MNRAS 000, 1–27 (2021) Preprint 24 June 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-Fernandez[1], F. R. Herpich[1], A. Kanaan[6], T. Ribeiro[7], W. Schoenell[8]

[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] *Departamento de Física, Universidade Federal de Santa Catarina, Florianópolis, SC, 88040-900, Brazil*
[7] *NOAO, P.O. Box 26732, Tucson, AZ 85726*
[8] *GMTO Corporation 465 N. Halstead Street, Suite 250 Pasadena, CA 91107*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

The ongoing multi-colour 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 area 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 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 - i) versus (g - z) diagram. Our Hα-emitter catalog contains 8,446 objects that exhibit excess in the narrow $J0660$ band. For 8,039 of them, 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 a red continuum. 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 84% 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 $n = 2$ 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, PNe, supernovae, novae, circumstellar enveloped among others). Hydrogen recombination lines trace a vast variety of sources such as 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

★ E-mail: gsoto.angel@gmail.com

experiencing mass exchange like symbiotic stars (SySt), cataclysmic Variables (CVs), among others.

At much larger scales, the Hα emission line can also be emitted by planetary nebulae, H II regions or star-forming regions in galaxies, novae and supernova remnants, as well as other galaxies. 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, revealed 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 Galaxy and sources in nearby galaxies. Examples 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 ongoing VST Photometric Hα Survey (VPHAS+; Drew et al. 2014).

Traditionally, Hα emitters are revealed directly from images and in colour-colour diagrams from photometric surveys observing the sky with at most five - generally broad-band or Hα - filters. For example, the (r - Hα) versus ($r - i$) colour-colour diagram or a similar diagram 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 80 cm telescope. These are 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 using colour-colour diagrams and unsupervised machine learning techniques, classifying them as blue or red sources and also proposing a class to which they belong. 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.

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

2 METHODOLOGY

2.1 Observations: the S-PLUS project

This paper uses data from the S-PLUS DR3, available in the database of the project, splus/cloud, and it covers 2,000 square degrees. 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 centered 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). Note that the area including the Galactic disk and bulge was not included in this study (S-PLUS plans to cover ~ 1,300 deg^2 of such areas), given that both Galactic areas will be available only in DR4. Amongst the different aperture photometry available in the catalog we have used the PStotal photometry, 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 in at least the r, i and $J0660$ bands, simultaneously, with errors less than 0.2 mag, were considered.

The first goal of this paper 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.

2.2 Selection of Hα emitters

Before search for the Hα emitters, we first divided our sample into four sub-samples based on their magnitudes in the r-band. Thus, we considered the following four sub-samples: (i) r-band < 16, (ii) $16 \leq r < 18$, (iii) $18 \leq r < 20$, and (Vi) $20 \leq r \leq 21$. In this way, we avoid mixing up bright and faint sources with low and high uncertainties, respectively. Otherwise, the selection criteria would be potentially 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 to latter are also available in S-PLUS: r, $J0660$, and i filters. The same technique was 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

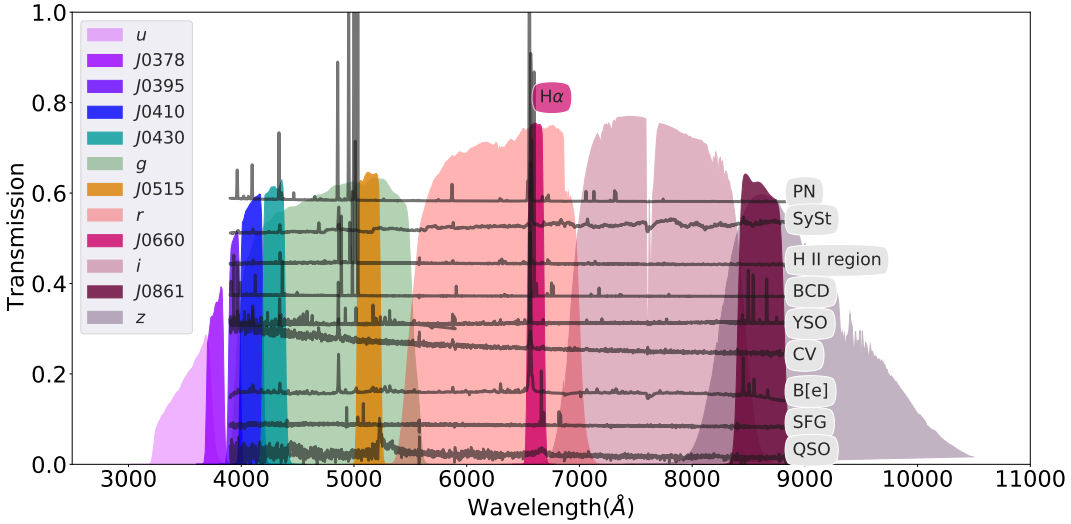


Figure 1. Transmission curves of the S-PLUS filter set. The narrow-band filter $J0660$ includes the Hα emission line. Over-plotted are spectra of different classes of emission line objects. From top to bottom: a PN, a symbiotic star, an extragalactic H II region, a blue compact/H II galaxy, a YSO, a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

magnitude bin and attempted to fit the loci mainly occupied by main-sequence and giant stars with a linear regression fit. We then implemented an iterative σ-clipping technique so that, by construction, Hα emitters should satisfy the condition:

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

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` [3]

Figure 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 should be noted that these cut-off lines are only approximations because only the residual around the fit is taken into account. The phothometric 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. 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 spectra that the excess in the $J0660$ band is linked with the Hα emission line.

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 transmission of the filters 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 Hα. 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.

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.

Our list of Hα emitters includes 8,446 sources. We now proceed to their classification into blue and red populations.

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 $(g - r)$ and $(z - g)$ 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 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

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

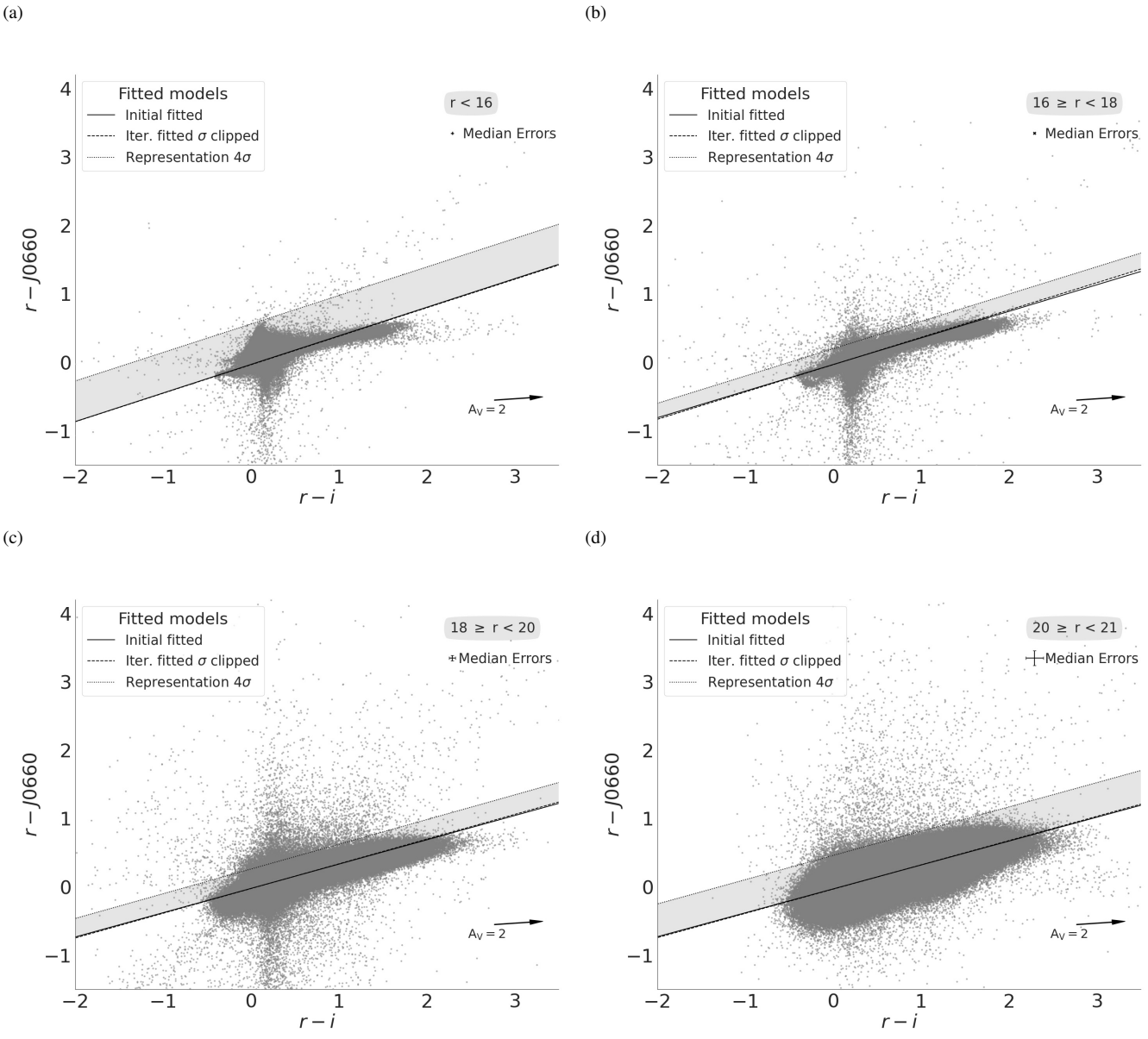


Figure 2. An illustration of the selection criteria used to identify strong emission-line objects via colour-colour plots. The data shown here are all from the S-PLUS DR3. The data are split up into four magnitude bins, as shown in the four panels. Objects with Hα excess should be located near the top 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.

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 between 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

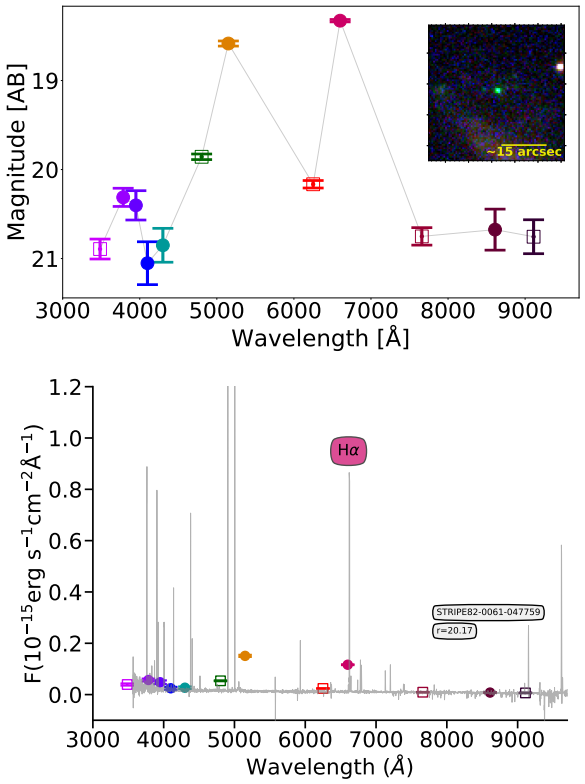


Figure 3. *Top panel:* S-spectrum of a random emitting object found by the method explained in Section 2.2. Open squares represent the SDSS-like broad-band filters. From left to right: u, g, r, i and z magnitudes. Circles represent the narrow-band filters, which from left to right correspond to $J0378$,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

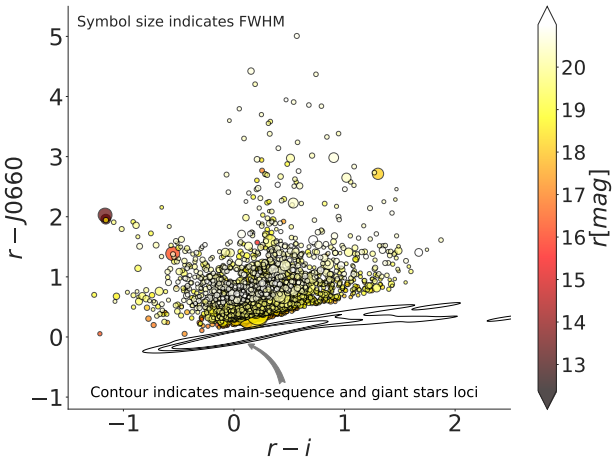


Figure 4. Colour-colour diagram with all the emission-line objects selected from the S-PLUS DR3. The size of the symbols represent the measured FWHM assuming a Gaussian core (for more detail see Almeida-Fernandes et al. 2022). Coloured bar indicates the magnitude values of the r-band. The contours represent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

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 variables 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 $d_\mathrm{m}(a,b) = \min\{\mathrm{core}_k(a), \mathrm{core}_k(b), d(a,b)\}$, where $d(a,b)$ is the distance between a and b according, for instance, 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 sort the edges of the tree by distance 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, effectively, 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 conserved.

(vi) The clusters are extracted on the notion of persistence in the hierarchy. The parameter $\lambda = d_\mathrm{m}^{-1}$ is defined, and each cluster has a λ_birth (the point at which the cluster split off) and λ_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 $\lambda_\mathrm{birth} \leq \lambda_p \leq \lambda_\mathrm{death}$. Cluster stability S is defined as the sum of $\lambda_p - \lambda_\mathrm{birth}$ for all points in the cluster. To extract clusters, the following procedure is implemented. First, each leave 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 $S_p = \sum_{i=0}^{n} S_d^i$.

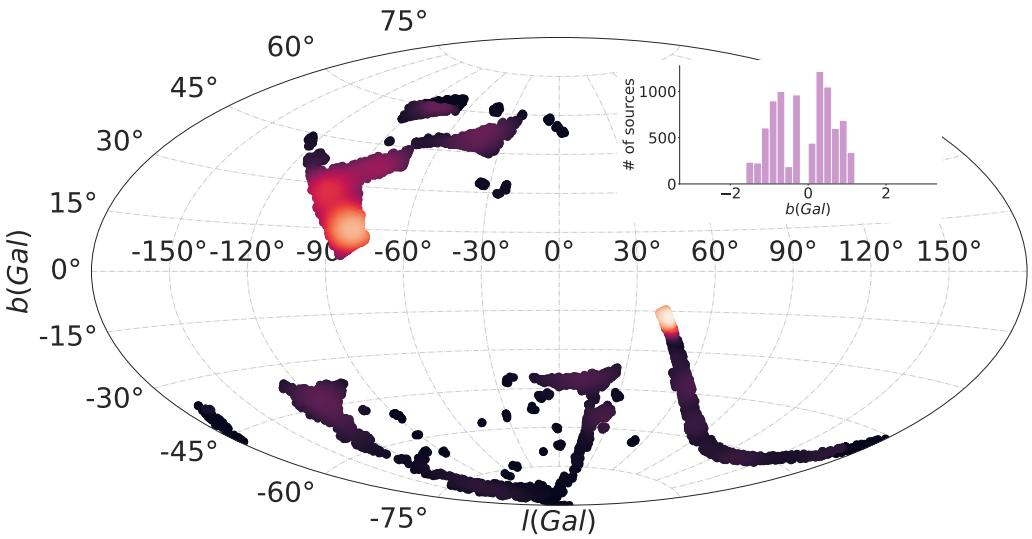


Figure 5. Distribution of the emission-line objects in galactic longitude and latitude coordinates. The inset figure represents the distribution of the objects in galactic latitude.

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 may solve the problem due to noise classification.

The HDBSCAN algorithm starts off much the same as DBSCAN: transforming the space according to density, exactly as DBSCAN does, and perform single linkage clustering on the transformed space. Instead of taking an epsilon[4] 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. 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).

[4] 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.

2.4 Splitting the Hα emitters into blue and red populations

For the selection of 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. Such diagram, the $(g-r)$ versus $(r-z)$ colour-colour diagram is displayed in Fig. 6. SySt and YSOs span a wide range on $(z-g)$, from -0.5 to 6.0. On the other hand, the PNe, H II regions, CVs, QSOs, and emission line galaxies are located on the lower-right region of the diagram. The dashed line in Fig. 6 highlights the blue and red zones.

Fig. 7 displays the $(g-r)$ versus $(z-g)$ diagram from the list of Hα emitters in S-PLUS. Obviously only such emitters with detections in the g and z filters are considered for this colour classification by making a cut in the magnitude errors at 0.2, totalizing 7086 objects. A bi-modal distribution is found for both colour indices (see inset plots of the Fig. 7. The two peaks on the $(g-r)$ and $(z-g)$ distributions have immediate correspondence with the blue and red zones pointed out from the synthetic diagram (Fig. 6). One can also see that the fraction of blue objects is considerable larger than that of the red ones.

2.4.1 HAC

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 $(z-g)$

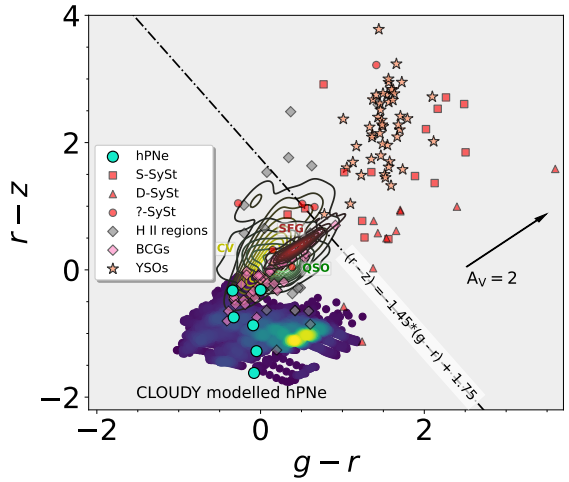


Figure 6. The $(g-r)$ versus $(z-g)$ synthetic colour-colour diagram of several classes of emission lines objects. Included in the diagrams, there are families of CLOUDY modelled halo PNe spanning a range of properties (density map region). Cyan circles represent S-PLUS photometry from observed spectra 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 circles 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.

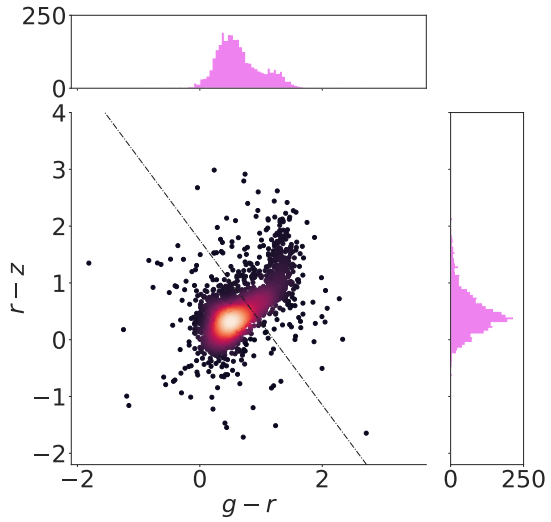


Figure 7. The $(g-r)$ versus $(z-g)$ colour-colour diagram with all the emission line objects selected in S-PLUS. The inset figures represent the $(g-r)$ and $(z-g)$ colour distributions.

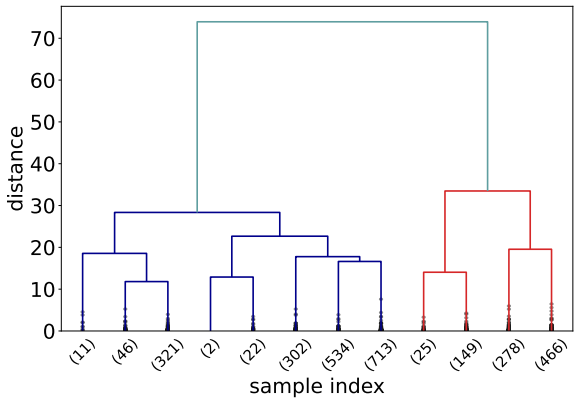


Figure 8. Truncated dendrogram of complete-linkage hierarchical clustered based on $(g-r)$ and $(z-g)$ colours. The cluster sizes are exposed in the brackets for the 12 truncated clusters.

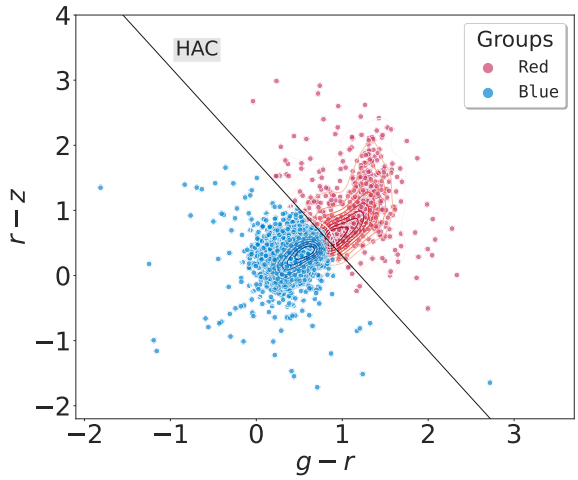


Figure 9. The $(g-r)$ versus $(z-g)$ colour-colour diagram with the two population found by implementing HAC algorithm. The blue and red symbols represent the sources with intense blue continuum and those with intense red continuum, respectively. The straight line is the same line of Figure 6.

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: coincidentally, they correspond to the red and blue populations of Hα emitters presented in Fig. 7 as it is showed 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

`Scikit-learn`[5] (Pedregosa et al. 2011). Then, the DENDROGRAM() function, which is included in the package `scipy`[6] was used in conjunctions wit 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`. `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 observation. 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 cluster 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 continuum, with the blue population (5,338 objects) being larger than the red one (1,748).

2.4.2 *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`[7] (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 "80" 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 "40" was adopted. With this model configuration , two clusters were identified. Several small clusters are found when the minimum number of samples values becomes smaller than 40.

Left panel of Fig. 10 shows the two clusters found with `HDBSCAN`. One cluster contains 192 red sources and the other one 3,825 blue sources. The number of objects classified as noise is 3,069. This result is fully consistent with those obtained from `HAC`. The two main clusters obtained with `HDBSCAN` are located in the same region in the $(g-r)$ versus $(z-g)$ diagram as those groups found based on the `HAC`. About 94% 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 noises) classifications is the implementation of the "soft clustering" (see section 2.3.2). Soft clustering from `HDBSCAN`[8] was used to assign every object to a cluster that they most likely belong to. According 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 measure 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 has been improved. Instead of forcing the algorithm to make a decision to which group a data point belongs to, just `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 that is focused on the identification of Hα emitters in the S-PLUS footprint, exploiting the unique filter system of the survey returned 8,446 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[9]. Optical spectra available in the SDSS DR16 (Ahumada et al. 2020) 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 ($d_{max,proj}$).

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

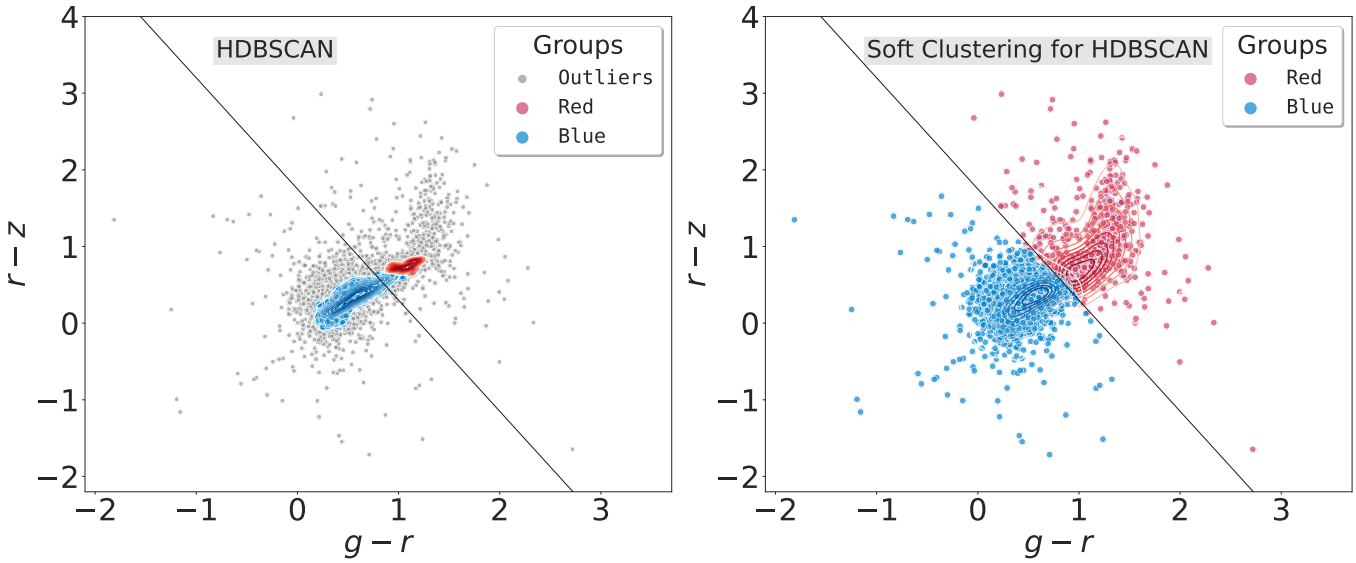


Figure 10. *Left panel*: as Figure 9 but with the results after to apply HDBSCAN to the sample of emission-line sources. Blue and red symbols correspond to blue and red sources, respectively, with gray symbols representing sources classified as noise by HDBSCAN. The straight line is the same line as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

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	HII, PN, SN, Candidate_SN*, Nova	32
Stellar binary system	CataclyV*, Candidate_CV*, EB*, HMXB	51
Star	star, WD*, Candidate_WD*, Blue, BlueSG*, PM*, low-mass*, Cl*, GlCl	60
Variable star	RRLyr, Candidate_RRLyr, V*, PulsV*	23
Galaxy	EmG, HII_G, StarburstG, BlueCompG, IG, PartofG, GinPair, GinGroup, GinCl, LSB_G, BClG, Galaxy	577
QSO	QSO, QSO_Candidate	209
AGN	AGN, AGN_Candidate, Seyfert_1, Seyfert_2, BLLac, RadioG	65
Other type	EmObj, FIR, MIR, MolCld, UV, Transient, Radio, X, Possible_lensImage, Unknown	25
Total		1,042

3.1 Matches with SIMBAD sources

We found 1,042 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 and in nearby galaxies are listed in our sample, most being H II regions, PNe, novae and SNe. H II regions are ionized by the UV light from early, massive stars (OB-type) and display an emission line spectrum. Unlike H II regions, 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. Supernovae or even novae also display emission line spectra and come from evolved stars through multiple channels. However, the energy-input mechanism is quite different in each case.

In our list of Hα emitters only one PN is catalogued in SIMBAD. Its S-spectrum and SDSS spectrum are displayed on panel (*a*) of Fig 11. Emission lines like Hα and [N II] are clearly perceptive in its spectra. This nebula belongs to the rare 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. 30 sources in our Hα emitters list are catalogued as H II regions based on the SIMBAD database. The S-PLUS photometry and SDSS spectrum of the extragalactic H II region GALEX 2417063145906373262 is illustrated in panel (*b*). Panel *c* of Figure 11 also shows the SDSS and S-spectrum, as well as the coloured image, of an extragalatic SN, with evident emission lines.

[8] https://hdbscan.readthedocs.io/en/latest/soft_clustering_explanation.html
[9] http://simbad.u-strasbg.fr/simbad/

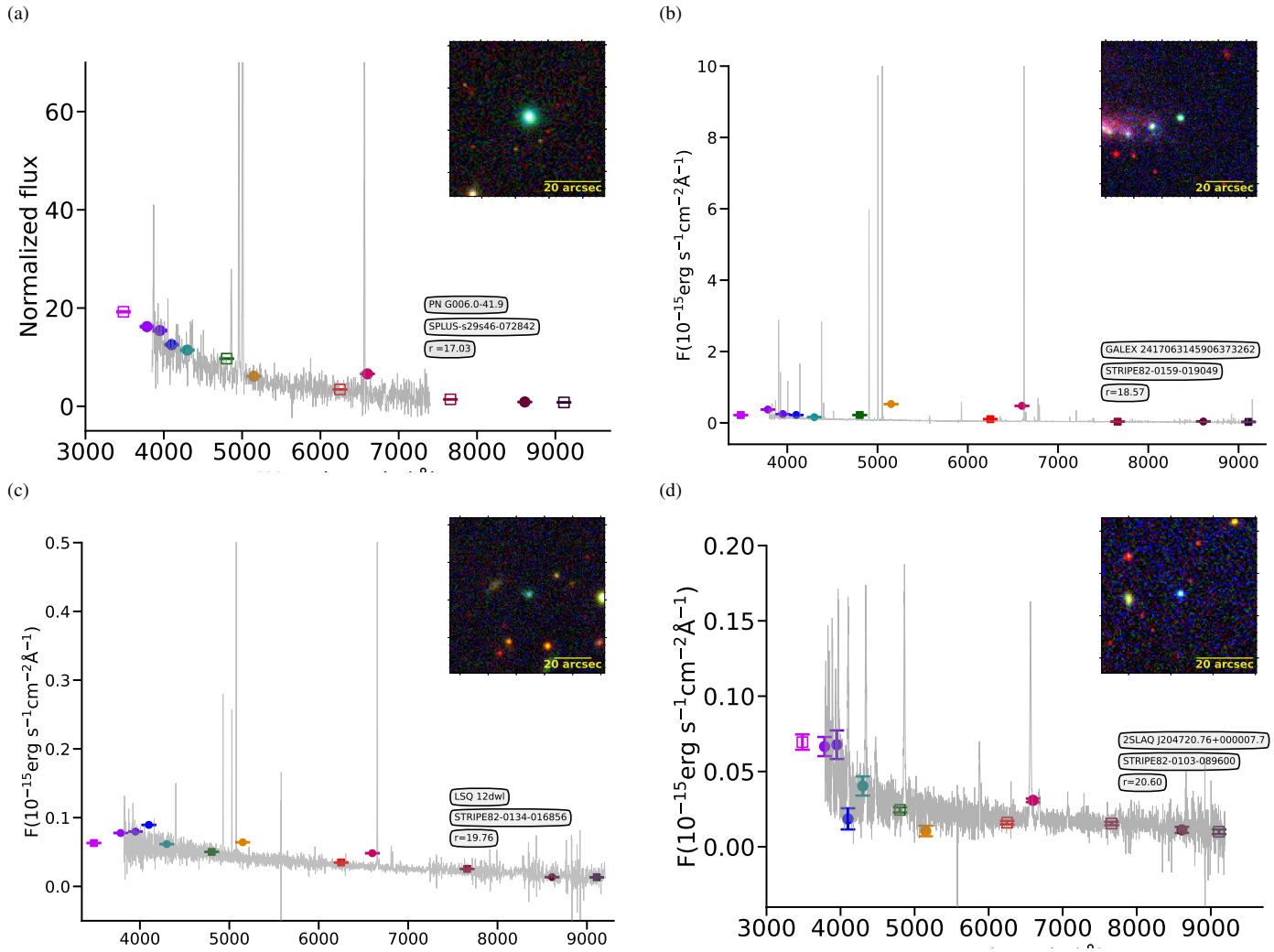


Figure 11. Summary of our selection results showing the spectrum (gray line en each panel) of different classes of emission line sources identified in our target list. A spectrum of a PN (*a*) from (Parker et al. 2016). The SDSS spectra of an external H II region (*b*), a super nova (*c*), a cataclysm variable star (*d*). As in Figure 3, coloured square and circles symbols represent the S-PLUS photometry. All these objects show a significance excesses on the *J*0660 filter in comparison with the broad-bands.

3.1.2 Interacting binary systems

Following the classification available in Simbad, 20 known and 7 candidate CVs 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. 11 (panel *d*) 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. X-ray binary systems as well as eclipsing binaries are also listed on our catalog.

3.1.3 Stars

Several objects in our sample of Hα emitters have been categorized, by SIMBAD, as: normal stars, white dwarf (WD*), white dwarf candidates (Candidate_WD*), blue stars, blue super-giant stars (BlueSG*), high proper-motion stars (PM*), variable stars of RR Lyr type and, low-mass star (low-mass*; M<1M$_\odot$),

3.1.4 Galaxies

Galaxies are also included in our catalog: emission-line galaxies (EmG), blue compact galaxies (BlueCompG), H II galaxies (HII G), starburst galaxies (StarburstG), galaxies in clusters (GinCl) and in groups (GinGroup), low surface brightness galaxies (LSB G), radio-galaxies (RadioG), interacting galaxies (IG), part of a Galaxy (PartofG), Seyfert types- 1 and 2, and other type of AGNs. Since, we are focusing on the Hα emission line, the emission line galaxies in the local Universe ($z \sim 0.02$) are of particular interest because their Hα line still falls into the wavelength range covered by the *J*0660 S-PLUS filter.

Fig. 12 shows the redshift distribution of the galaxies in our sample that have SIMBAD correspondents. About 60% of the galaxies have small redshift values ($z < 0.02$) probing that the emission detected in the *J*0660 filters is associated to the real Hα λ6563 emission line. On the other hand, Fig. 12 also shows that there is an increment of galaxies with redshift between ~0.31 and ~0.38. This particular population of Hα emitters is represented by AGN, Seyfert 1 and Seyfert 2 galaxies. In fact, at the accumulative redshift range, 0.306 <

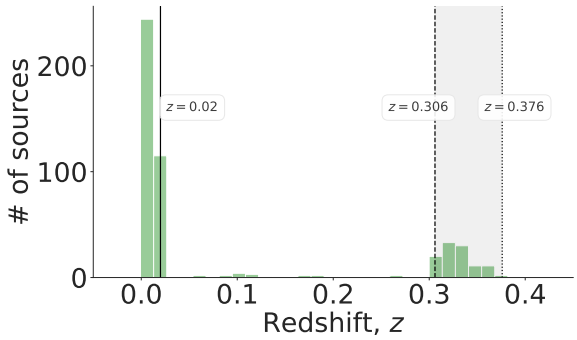


Figure 12. Histogram of redshift for the galaxies with SIMBAD coincidence. The black vertical continuous line indicates the threshold value ($z \sim 0.02$) on which the Hα emission-line is detected in $J0660$ filter. The black vertical dashed and dotted lines represent the accumulative redshift range where the Hβ and [O III] 4959, 5007 Å emission lines are detected on the $J0660$-band.

$z < 0.376$, represented by the filled area in the figure, the Hβ and [O III] 4959, 5007Å emission lines are the ones detected by the $J0660$ filter.

It should be noted here that the majority of the sources with a classification of "part of a Galaxy" (PartofG) in SIMBAD actually designates galaxies with Wolf-Rayet (WR) signature in the low redshift Universe also named "WR galaxy" (Osterbrock & Cohen 1982). The presence of WR stars in galaxies is perceptible from their spectra, as suggested by strong emission lines such as Hα and [N II]. However, these spectral features can also be indicative of extragalactic H II regions which are typically present in the outskirt of spiral galaxies. Panel (*e*) of Fig. 13 displays the S-PLUS and SDSS spectra of the galaxy with Wolf-Rayet signature [BKD2008] WR 14.

Some Hα emitters are classified in SIMBAD as "galaxy". These galaxies can in fact be spiral galaxies in which star formation is still active. Panel *c* of Fig. 13 shows the S-PLUS photometry and the SDSS spectrum of a star-forming galaxy. The spectrum clearly exhibits strong emission lines. Note that almost all these galaxies are classified as blue objects by `HAC` and `HDBSCAN`. However, most of the Seyfert type-2 galaxies and radio-galaxies as well as a handful of other galaxies are found to fill up the population of red sources.

3.1.5 QSOs: false positive detection of Hα emission

Following the classification available in the literature, about 3% of blue Hα emitters sources in our sample are found to be QSOs. We have to point out here that the excess in the $J0660$ 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, C IV 1550 Å, C III] 1909 Å, and Mg II 2798 Å (see Gutiérrez-Soto et al. 2020 and Nakazono et al. 2021). QSO 2SLAQ J220529.34-003110.6, shown in Panel *f* of Figure 13, is an example of a QSO at redshift ~2.45, for which the C III] line falls at the range covered by the $J0660$ filter.

3.1.6 Other type 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 clusters of stars, far- and mid-infrared sources, molecular clouds, UV sources, 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 479 sources belonging to both catalogues. And about 96% of them display strong emission lines spectra.

Most of the Hα emitters with available spectroscopic information correspond to H II regions, CVs, SNe, emission-line galaxies (blue compact galaxies, H II galaxies, star-forming galaxies, among others), AGN (Seyfert 1 and 2), 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 projects provide a good validation to our approach, clearly showing that the methodology is actually effective for selecting sources with the Hα emission line in emission.

3.3 Magnitudes and colour distributions

In Fig. 14, we demonstrate the distribution of the blue and red population of S-PLUS Hα emitters in term of their r magnitude and their $(r - i)$ and $(r\text{-}J0660)$ colours.

Both, blue and red sources can be as bright as 16 mag in the r filter, while they show a peak at ~20 mag. The fraction of blue sources in the $16.0 \leq r \leq 19.0$ magnitude range is considerable higher in comparison with the red group. Therefore, the blue sample tends to be brighter than the red one in the r-band.

Middle panel of Fig. 14 displays the $(r - i)$ distribution of the blue and red Hα sources which peak at distinct values of -0.9 and 0.5, 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 bottom panel of Fig. 14 shows the $(r - J0660)$ colour index distribution of the blue and red objects with peak at 0.5 and 0.7, respectively. This result implies a strong 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 databases.

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 in the local Universe with an Hα-excess ($r - J0660 > 0$) (see Fig. 4). 8,446 sources matched this criterion, with 407 of them claimed in the literature as QSOs and non-local galaxies, and therefore being false positive identifications of Hα (see sections 3.1.4 and 3.1.5).

The $(g - r)$ and $(z - g)$ 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

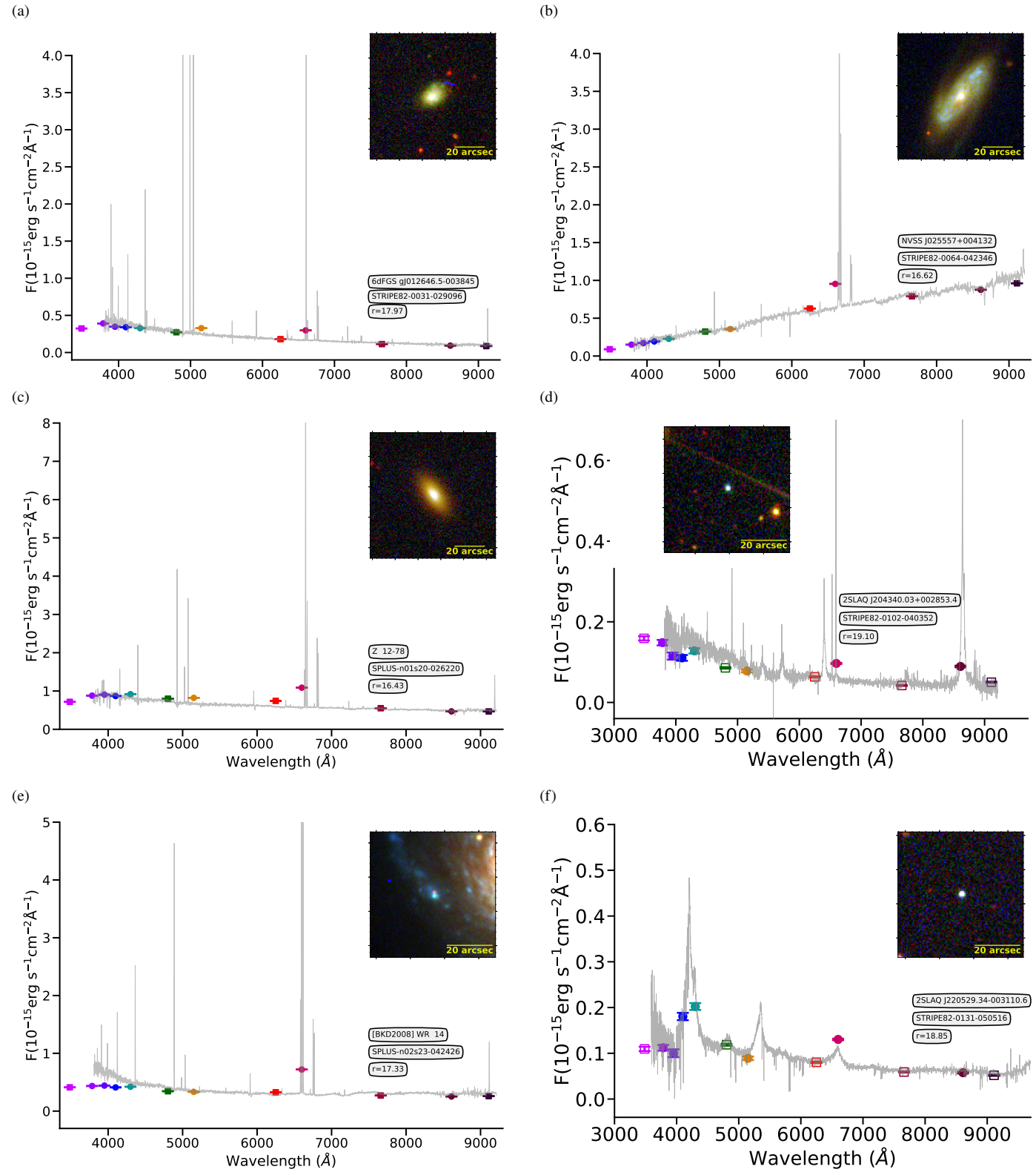


Figure 13. SDSS spectra of a H II galaxy with $z = 0.006$ (*a*), a radio galaxy with $z = 0.014$ (*b*), a star-forming galaxy with $z = 0.013$(*c*). For this object, the Hα line is responsible for the $J0660$ magnitude. And a Seyfert 1 with $z = 0.317$ (*d*). For this last object, the excess on the $J0660$ is due to the [O III] 4959, 5007 Å emission lines. a WR in a galaxy (*e*) and a QSO (*f*) with redshift of ~ 2.45. As in Figure 11, coloured symbols indicate the S-PLUS photometry.

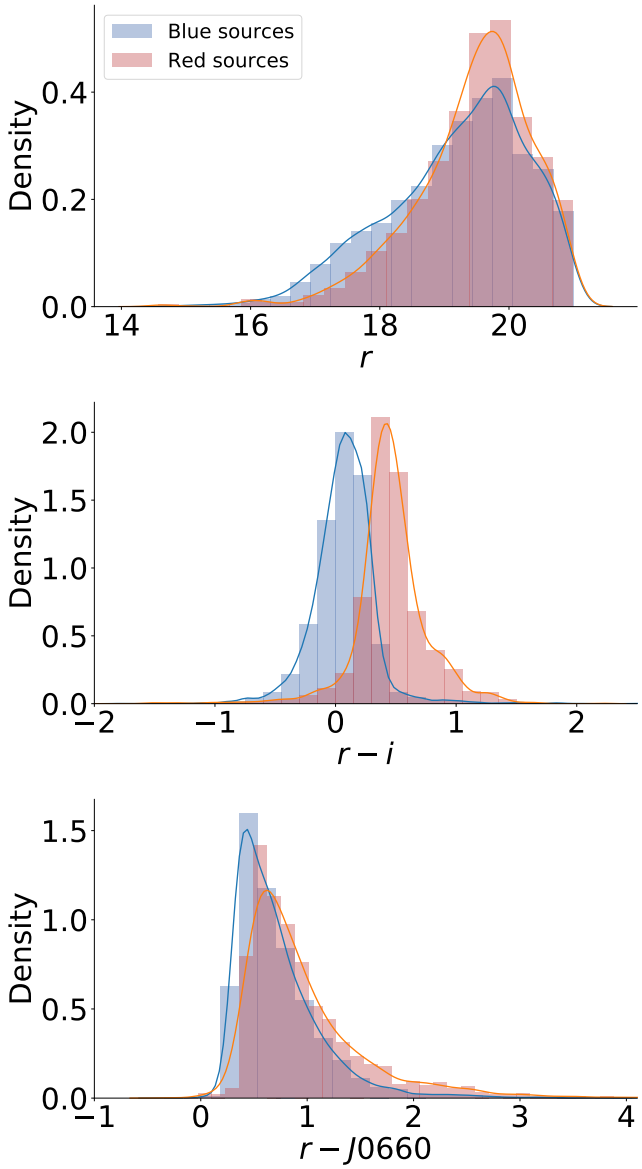


Figure 14. Distribution of r magnitude (*top panel*), $(r - i)$ colour (*middle panel*), and $(r - J0660)$ colour (*bottom panel*) for the blue and red sources of the sample of Hα emitters. The histogram heights show density normalization scales. The smooth curves represent a Kernel density estimation for both samples.

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 $(g - r)$ and $(z - g)$ 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. We, therefore, reckon that the $(g - r)$ and $(z - g)$ colours are ideal for separating objects into the bluer and redder populations and attribute their colour to the nature of the sources. In particular, the bluer objects were found to be mainly CVs, PNe, H II regions, dwarf compact galaxies, and QSOs, among others, while the redder sources are early type galaxies with emission lines (for instances, radio-galaxies and Seyfert 2 galaxies), probably young/active late-type stars or even symbiotic stars (in fact, evolved binary systems hosting a red giant star).

Finally, we also cross-matched our catalog 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 Hα emitters in the S-PLUS project. The spectroscopic sample of Hα emitters lists 240 sources of the local Universe (with $z < 0.02$) indicating that the emission on the $J0660$ filter corresponds to the Hα line, 239 sources with redshift larger than 0.02, indicating that they are very likely QSOs and non-local galaxies on which the excess of the $J0660$ filter is due to C IV 1550 Å, C III] 1909 Å, and Mg II 2798 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.

As a practical result, here we make public a catalog from SPLUS/DR3 that can be explored by the community in the identification and investigation of sources in twelve photometric bands in a systematic and homogeneous way.

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.

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 Uni-

versity, 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[10] (Taylor 2005), simbad and vizier from Strasbourg Astronomical Data Center (CDS)[11] and the following python packages: numpy, astropy, matplotlib, seaborn, scikit-learn.

DATA AVAILABILITY

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
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
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
Osterbrock D. E., Cohen R. D., 1982, ApJ, 261, 64
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

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

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 $(r - g)$ and $(g - z)$ 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

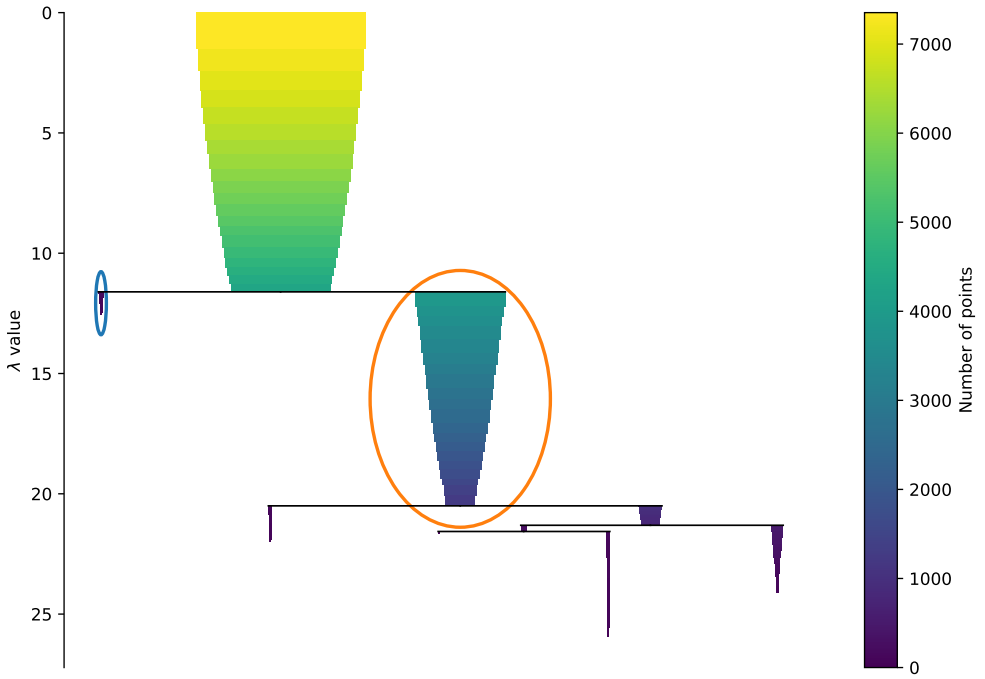


Figure A1. The condensed Trees for our sample of Hα emitters. The width and colour of each branch represent the number of points in the cluster at that level. The orange and blue ellipses represent the branches selected by the `HDBSCAN` algorithm.

Table B1: 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.

Id Object	RA	Dec	Type	Redshift	Group HAC	P(Blue) HDBSCAN	P(Red) HDBSCAN
QSO B2359+005	00:02:30.71	00:49:59.2	QSO	1.354	Blue	0.68	0.08
SDSS J000637.99-003656.3	00:06:37.99	−00:36:56.2	QSO	4.435	–	–	–
LBQS 0004+0036	00:07:10.00	00:53:29.1	QSO	0.316	Blue	0.47	0.23
SDSS J000809.34+004935.5	00:08:09.34	00:49:35.3	QSO	3.293	–	–	–
2SLAQ J000918.74-003907.2	00:09:18.76	−00:39:07.0	Galaxy	–	Blue	0.68	0.11
[VV2006] J001040.1-294428	00:10:40.08	−29:44:27.3	QSO	1.361	Blue	0.22	0.08
2QZ J001055.3-304423	00:10:55.37	−30:44:23.5	Galaxy	0.307	Blue	0.47	0.16
[VV2006] J001228.8-310241	00:12:28.78	−31:02:40.0	QSO	1.360	Blue	0.54	0.16
LBQS 0010+0035	00:13:27.32	00:52:32.2	Seyfert 1	0.363	Blue	0.70	0.22
2SLAQ J001455.99+001903.5	00:14:55.99	00:19:03.7	Star	–	Blue	0.48	0.13
SDSS J001526.52+001813.2	00:15:26.52	00:18:13.4	QSO	1.362	Blue	0.32	0.12
[VV2006] J001535.5+005355	00:15:35.55	00:53:56.1	QSO	1.358	Blue	0.61	0.06
[VV2006] J001641.9-312657	00:16:41.87	−31:26:56.6	QSO	0.360	Blue	0.43	0.14
SDSS J001731.27-004859.3	00:17:31.26	−00:48:59.2	QSO	1.357	Blue	0.55	0.18
SDSS J001753.82+005057.6	00:17:53.82	00:50:57.7	QSO	1.358	–	–	–
SDSS J001912.39+000319.6	00:19:12.39	00:03:19.8	QSO	1.372	Blue	0.74	0.15
SDSS J001940.23-005435.9	00:19:40.24	−00:54:35.8	QSO	1.374	Blue	0.49	0.12
SDSS J002237.90+000519.0	00:22:37.90	00:05:19.2	QSO	1.373	Blue	0.58	0.15
SDSS J002940.02+010528.5	00:29:40.02	01:05:28.7	QSO	1.387	Blue	0.61	0.13
[VV2010c] J002951.5+004159	00:29:51.45	00:42:00.0	AGN	0.315	Red	0.22	0.31
SDSS J003117.70+001705.0	00:31:17.69	00:17:05.1	QSO	4.335	–	–	–
2QZ J003137.5-292815	00:31:37.50	−29:28:15.3	Unknown Candidate	–	Blue	0.80	0.06
2QZ J003152.5-293534	00:31:52.56	−29:35:33.3	Galaxy	0.313	Blue	0.73	0.06
SDSS J003208.53-005303.7	00:32:08.53	−00:53:03.6	QSO	1.344	Blue	0.35	0.12
SDSS J003234.62-001557.1	00:32:34.62	−00:15:57.1	QSO	3.243	Blue	0.32	0.13
[VV2006] J003242.7+003111	00:32:42.74	00:31:11.1	QSO	0.360	Blue	0.13	0.10
[VV2006] J003545.9+002306	00:35:45.86	00:23:06.0	QSO	3.237	Blue	0.62	0.11
2MASS J00362543-0029075	00:36:25.39	−00:29:07.1	AGN	0.308	Red	0.06	0.55
[VV2006] J003722.2-001140	00:37:22.17	−00:11:40.6	QSO	1.370	Blue	0.60	0.10
SDSS J003859.33-004252.3	00:38:59.35	−00:42:52.0	QSO	2.502	Blue	0.07	0.04
SDSS J004243.87+011701.9	00:42:43.87	01:17:02.2	QSO	1.366	Blue	0.61	0.19
2dFGRS TGS501Z235	00:43:21.88	−33:19:02.9	Galaxy	0.015	Blue	0.19	0.08
2SLAQ J004335.13-003729.7	00:43:35.16	−00:37:29.6	CataclyV*	–	Blue	0.23	0.10
SDSS J004400.26+004724.6	00:44:00.26	00:47:24.7	Galaxy	0.345	Blue	0.23	0.10
SDSS J004415.82-004303.1	00:44:15.81	−00:43:03.1	QSO	3.248	–	–	–
[VV2006] J004544.4-315729	00:45:44.35	−31:57:29.2	QSO	1.344	Blue	0.61	0.19
2SLAQ J004626.30-011417.0	00:46:26.30	−01:14:16.8	Galaxy	–	Blue	0.53	0.08
[VV98] J004826.9-341340	00:48:26.97	−34:13:38.7	QSO	1.910	Blue	0.48	0.22
SDSS J004918.52+011308.9	00:49:18.52	01:13:09.1	QSO	1.339	Blue	0.52	0.18
CRTS J005019.2-000241	00:50:19.30	−00:02:40.6	EB*	–	Red	0.08	0.77
QSO B0049-272	00:51:55.64	−26:57:43.3	QSO	2.484	Blue	0.17	0.08
SDSS J005343.78+012147.6	00:53:43.76	01:21:47.5	QSO	1.358	Blue	0.64	0.20
QSO B0052-307	00:54:43.95	−30:30:54.1	QSO	2.450	Blue	0.43	0.15
[VV2006] J005532.1-311538	00:55:32.08	−31:15:37.8	QSO	1.350	Blue	0.62	0.18
[CT83] 219	00:55:51.35	−30:56:42.8	UV	–	Blue	0.43	0.15
[VV2006] J005609.9-312209	00:56:09.93	−31:22:08.6	QSO	2.460	Blue	0.26	0.12
[VV2006] J005639.0-315759	00:56:39.05	−31:57:58.6	QSO	1.350	Blue	1.00	0.00
[VV2000] J005840.5-300203	00:58:40.42	−30:02:00.1	QSO	1.361	Blue	0.30	0.12
LBQS 0057-0135	00:59:48.81	−01:19:05.2	QSO	0.325	Blue	0.43	0.23
QSO B0057-3948	00:59:53.21	−39:31:57.3	QSO	3.240	Blue	0.61	0.11
SCMS 679	01:00:04.44	−33:39:32.5	Star	–	Blue	0.49	0.12

Table B1: –continued

Id Object	RA	Dec	Type	Redshift	Group HAC	P(Blue) HDBSCAN	P(Red) HDBSCAN
[MFW2011] 24026	01:01:27.67	−33:47:30.7	Star	–	–	–	–
QSO B0059-304B	01:02:14.65	−30:07:53.8	QSO	3.240	Blue	0.52	0.10
2SLAQ J010230.03-003206.8	01:02:30.02	−00:32:06.8	Seyfert 1	0.343	Blue	0.47	0.23
[VV2006] J010336.4-005508	01:03:36.39	−00:55:08.8	QSO	2.443	Blue	0.48	0.20
SDSS J010413.86-011552.1	01:04:13.86	−01:15:52.0	QSO	1.366	Blue	0.91	0.09
QSO B0103+00	01:06:19.23	00:48:23.4	QSO	4.435	Red	0.03	0.05
2MASS J01065344-3243420	01:06:53.44	−32:43:41.9	AGN	0.371	Blue	0.54	0.13
[VV2006] J010705.6+000609	01:07:05.55	00:06:09.0	QSO	1.357	Blue	0.62	0.18
SDSS J010907.59+000649.8	01:09:07.59	00:06:50.0	QSO	1.372	Blue	0.53	0.17
SDSS J010918.56+005419.3	01:09:18.56	00:54:19.4	QSO	1.356	Blue	0.59	0.19
SDSS J010925.95-003739.0	01:09:25.96	−00:37:39.0	QSO	1.360	Blue	0.59	0.07
2QZ J011014.0-302445	01:10:13.97	−30:24:44.5	EmG	0.313	Blue	0.49	0.12
SDSS J011128.38+000143.7	01:11:28.35	00:01:43.3	QSO	0.765	Red	0.12	0.16
SDSS J011230.55+001441.5	01:12:30.55	00:14:41.7	QSO	3.259	Blue	0.71	0.11
2SLAQ J011402.35-004750.9	01:14:02.35	−00:47:50.8	Seyfert 1	0.350	Blue	0.50	0.19
[VV2006] J011405.3-310903	01:14:05.25	−31:09:02.8	QSO	1.333	Blue	0.16	0.08
SDSS J011542.18+002300.2	01:15:42.18	00:23:00.4	QSO	1.373	Blue	0.61	0.13
[VV2006] J011553.9-312439	01:15:53.96	−31:24:38.9	QSO	1.340	Blue	0.37	0.13
SDSS J011818.13+001455.2	01:18:18.12	00:14:55.5	QSO	1.372	Blue	0.76	0.10
2SLAQ J011829.63+004549.4	01:18:29.62	00:45:49.4	Seyfert 1	0.314	Blue	0.42	0.28
SDSS J012110.74-005037.2	01:21:10.74	−00:50:37.1	QSO	1.352	Blue	0.53	0.12
[HB93] 0119-341B	01:21:52.19	−33:56:15.8	Star	–	Red	0.22	0.31
QSO B0120-002	01:23:01.78	00:03:23.6	QSO	1.356	Blue	0.67	0.17
2dFGRS TGS297Z222	01:23:50.87	−29:11:46.4	Galaxy	0.000	Blue	0.06	0.04
QSO B0121-324	01:24:16.18	−32:12:21.7	QSO	1.358	Blue	0.55	0.18
QSO B0122-3232	01:25:04.59	−32:17:14.6	QSO	2.450	Blue	0.32	0.13
2QZ J012526.2-304433	01:25:26.24	−30:44:32.8	EmG	0.311	Blue	0.65	0.16
2QZ J012549.3-280944	01:25:49.29	−28:09:43.6	Galaxy	0.324	–	–	–
SDSS J012753.04+003817.3	01:27:52.94	00:38:17.1	Galaxy	0.152	Red	0.10	0.20
SDSS J013034.18-002106.6	01:30:34.17	−00:21:06.5	QSO	3.234	Blue	0.62	0.12
SDSS J013303.18+005102.6	01:33:03.19	00:51:02.8	Galaxy	0.313	Blue	0.49	0.29
[VV2006] J013500.8-004054	01:35:00.83	−00:40:54.2	QSO	1.007	Blue	0.11	0.06
FBQS J0135-0019	01:35:17.53	−00:19:39.0	Seyfert 1	0.312	Blue	0.36	0.22
SDSS J013701.72-012059.3	01:37:01.71	−01:20:59.1	QSO	2.496	Blue	0.45	0.13
[VV2006] J013729.4-320715	01:37:29.40	−32:07:15.7	QSO	1.368	Blue	0.49	0.13
[VV2006] J013837.3+002818	01:38:37.28	00:28:18.5	QSO	1.348	Blue	0.47	0.13
SDSS J013951.07+002537.9	01:39:51.07	00:25:38.0	QSO	1.342	Blue	0.82	0.07
LEDA 6178	01:40:17.06	−00:50:03.0	Seyfert 1	0.334	Blue	0.19	0.10
[VV2006] J014224.7-320414	01:42:24.73	−32:04:13.7	QSO	2.460	Blue	0.31	0.11
2SLAQ J014227.07+001729.8	01:42:27.08	00:17:30.0	CataclyV*	-0.000	Blue	0.04	0.04
[VV2006] J014303.6-295255	01:43:03.49	−29:52:54.8	QSO	2.450	Blue	0.34	0.14
SDSS J014721.12-004505.3	01:47:21.12	−00:45:05.3	QSO	1.348	Red	0.35	0.43
[VV2006] J014739.2-285259	01:47:39.21	−28:52:59.2	QSO	0.360	Blue	0.43	0.13
[VV2006] J014921.5-003220	01:49:21.53	−00:32:20.9	QSO	1.379	Blue	0.62	0.12
2QZ J015257.7-284838	01:52:57.76	−28:48:37.8	Seyfert 1	0.326	Blue	0.47	0.23
SDSS J015331.85+002252.8	01:53:31.85	00:22:53.0	QSO	1.367	Blue	0.67	0.12
SDSS J015409.27+002645.2	01:54:09.27	00:26:45.3	QSO	1.355	Blue	0.58	0.14
[VV2006] J015832.1-301703	01:58:32.16	−30:17:02.7	QSO	1.380	Blue	0.65	0.07
[VV2006] J015935.4+000401	01:59:35.48	00:04:01.5	QSO	3.277	Blue	0.49	0.15
SDSS J020025.40+002916.6	02:00:25.40	00:29:16.8	QSO	0.313	Red	0.07	0.51
[VV2006] J020055.0-293527	02:00:55.02	−29:35:26.5	QSO	1.349	Blue	0.63	0.11
[VV98] J020115.4+003136	02:01:15.53	00:31:35.1	QSO	0.362	Blue	0.26	0.12
SDSS J020200.06-000921.2	02:02:00.06	−00:09:21.2	QSO	1.359	Blue	0.73	0.18
[VV96] J020435.5-455923	02:04:35.46	−45:59:24.0	QSO	3.240	Blue	0.60	0.14
SDSS J020921.99-005455.5	02:09:22.00	−00:54:55.4	QSO	1.367	Blue	0.87	0.09
SDSS J021529.02-005314.8	02:15:29.02	−00:53:14.9	QSO	1.369	Blue	0.63	0.10
SDSS J021617.19-011046.8	02:16:17.19	−01:10:46.7	QSO	3.264	Blue	0.95	0.05
SDSS J021810.52-010147.4	02:18:10.52	−01:01:47.2	QSO	1.353	Blue	0.51	0.10

Table B1: –continued

Id Object	RA	Dec	Type	Redshift	Group HAC	P(Blue) HDBSCAN	P(Red) HDBSCAN
V* AX For	02:19:28.00	−30:45:46.0	CataclyV*	–	Blue	0.39	0.16
SDSS J022010.02-005646.6	02:20:10.02	−00:56:46.5	QSO	0.338	Red	0.12	0.16
2QZ J022112.5-302559	02:21:12.54	−30:25:59.0	EmG	0.315	Blue	0.71	0.22
SDSS J022714.48+010536.1	02:27:14.47	01:05:36.3	EmG	0.349	Blue	0.54	0.12
[VV2006] J022738.3-313627	02:27:38.28	−31:36:26.4	QSO	1.350	Blue	0.55	0.16
2SLAQ J022945.34+000856.2	02:29:45.34	00:08:56.4	Star	–	Blue	0.48	0.22
2QZ J022954.6-303558	02:29:54.69	−30:35:58.4	Seyfert 1	0.372	Blue	0.30	0.14
SDSS J023020.93+001355.5	02:30:20.93	00:13:55.8	Seyfert 1	0.335	Blue	0.14	0.10
SDSS J023230.64-011654.5	02:32:30.63	−01:16:54.5	QSO	1.364	Blue	0.61	0.12
SDSS J023248.71+005138.8	02:32:48.71	00:51:38.8	Galaxy	0.344	Blue	0.35	0.08
V* HP Cet	02:33:22.62	00:50:59.4	Nova	-0.000	Blue	0.32	0.11
[VV2006] J023335.4-010744	02:33:35.37	−01:07:44.6	QSO	0.367	Blue	0.49	0.21
[VV2006] J023635.7-003203	02:36:35.69	−00:32:03.4	QSO	1.362	–	–	–
SDSS J024059.15+004545.8	02:40:59.14	00:45:45.9	QSO	3.233	Blue	0.86	0.05
CRTS J024109.5+004813	02:41:09.54	00:48:13.5	EB*	–	Red	0.32	0.58
[VV2006] J024235.0-010351	02:42:34.91	−01:03:51.9	QSO	1.373	Blue	0.65	0.07
CRTS J024408.0+000324	02:44:08.08	00:03:24.2	EB*	–	Blue	1.00	0.00
SDSS J025100.65+001707.2	02:51:00.64	00:17:07.3	QSO	2.466	Blue	0.46	0.14
SDSS J025252.02-002211.7	02:52:52.00	−00:22:11.6	QSO	1.370	Blue	0.54	0.13
QSO B0253+0058	02:56:07.25	01:10:38.8	QSO	1.349	Blue	0.70	0.12
2MASSI J0259103-002239	02:59:10.38	−00:22:39.8	Seyfert 1	0.360	Blue	0.47	0.23
LBQS 0302-0019	03:04:49.85	−00:08:13.4	QSO	3.295	Blue	0.66	0.12
LBQS 0303+0110	03:06:12.72	01:21:57.3	QSO	1.335	Blue	0.51	0.11
SDSS J030757.55+000712.0	03:07:57.55	00:07:12.1	QSO	1.343	Blue	0.61	0.12
SDSS J031129.69-001701.4	03:11:29.70	−00:17:01.5	QSO	1.357	Blue	0.55	0.14
2QZ J031130.9-315250	03:11:30.92	−31:52:51.1	WD*	–	Blue	0.24	0.10
SDSS J031258.36-000453.6	03:12:58.36	−00:04:53.6	Galaxy	0.117	Blue	0.82	0.09
2SLAQ J031428.25+004506.6	03:14:28.25	00:45:07.0	Galaxy	–	Blue	1.00	0.00
2SLAQ J031618.00-003025.2	03:16:18.01	−00:30:24.9	Galaxy	–	Blue	1.00	0.00
2SLAQ J031829.06-000040.3	03:18:29.06	−00:00:40.5	Galaxy	–	Blue	0.59	0.08
[VV2006] J031845.2-001844	03:18:45.17	−00:18:45.3	QSO	3.224	Blue	0.92	0.08
SDSS J031937.30-002641.1	03:19:37.29	−00:26:41.0	QSO	1.371	Blue	0.28	0.10
SDSS J032244.90+004442.4	03:22:44.90	00:44:42.3	QSO Candidate	0.304	Blue	1.00	0.00
QSO B0323-381	03:24:54.31	−37:57:00.1	QSO	0.341	Blue	0.49	0.16
SDSS J033226.29-011126.2	03:32:26.29	−01:11:26.0	QSO	1.361	Blue	0.57	0.17
2MASS J03342942+0006112	03:34:29.44	00:06:11.1	Seyfert 1	0.348	Blue	0.70	0.23
[VV2006] J033458.5-000744	03:34:58.48	−00:07:43.9	QSO	1.357	Blue	0.86	0.05
[VV2006] J033821.6+003106	03:38:21.51	00:31:06.6	QSO	1.349	Blue	0.63	0.11
SDSS J034019.89+010330.7	03:40:19.89	01:03:30.7	EmG	0.322	Blue	0.43	0.21
2MASS J03424773+0109331	03:42:47.72	01:09:33.0	Seyfert 1	0.360	Red	0.09	0.09
2SLAQ J034304.64+002512.1	03:43:04.65	00:25:12.3	Star	–	Blue	0.45	0.25
[VV2006] J034408.3-003106	03:44:08.25	−00:31:05.8	QSO	1.646	Blue	0.81	0.12
SDSS J034517.02-001549.8	03:45:17.01	−00:15:49.7	QSO	1.335	Blue	1.00	0.00
FASTT 83	03:51:19.36	00:32:16.6	EB*	–	Red	0.14	0.56
Gaia EDR3 4857261601188886016	03:55:16.01	−37:29:44.7	WD* Candidate	–	–	–	–
[ZJM2003] SA 95-2230	03:55:38.45	00:28:34.9	Star	–	Blue	0.98	0.02
[VV96] J041130.5-335331	04:11:30.51	−33:53:31.1	QSO	1.350	Blue	0.65	0.09
Gaia EDR3 4872129059981617536	04:20:06.78	−32:51:20.0	WD* Candidate	–	–	–	–
6dFGS gJ043139.6-301514	04:31:39.57	−30:15:14.1	CataclyV*	-0.000	Blue	0.22	0.17
CRTS J095754.0-384019	09:57:54.08	−38:40:19.0	EB*	–	Blue	0.68	0.04
CRTS J095950.7-383024	09:59:50.88	−38:30:22.9	RRLyr	–	Blue	0.92	0.08
CRTS J100044.3-352518	10:00:44.35	−35:25:18.0	EB*	–	Blue	0.82	0.11
COSMOS 1949846	10:01:01.83	02:20:18.3	Star	0.040	Red	0.07	0.71
CRTS J100201.2-390943	10:02:01.37	−39:09:42.5	EB*	–	Blue	0.93	0.07
RE J1002-19	10:02:11.73	−19:25:37.1	CataclyV*	–	Blue	0.22	0.09
SDSS J100215.83-001056.1	10:02:15.83	−00:10:55.8	QSO	0.353	Blue	0.45	0.20
CRTS J100246.0-370053	10:02:46.10	−37:00:53.9	EB*	–	Blue	0.93	0.07
[VV96] J100342.1-150808	10:03:41.93	−15:08:08.9	QSO	0.342	Blue	0.38	0.19

Table B1: –continued

Id Object	RA	Dec	Type	Redshift	Group HAC	P(Blue) HDBSCAN	P(Red) HDBSCAN
CRTS J100414.7-292007	10:04:14.73	−29:20:07.4	RRLyr	–	Blue	0.91	0.09
[VV2006] J100539.9+040914	10:05:39.88	04:09:14.7	QSO	1.355	Blue	0.52	0.13
CRTS J100603.2-402957	10:06:03.37	−40:29:56.0	EB*	–	Red	0.68	0.22
CRTS J100733.7-301921	10:07:33.84	−30:19:19.4	EB*	–	Blue	0.66	0.10
RX J1007.5-2017	10:07:34.65	−20:17:32.4	CataclyV*	–	Blue	0.30	0.18
CRTS J100734.9-383117	10:07:35.06	−38:31:17.1	EB*	–	Red	0.17	0.36
CRTS J101200.8-365725	10:12:00.81	−36:57:25.2	EB*	–	Red	0.18	0.45
Gaia EDR3 5407412036686860672	10:12:47.58	−47:33:51.1	Star	–	Blue	0.69	0.13
Gaia EDR3 5414200039210586240	10:15:42.13	−45:15:30.8	RRLyr Candidate	–	Blue	0.26	0.18
V* KO Vel	10:15:58.31	−47:58:09.1	CataclyV*	-0.000	Blue	0.19	0.08
Gaia EDR3 5413356443209110912	10:16:24.12	−47:35:21.4	Star	–	Blue	0.93	0.07
CRTS J101853.5-400644	10:18:53.51	−40:06:43.7	CV* Candidate	–	Blue	0.18	0.16
CRTS SSS110628 J102019-300035	10:20:18.61	−30:00:35.3	CV* Candidate	–	–	–	–
CRTS J102042.2-335002	10:20:42.16	−33:50:02.4	CV* Candidate	–	Blue	0.42	0.13
Gaia EDR3 5668001579559758720	10:20:43.31	−20:47:54.6	Star	–	Blue	0.39	0.13
CRTS J102206.2-252159	10:22:06.22	−25:21:59.2	RRLyr	–	Blue	0.89	0.07
2MASS J10223994-3029305	10:22:39.94	−30:29:30.6	AGN Candidate	0.317	Blue	0.41	0.16
CRTS J102424.0-164933	10:24:24.03	−16:49:33.3	EB*	–	Red	0.24	0.39
2MASX J10244566-1838271	10:24:45.66	−18:38:26.6	Galaxy	0.051	Blue	0.02	0.01
NGC 3242	10:24:46.13	−18:38:32.3	PN	0.000	Blue	0.04	0.02
CRTS J102513.4-354014	10:25:13.46	−35:40:16.7	EB*	–	Blue	0.86	0.02
NAME OT J102705.8-434341	10:27:05.83	−43:43:41.3	CataclyV*	–	Blue	0.43	0.11
CRTS CSS140309 J102844-161303	10:28:43.86	−16:13:03.3	CataclyV*	–	Red	0.07	0.07
ATO J158.2117-27.8636	10:32:50.82	−27:51:49.2	EB* Candidate	–	Red	0.19	0.46
6dFGS gJ103530.3-182048	10:35:30.33	−18:20:47.6	Galaxy	0.344	Blue	0.44	0.22
CRTS J103634.8-262023	10:36:34.73	−26:20:21.9	EB*	–	Blue	0.99	0.01
WISE J103754.92-242544.5	10:37:54.92	−24:25:44.6	MIR	–	Red	0.16	0.53
2MASS J10395999-4701261	10:39:59.97	−47:01:26.3	CataclyV*	–	Blue	0.29	0.11
CRTS J104104.0-341120	10:41:03.86	−34:11:23.4	RRLyr	–	Blue	0.65	0.10
ATO J160.9042-19.0551	10:43:37.04	−19:03:18.5	EB* Candidate	–	Red	0.04	0.76
Gaia EDR3 5391507429181636352	10:47:23.91	−41:59:49.3	Star	–	Blue	0.50	0.20
6dFGS gJ105233.0-230900	10:52:33.04	−23:08:59.6	Galaxy	0.318	Blue	0.18	0.12
CRTS J105653.2-353907	10:56:53.29	−35:39:07.0	EB*	–	Blue	0.94	0.06
CRTS J105702.6-264020	10:57:02.63	−26:40:19.6	RRLyr	–	Blue	0.69	0.03
CRTS J105706.2-335338	10:57:06.29	−33:53:38.9	RRLyr	–	Blue	0.90	0.10
EC 10566-3120	10:58:59.03	−31:36:34.1	CataclyV*	–	Blue	0.29	0.10
Gaia EDR3 5386613537284200960	11:01:51.26	−46:53:04.5	RRLyr Candidate	–	Blue	0.84	0.04
Gaia EDR3 3537117430403448320	11:01:57.97	−23:47:27.3	PM*	–	Blue	0.25	0.10
V* TU Crt	11:03:36.57	−21:37:45.9	CataclyV*	–	Blue	0.48	0.22
[VV96] J111644.8-171127	11:16:43.58	−17:11:41.5	QSO	0.375	Blue	0.43	0.23
CRTS J111737.6-171934	11:17:37.72	−17:19:33.2	RRLyr	–	Blue	0.68	0.10
CRTS J112256.0-242841	11:22:56.09	−24:28:40.0	EB*	–	Blue	0.57	0.11
[VV2010c] J113128.4-195903	11:31:28.46	−19:59:02.8	AGN	0.363	Blue	0.51	0.19
CRTS SSS110509 J113219-213943	11:32:19.01	−21:39:42.9	CV* Candidate	–	–	–	–
V* RZ Leo	11:37:22.18	01:48:58.9	CataclyV*	-0.000	Blue	0.10	0.08
Gaia EDR3 3541998025080414336	11:37:49.97	−20:07:37.1	WD* Candidate	–	Blue	0.15	0.07
CRTS J113855.5-211148	11:38:55.60	−21:11:47.7	RRLyr	–	Blue	0.92	0.08
LBQS 1136-0109	11:39:04.35	−01:26:25.0	QSO	1.375	Blue	0.61	0.09
2QZ J114214.5-023154	11:42:14.64	−02:31:53.3	Galaxy	0.319	Blue	0.77	0.13
CRTS J114238.0-202722	11:42:37.96	−20:27:21.8	RRLyr	–	Blue	0.72	0.09
2QZ J114250.9+013057	11:42:50.95	01:30:58.2	Seyfert 1	0.361	Blue	0.67	0.21
SDSS J114329.34-020319.7	11:43:29.34	−02:03:19.5	QSO	3.304	–	–	–
SDSS J114408.82+012420.5	11:44:08.82	01:24:20.7	RRLyr	0.001	Blue	0.66	0.12
Gaia EDR3 3544179185567992320	11:44:55.76	−17:56:39.4	WD* Candidate	–	Blue	0.20	0.08
SDSS J114643.11+011118.6	11:46:43.12	01:11:18.8	QSO	3.220	Blue	0.58	0.14
[VV2006] J114939.6+014624	11:49:39.60	01:46:25.5	QSO	1.362	Blue	0.50	0.20
[VV2006] J115049.2-005149	11:50:49.29	−00:51:49.1	QSO	1.354	Blue	0.18	0.08
SDSS J115129.42-000333.8	11:51:29.45	−00:03:33.6	Galaxy	0.326	Red	0.08	0.28

Table B1: –continued

Id Object	RA	Dec	Type	Redshift	Group HAC	P(Blue) HDBSCAN	P(Red) HDBSCAN
[VV2006] J115345.5-024320	11:53:45.44	−02:43:20.4	QSO	1.347	Blue	0.54	0.06
2QZ J115737.0-020138	11:57:37.09	−02:01:37.2	Galaxy	0.328	Blue	0.72	0.11
[VV2006] J115748.0+014320	11:57:48.02	01:43:20.9	QSO	1.364	Blue	0.67	0.16
GAMA 137854	11:59:23.49	−01:43:22.3	Galaxy	0.304	Blue	0.38	0.32
GAMA 584657	12:00:06.54	−00:10:42.6	Galaxy	0.166	–	–	–
SDSS J120021.76-024331.0	12:00:21.77	−02:43:30.9	QSO	3.248	Blue	0.61	0.12
[VV2006] J120038.3+011246	12:00:38.29	01:12:46.5	QSO	1.358	Blue	0.65	0.18
QSO B1158-1842	12:00:44.95	−18:59:44.5	QSO	2.453	Blue	0.42	0.16
QSO B1158+007	12:01:23.26	00:28:28.5	QSO	1.369	Blue	0.62	0.18
CRTS J120206.7-230305	12:02:06.75	−23:03:06.0	EB*	–	Blue	0.91	0.09
MGC 28924	12:05:32.33	−00:16:57.4	Star	–	Blue	0.01	0.01
[VV2006] J120700.4+011155	12:07:00.41	01:11:56.4	QSO	1.520	Blue	0.49	0.14
[VV2006] J120825.7+010354	12:08:25.72	01:03:55.5	QSO	1.340	Blue	0.68	0.10
SDSS J120920.53-002855.3	12:09:20.55	−00:28:55.3	QSO	3.237	Blue	0.72	0.17
LINEAR 3056354	12:12:59.78	01:49:23.2	EB*	–	Blue	0.38	0.29
6dFGS gJ121348.2-143140	12:13:48.16	−14:31:39.8	Galaxy	0.330	Blue	0.48	0.16
SDSS J121435.25-015924.4	12:14:35.26	−01:59:24.4	QSO	3.233	Blue	0.44	0.13
[VV2006] J121515.2-013542	12:15:15.23	−01:35:40.8	QSO	1.350	Blue	0.55	0.13
QSO B1216+0216	12:18:55.80	02:00:02.1	QSO	0.327	Blue	0.27	0.20
[VV2006] J121942.5-001821	12:19:42.47	−00:18:21.4	QSO	1.337	Blue	0.83	0.11
SDSS J122003.72+010632.0	12:20:03.73	01:06:32.4	Galaxy	0.315	Red	0.10	0.25
[VV2006] J122130.9+010727	12:21:30.97	01:07:28.1	QSO	1.370	Blue	0.61	0.13
Gaia EDR3 3521773745637847552	12:21:34.41	−14:57:50.5	Star	–	Blue	0.05	0.04
2SLAQ J122421.12+002354.1	12:24:21.13	00:23:54.4	QSO	0.334	Blue	0.46	0.24
[VV2006] J122625.7+011604	12:26:25.67	01:16:04.6	QSO	2.478	Blue	0.34	0.12
2SLAQ J122641.43-002005.1	12:26:41.45	−00:20:05.1	Seyfert 1	0.353	Blue	0.48	0.18
GALEX 2414740977348515009	12:32:36.17	−03:18:39.4	Blue	–	Blue	0.49	0.13
[DCD2013] CSS J123702.3-151643	12:37:02.41	−15:16:43.5	RRLyr	–	Blue	0.66	0.13
6dFGS gJ125440.4-142244	12:54:40.42	−14:22:44.4	Galaxy	0.861	Blue	1.00	0.00
2MASS J12551223-1814542	12:55:12.23	−18:14:54.1	X	–	Blue	0.34	0.15
HE 1256-1805	12:58:42.99	−18:21:36.5	Unknown Candidate	0.014	Blue	0.76	0.11
PSO J194.7124-18.9084	12:58:50.98	−18:54:30.5	QSO	3.255	Blue	0.91	0.09
CRTS J125900.8-133442	12:59:00.82	−13:34:42.0	CV* Candidate	–	Blue	0.38	0.15
[VV96] J125914.0-192508	12:59:14.03	−19:25:08.4	QSO	1.140	Blue	0.63	0.11
[VV96] J130243.5-135553	13:02:43.59	−13:55:52.8	QSO	1.391	Blue	0.64	0.20
[VV2006] J131712.7-000200	13:17:12.74	−00:01:59.4	QSO	1.360	Blue	0.68	0.13
2SLAQ J131957.59-003446.7	13:19:57.60	−00:34:46.6	Star	–	Blue	0.91	0.09
SDSS J132023.46-004730.9	13:20:23.47	−00:47:30.8	QSO	3.255	–	–	–
[SHM2017] J200.93368-12.05326	13:23:44.09	−12:03:11.8	RRLyr	–	Blue	0.61	0.08
CRTS J132418.4-114734	13:24:18.48	−11:47:34.3	RRLyr	–	Blue	0.93	0.07
ATO J201.6140-13.6964	13:26:27.37	−13:41:47.0	V*	–	Red	0.07	0.75
6dFGS gJ132652.1-150639	13:26:52.11	−15:06:38.4	Galaxy	0.323	Red	0.08	0.17
BPS CS 22889-0007	13:31:59.47	−09:53:02.6	RRLyr	0.001	Blue	0.68	0.11
NVSS J133618-072251	13:36:18.64	−07:22:51.8	Radio	–	Blue	0.68	0.24
BRI B1335-0417	13:38:03.41	−04:32:34.7	Galaxy	4.396	Red	0.02	0.03
GALEX 2697385722761974216	13:39:09.19	−08:19:40.8	Blue	–	Blue	0.55	0.07
[DCD2013] CSS J134330.9-151858	13:43:31.01	−15:18:58.9	RRLyr	–	Blue	0.63	0.10
Gaia 18dwd	13:46:39.20	−09:38:36.0	Transient	–	Blue	0.71	0.09
GALEX 2697315366902694354	13:47:49.82	−04:10:10.6	Blue	–	Blue	0.62	0.12
GALEX 2699039396840082228	13:50:33.33	−12:16:42.9	Blue	–	Blue	0.53	0.17
QSO B1352-104	13:54:46.53	−10:41:02.6	QSO	0.330	Blue	0.28	0.14
[VV2006] J135602.8-022624	13:56:02.79	−02:26:23.3	QSO	1.373	–	–	–
LCRS B135623.8-061854	13:59:01.29	−06:33:27.0	Galaxy	–	Red	0.12	0.32
2MASS J14265388+0525172	14:26:53.89	05:25:17.4	QSO	0.323	Blue	0.44	0.26
[LAM2019] J1428+0500 B	14:28:55.39	05:00:21.9	lensImage Candidate	–	Blue	0.50	0.18
GALEX 2429518413625830432	14:28:55.46	05:00:19.9	Blue	–	Blue	0.57	0.15
SDSS J145344.51+045645.8	14:53:44.52	04:56:46.0	QSO	3.328	Blue	0.63	0.11
SDSSCGB 43444.3	14:55:33.70	04:46:43.2	AGN	0.334	Red	0.37	0.47

Table B1: –continued

Id Object	RA	Dec	Type	Redshift	Group HAC	P(Blue) HDBSCAN	P(Red) HDBSCAN
CRTS J145640.1-211601	14:56:40.15	−21:16:01.1	EB*	–	Red	0.22	0.42
ATO J299.9677+00.4034	19:59:52.25	00:24:12.7	EB* Candidate	–	Red	0.18	0.49
ATO J300.2623+00.3134	20:01:02.96	00:18:48.5	V*	–	Red	0.00	1.00
SDSS J200143.74+004918.4	20:01:43.73	00:49:18.4	QSO	–	Blue	0.37	0.21
ATO J300.4738+01.2397	20:01:53.72	01:14:23.1	EB* Candidate	–	Blue	0.89	0.11
SDSS J200432.38+001041.3	20:04:32.39	00:10:41.4	low-mass*	–	–	–	–
[SHM2017] J302.70083-00.21773	20:10:48.20	−00:13:03.9	RRLyr	–	Blue	1.00	0.00
ATO J304.2000+00.9027	20:16:48.00	00:54:09.9	EB* Candidate	–	Blue	0.43	0.27
ATO J305.6479-00.6694	20:22:35.51	−00:40:09.9	RRLyr	–	Blue	0.66	0.12
ATO J305.6574-00.0473	20:22:37.80	−00:02:50.5	RRLyr	–	Blue	0.96	0.04
SDSS J202906.80+005453.5	20:29:06.81	00:54:53.6	QSO	–	Blue	0.79	0.11
SDSS J203521.96-011413.5	20:35:21.96	−01:14:13.4	low-mass*	–	Red	0.09	0.48
[SHM2017] J309.71625+00.24532	20:38:51.91	00:14:43.1	RRLyr	–	Blue	0.75	0.09
2SLAQ J204340.03+002853.4	20:43:40.04	00:28:53.6	Seyfert 1	0.317	Blue	0.48	0.22
SDSS J204626.10+002337.7	20:46:26.11	00:23:37.8	QSO	0.332	Red	0.10	0.38
2SLAQ J204720.76+000007.7	20:47:20.76	00:00:07.7	CataclyV*	0.001	Blue	0.40	0.13
2SLAQ J204910.96+001557.2	20:49:10.95	00:15:57.5	Seyfert 1	0.363	Blue	0.42	0.23
[VV2006] J204956.6-001201	20:49:56.62	−00:12:01.7	QSO	0.369	Blue	0.37	0.21
CRTS J205007.0-002119	20:50:07.00	−00:21:18.5	EB*	–	Blue	0.96	0.04
SDSS J205352.03-001601.5	20:53:52.04	−00:16:01.5	QSO	0.363	Blue	0.51	0.21
2SLAQ J205614.55-004050.9	20:56:14.55	−00:40:50.6	Star	–	–	–	–
SDSS J205703.28+000945.9	20:57:03.28	00:09:45.8	low-mass*	–	Red	0.08	0.17
2SLAQ J205712.69+001211.3	20:57:12.69	00:12:11.4	QSO	0.335	Blue	0.45	0.25
Gaia EDR3 6794425304909258752	20:58:06.45	−30:08:18.1	WD* Candidate	–	Blue	0.29	0.11
CRTS J205942.7-214038	20:59:42.85	−21:40:37.6	EB*	–	Blue	0.65	0.20
6dFGS gJ205957.5-213935	20:59:57.53	−21:39:34.9	Galaxy	-0.001	Blue	0.58	0.14
SDSS J210014.12+004446.0	21:00:14.11	00:44:45.9	CataclyV*	0.000	Blue	0.36	0.13
QSO B2059-330	21:02:41.71	−32:52:44.1	QSO	3.280	Blue	0.52	0.09
Gaia EDR3 6808104805812408064	21:03:56.66	−21:47:27.1	Star	–	Blue	0.79	0.08
[GPM2009] J2104-0035 2	21:04:55.31	−00:35:21.8	EmG	0.005	Blue	0.31	0.11
[VV2006] J210514.1-004326	21:05:14.04	−00:43:26.4	QSO	3.250	Blue	0.79	0.03
PN G006.0-41.9	21:05:53.57	−37:08:40.4	PN	–	Blue	0.04	0.03
EC 21035-4032	21:06:48.02	−40:20:03.7	Star	–	Blue	0.15	0.07
CRTS J210728.9-373104	21:07:28.93	−37:31:03.0	RRLyr	–	Blue	0.69	0.10
[GPM2009] J2112-0016 1	21:12:00.92	−00:16:49.2	EmG	0.012	Blue	1.00	0.00
2MASS J21122459-4128534	21:12:24.59	−41:28:53.3	AGN	0.349	Blue	0.49	0.16
CRTS J211328.1+000332	21:13:28.17	00:03:32.6	EB*	–	Blue	0.95	0.05
CRTS J211639.4+010627	21:16:39.48	01:06:27.3	EB*	–	Blue	0.90	0.10
CRTS J211751.2-343932	21:17:51.29	−34:39:30.1	RRLyr	–	Blue	0.98	0.02
1RXS J211805.2-341343	21:18:04.28	−34:13:43.3	CataclyV*	–	Blue	0.62	0.12
SDSS J212225.23-005327.0	21:22:25.23	−00:53:27.1	low-mass*	-0.000	Red	0.13	0.46
AT20G J212302-291504	21:23:02.82	−29:15:04.0	Radio(cm)	–	Blue	0.55	0.16
CRTS J212609.1+011147	21:26:09.16	01:11:47.9	EB*	–	Blue	1.00	0.00
CRTS J212654.5-012054	21:26:54.54	−01:20:54.1	CV* Candidate	–	Blue	0.24	0.14
LBQS 2128-4555	21:31:29.53	−45:41:50.5	QSO	0.623	Blue	0.58	0.14
2MASS J21333817+0126291	21:33:38.14	01:26:29.0	QSO	1.010	Blue	0.93	0.07
SDSS J213455.08+001056.9	21:34:55.09	00:10:56.8	QSO	3.289	Blue	0.62	0.12
WiggleZ R22J213601526-03023619	21:36:01.52	−03:02:36.3	Galaxy	3.239	–	–	–
WISEA J213649.75-012852.2	21:36:49.75	−01:28:52.2	QSO	3.280	Blue	0.64	0.11
CRTS J213712.6-223229	21:37:12.63	−22:32:28.1	RRLyr	–	Blue	0.90	0.10
2MASS J21381896+0112224	21:38:18.96	01:12:22.5	Seyfert 1	0.344	Blue	0.32	0.19
CRTS J213937.6-023913	21:39:37.58	−02:39:13.0	CV* Candidate	–	Blue	0.27	0.17
SN 2017hxv	21:44:22.94	−29:54:59.0	SN	0.019	Red	0.09	0.12
6dFGS gJ214540.0-291937	21:45:40.01	−29:19:36.9	Galaxy	0.341	Red	0.09	0.20
2SLAQ J214830.60-004752.6	21:48:30.61	−00:47:52.5	EmG	0.332	–	–	–
SDSS J215002.70+011343.8	21:50:02.70	01:13:43.8	QSO	3.267	Blue	0.55	0.06
2MASS J21501054-0010002	21:50:10.53	−00:10:00.6	QSO	0.335	Blue	0.45	0.25
SDSS J220242.61-012528.0	22:02:42.61	−01:25:28.1	QSO	1.376	Blue	0.49	0.11

Table B1: –continued

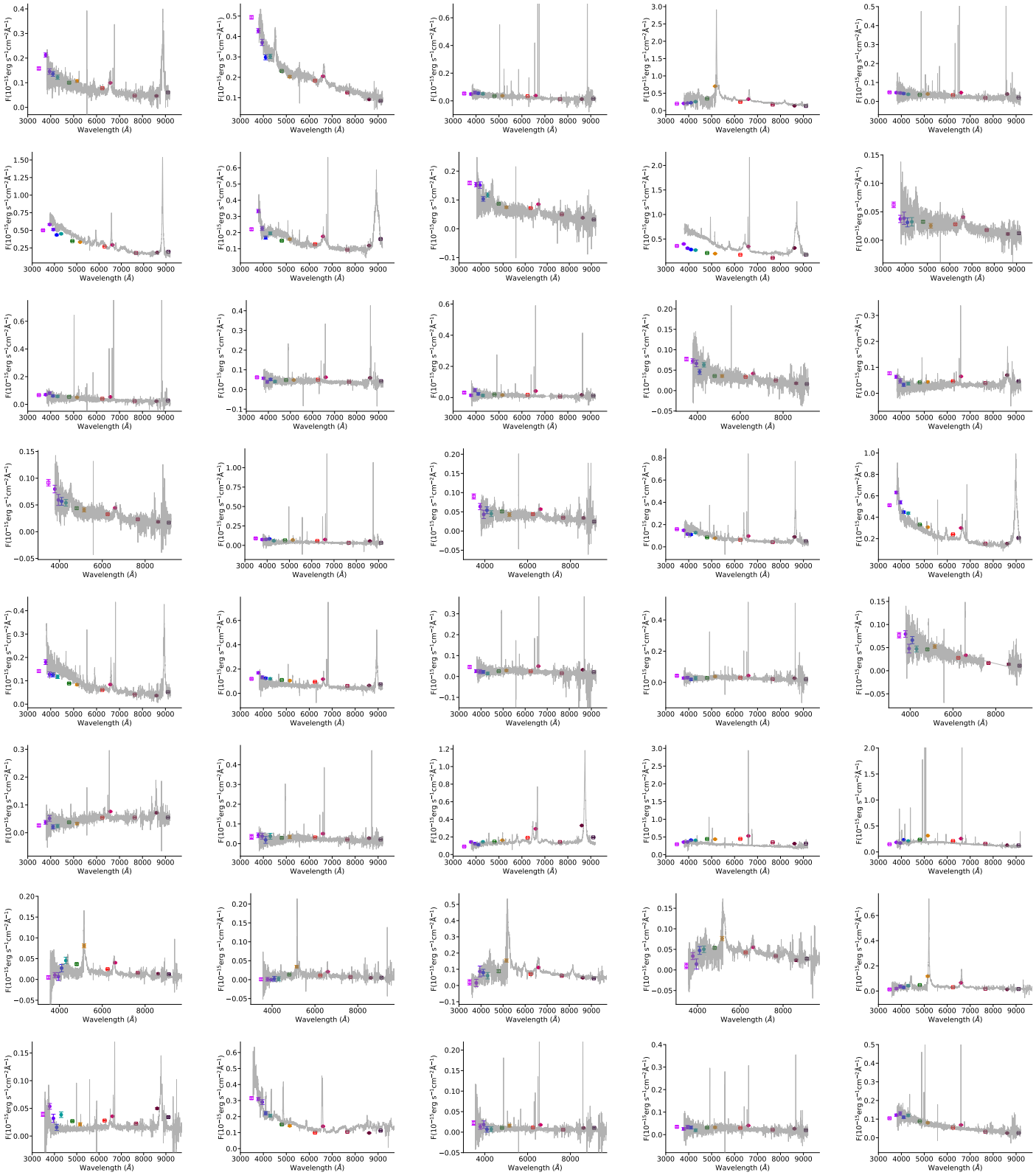
Id Object	RA	Dec	Type	Redshift	Group HAC	P(Blue) HDBSCAN	P(Red) HDBSCAN
PB 5049	22:03:15.14	01:17:21.0	Star	–	Blue	0.06	0.04
SDSS J220529.34-003110.6	22:05:29.34	−00:31:10.7	QSO	2.454	Blue	0.40	0.12
2MASS J22085196-0106038	22:08:51.97	−01:06:03.7	QSO	0.351	Blue	0.42	0.21
2dFGRS TGS061Z180	22:09:19.05	−24:07:12.4	QSO	0.320	Red	0.09	0.16
2QZ J220948.6-301357	22:09:48.63	−30:13:55.8	WD*	–	Blue	0.08	0.04
SDSS J220954.57-012717.6	22:09:54.57	−01:27:17.6	QSO	3.296	Blue	0.66	0.06
2QZ J221000.7-311400	22:10:00.75	−31:14:00.0	EmG	0.328	Blue	0.49	0.14
2QZ J221005.7-275439	22:10:05.76	−27:54:38.7	Galaxy	0.330	Blue	0.80	0.06
2QZ J221058.3-273930	22:10:58.33	−27:39:29.4	Galaxy	0.313	Blue	0.76	0.13
[VV2006] J221335.7-282542	22:13:35.65	−28:25:41.7	QSO	2.469	Blue	0.30	0.12
[VV2006] J221532.6-281805	22:15:32.58	−28:18:03.9	QSO	1.330	Blue	0.62	0.12
SDSS J221546.92-015906.6	22:15:46.92	−01:59:06.6	QSO	1.361	Blue	0.53	0.08
6dFGS gJ221706.5-303447	22:17:06.52	−30:34:46.1	Galaxy	0.337	Red	0.19	0.34
SDSS J221722.45+010436.3	22:17:22.44	01:04:36.3	QSO	1.403	Blue	0.07	0.05
2QZ J221819.4-271544	22:18:19.39	−27:15:44.2	Seyfert 1	0.355	Blue	0.45	0.25
2QZ J221945.1-293414	22:19:45.08	−29:34:13.4	EmG	0.343	Blue	0.92	0.08
2QZ J222113.6-280421	22:21:13.62	−28:04:20.9	Seyfert 1	0.332	Red	0.20	0.33
2QZ J222336.0-283140	22:23:36.03	−28:31:39.6	EmG	0.333	Blue	0.63	0.07
2MASS J22240340-0057241	22:24:03.35	−00:57:24.2	QSO	0.313	Blue	0.37	0.18
2QZ J222416.2-292421	22:24:16.26	−29:24:21.7	CataclyV*	–	Blue	0.26	0.09
2MASX J22263365-2917276	22:26:33.64	−29:17:28.0	Galaxy	0.003	Red	0.15	0.49
2SLAQ J222825.11-002217.4	22:28:25.12	−00:22:17.2	Galaxy	–	Red	0.47	0.22
SDSS J222956.54+003126.3	22:29:56.54	00:31:26.5	QSO	1.340	Blue	0.44	0.14
2QZ J223114.0-312005	22:31:13.95	−31:20:04.4	Star	–	Blue	0.66	0.18
[VV2006] J223251.7-303250	22:32:51.74	−30:32:49.6	QSO	0.350	Blue	0.26	0.12
2QZ J223342.5-301936	22:33:42.56	−30:19:35.4	Galaxy	0.324	Blue	0.49	0.13
FASTT 1560	22:34:39.93	00:41:27.5	CataclyV*	0.001	Blue	0.49	0.21
2SLAQ J223543.05-005436.5	22:35:43.05	−00:54:36.6	Galaxy	–	Blue	0.59	0.14
[VV2006] J223633.5+002652	22:36:33.54	00:26:52.8	QSO	1.354	Blue	0.65	0.18
SDSS J223649.60+005413.5	22:36:49.60	00:54:13.8	QSO	3.313	–	–	–
PHL 354	22:38:23.25	−00:57:08.2	QSO	0.361	Blue	0.33	0.15
2SLAQ J223844.30-005655.3	22:38:44.29	−00:56:55.3	QSO	1.357	Blue	0.46	0.15
2SLAQ J224531.20-004509.4	22:45:31.20	−00:45:09.3	QSO	1.368	Blue	0.21	0.09
SDSS J224539.94-002419.7	22:45:39.94	−00:24:19.6	QSO	3.280	Blue	0.93	0.07
2MASS J22495608+0002182	22:49:56.08	00:02:18.4	QSO	3.307	Blue	0.72	0.08
SDSS J225149.74-002811.7	22:51:49.75	−00:28:11.4	QSO	3.228	Blue	0.93	0.07
2QZ J225157.1-292451	22:51:57.10	−29:24:50.8	EmG	0.318	Blue	0.45	0.15
2SLAQ J225257.45+002731.5	22:52:57.44	00:27:31.6	Star	–	Blue	0.23	0.09
2QZ J225352.9-300944	22:53:52.96	−30:09:43.7	Seyfert 1	0.326	Blue	0.19	0.15
[VV2006] J225411.2-312712	22:54:11.15	−31:27:11.3	QSO	1.360	Blue	0.63	0.11
SDSS J225411.96-004949.5	22:54:11.96	−00:49:49.4	QSO	3.297	Blue	0.31	0.09
2QZ J225908.1-312717	22:59:08.12	−31:27:16.7	Star	–	Blue	0.23	0.09
2SLAQ J230201.20+003047.2	23:02:01.20	00:30:47.3	QSO	1.344	Blue	0.49	0.13
[VV2006] J230235.5-285630	23:02:35.44	−28:56:29.7	QSO	0.368	Blue	0.52	0.18
2SLAQ J230316.40-001211.5	23:03:16.41	−00:12:11.4	QSO	1.516	Blue	0.42	0.13
V* HY Psc	23:03:51.63	01:06:51.4	CataclyV*	-0.000	Blue	0.70	0.22
SDSS J230428.31+005701.2	23:04:28.34	00:57:01.2	QSO	0.317	Blue	0.36	0.19
2SLAQ J230444.16-010251.7	23:04:44.16	−01:02:51.5	QSO	1.377	Blue	0.40	0.15
SDSS J230855.49+003705.6	23:08:55.49	00:37:05.7	QSO	1.784	Blue	0.58	0.09
[VV2006] J230914.4-305913	23:09:14.31	−30:59:12.5	QSO	1.380	Blue	0.62	0.12
2MASS J23094616+0000496	23:09:46.16	00:00:49.0	Seyfert 1	0.352	Blue	0.46	0.20
[VV2006] J231135.1-312644	23:11:35.12	−31:26:44.1	QSO	1.350	Blue	0.52	0.13
2SLAQ J231231.36-011137.5	23:12:31.36	−01:11:37.3	QSO	1.360	Blue	0.58	0.14
SDSS J231259.07+010805.6	23:12:59.06	01:08:05.9	QSO	3.295	Blue	0.61	0.13
[VV2006] J231311.9-004538	23:13:11.91	−00:45:38.0	QSO	1.364	Blue	0.53	0.09
[VV2006] J231519.4-303857	23:15:19.39	−30:38:57.2	QSO	1.356	Blue	0.56	0.14
V* CC Scl	23:15:31.78	−30:48:48.7	CataclyV*	–	Blue	0.43	0.15
[VV2006] J231652.0+005125	23:16:52.04	00:51:25.9	QSO	3.229	Blue	0.62	0.11

Table B1: –continued

Id Object	RA	Dec	Type	Redshift	Group HAC	P(Blue) HDBSCAN	P(Red) HDBSCAN
3XMM J231742.5+000535	23:17:42.61	00:05:35.3	Seyfert 1	0.321	Blue	0.49	0.18
[VV2006] J231942.8-302629	23:19:42.76	−30:26:29.5	QSO	2.473	Blue	0.39	0.12
2QZ J232126.5-310730	23:21:26.51	−31:07:29.5	Galaxy	0.309	Blue	1.00	0.00
CRTS J232435.1-000212	23:24:35.18	−00:02:11.8	EB*	–	Blue	0.86	0.05
2SLAQ J232457.75+002153.2	23:24:57.75	00:21:53.4	QSO	0.345	Blue	0.14	0.11
CRTS J232551.5-014024	23:25:51.48	−01:40:23.8	CataclyV*	–	Blue	0.24	0.19
[VV2006c] J232555.5-003710	23:25:55.51	−00:37:10.7	Seyfert 1	0.332	–	–	–
SDSS J233104.38-004237.2	23:31:04.40	−00:42:37.1	QSO	1.353	Blue	0.41	0.13
2QZ J233254.8-305844	23:32:54.78	−30:58:43.8	EmG	0.329	Blue	0.47	0.16
SDSS J233300.21-002030.5	23:33:00.22	−00:20:30.5	QSO	3.328	Blue	1.00	0.00
CRTS J233408.7+002704	23:34:08.78	00:27:04.6	low-mass*	–	Red	0.16	0.37
[VV2006] J233438.5+002341	23:34:38.55	00:23:41.9	QSO	1.385	Blue	0.54	0.17
2SLAQ J233522.69-000635.2	23:35:22.69	−00:06:35.2	QSO	1.373	Blue	0.52	0.17
[VV2006] J233722.0+002239	23:37:22.02	00:22:39.2	QSO	1.377	Blue	0.55	0.16
2XMM J233731.7+002559	23:37:31.79	00:25:59.9	AGN	0.314	Blue	0.44	0.26
[VV2006] J234329.1-300200	23:43:29.16	−30:02:00.1	QSO	1.358	Blue	0.53	0.17
2SLAQ J234440.53-001205.8	23:44:40.53	−00:12:06.1	CataclyV*	–	Blue	0.40	0.21
PB 5574	23:53:14.82	−00:18:20.5	WD*	-0.000	Blue	0.07	0.04
[VV2006] J235546.2-002342	23:55:46.14	−00:23:42.8	QSO	3.245	Blue	0.56	0.08
[VV2006] J235718.4+004350	23:57:18.37	00:43:50.5	QSO	4.366	Red	0.04	0.08
SDSS J235805.25-012153.9	23:58:05.25	−01:21:53.9	QSO	1.368	Blue	0.82	0.11

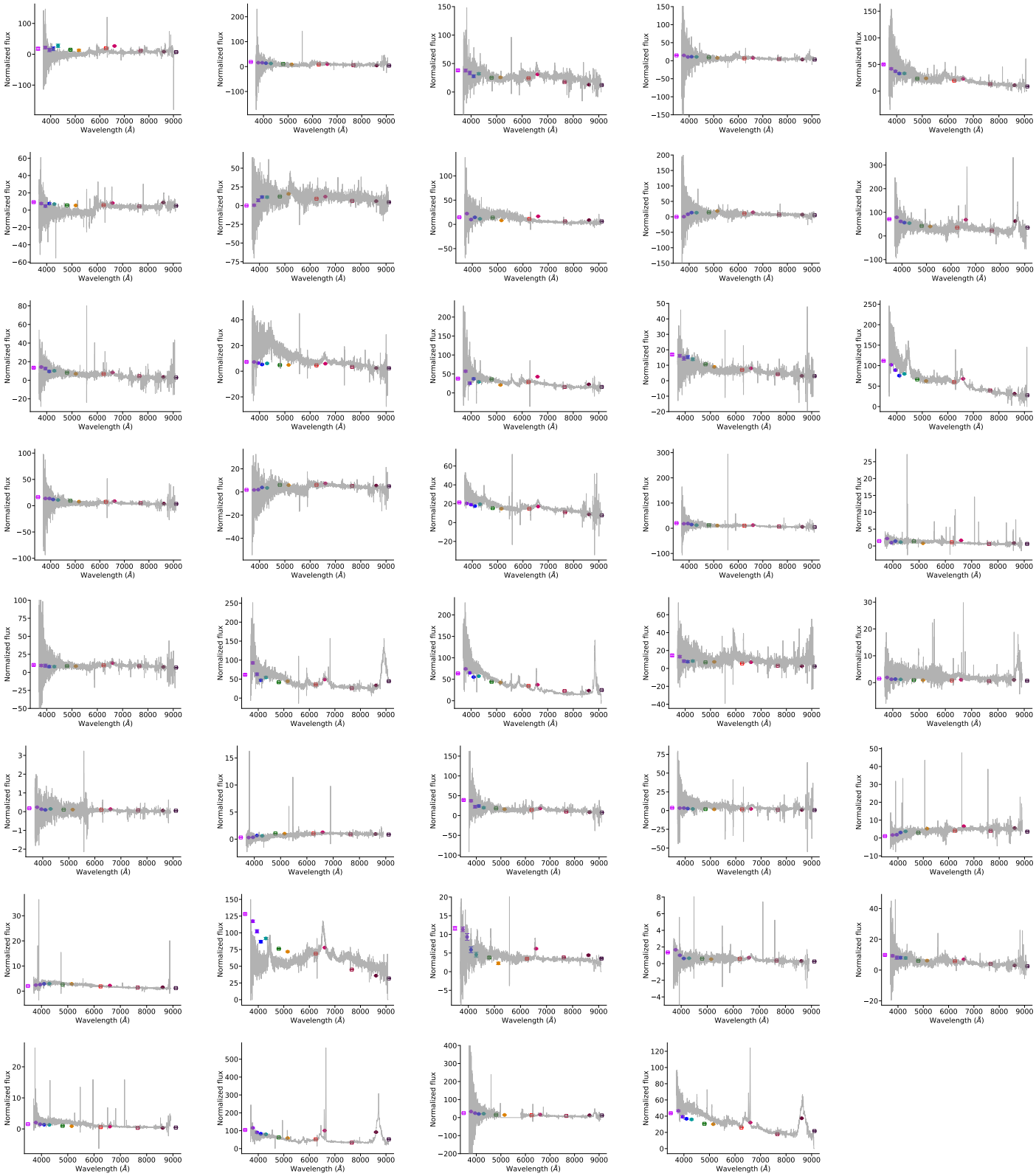
APPENDIX C: SDSS SPECTRA

Table C1: Espectra from SDSS DR16



APPENDIX D: LAMOST SPECTRA

Table D1: Espectra from LAMOST DR6



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