:45.08	−29:34:13.4	EmG	0.343	Blue	0.88	0.03
SDSS J222011.76-015930.8	22:20:11.76	−01:59:30.7	QSO	1.344	Blue	1.00	0.00

APPENDIX C: SDSS AND LAMOST SPECTRA

Table C1: Examples of spectra from SDSS DR16. Coloured symbols represent the S-PLUS photometry as Fig. 3.

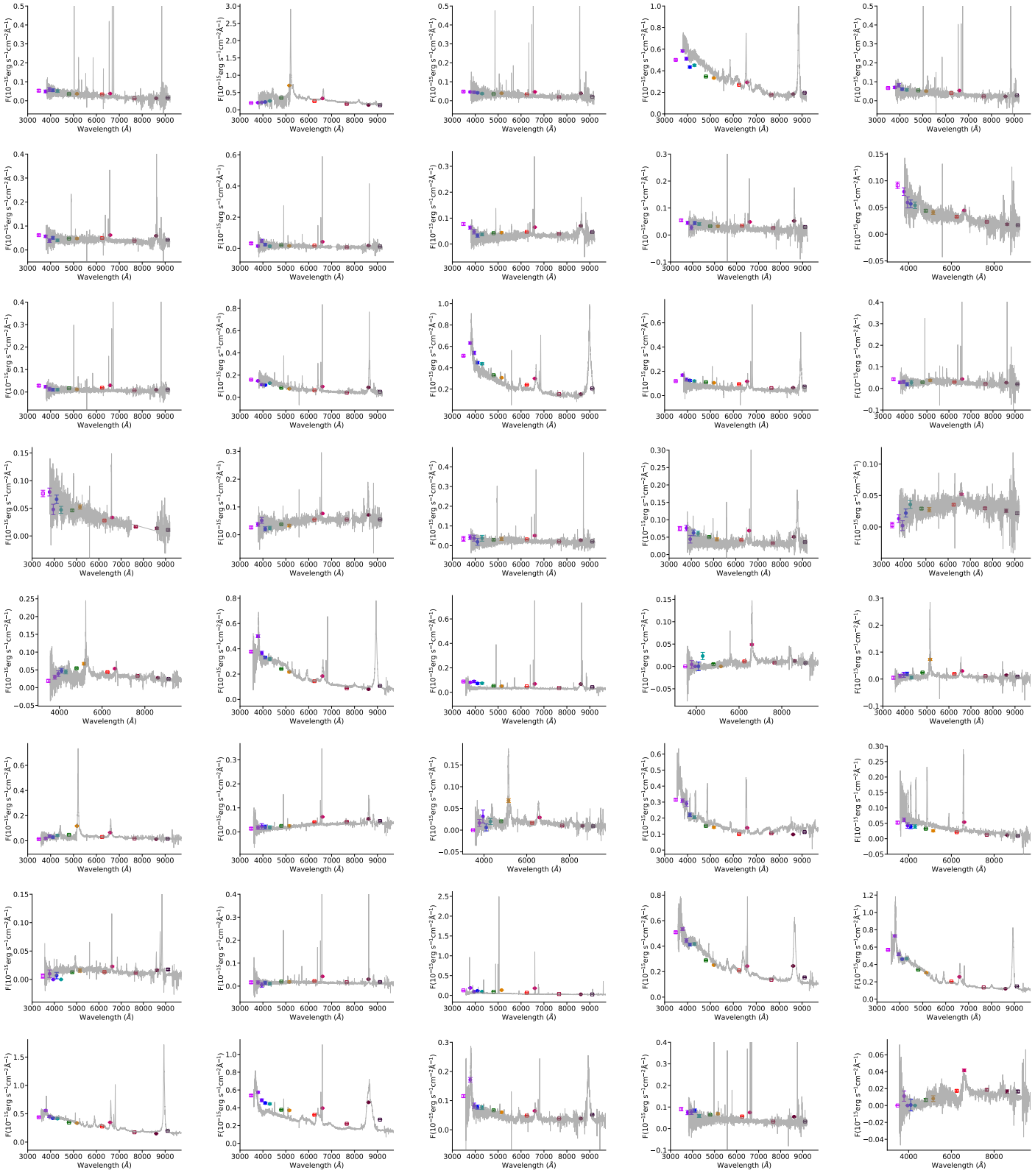
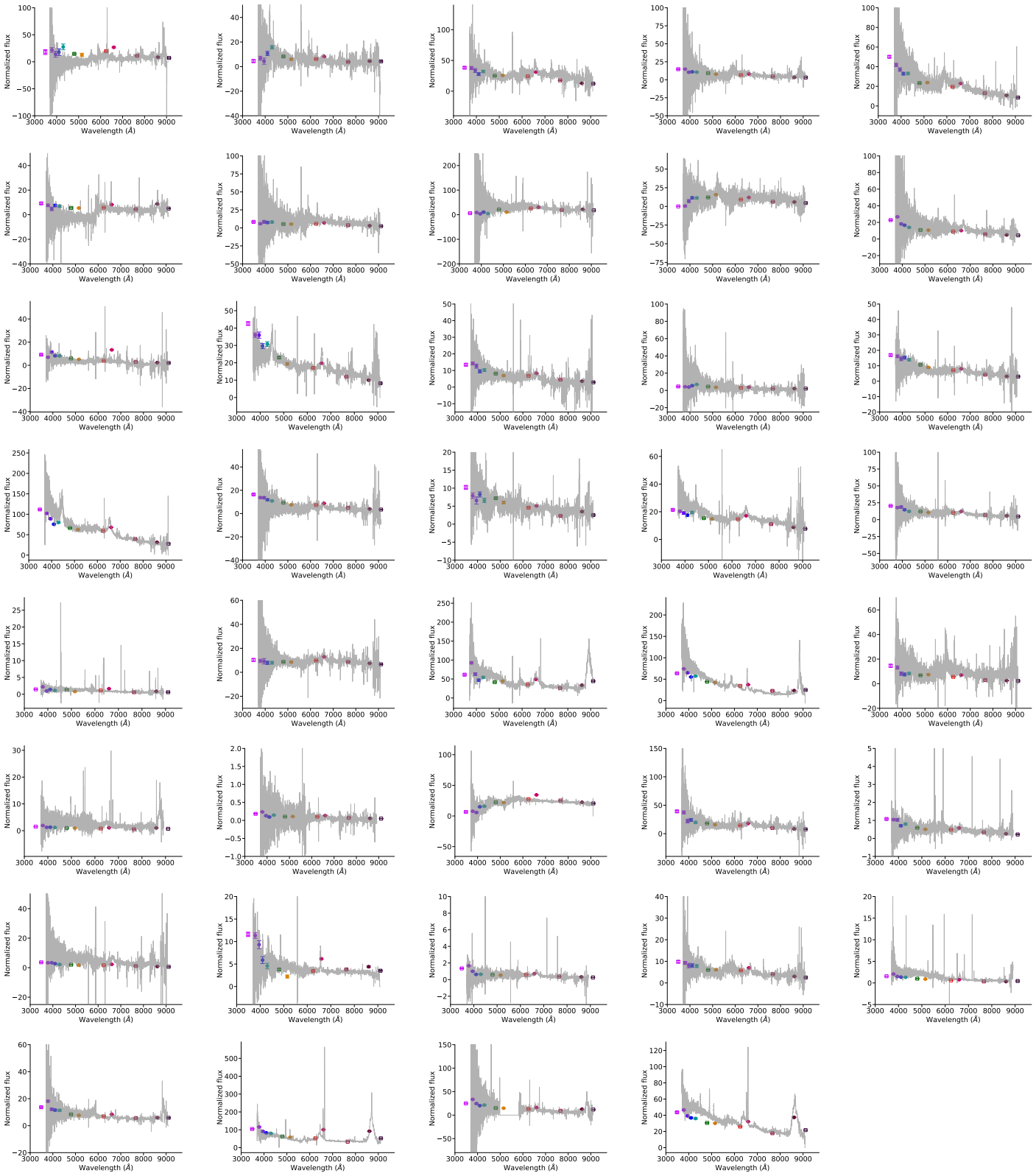


Table C2: Spectra from LAMOST DR7, scaled to match the *r*-band total magnitude from S-PLUS.



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