Report on: H-alpha emitters from the Southern Photometric Local Universe Survey (S-PLUS)

By Gutierrez-Soto et al.

This paper is about the selection and further analysis of candidate H-alpha emission line objects within DR3 of S-PLUS, a survey that is still underway. The first part of the work is very much along the lines of earlier papers, based on other surveys, that have used the combination of a narrow-band H-alpha filter with broadband r and I (most notably Witham et al 2008 and Wevers et al 2017) to pull out candidate emission line objects. With a selection made, attention then turns to analysis of the r-z,g-r colour-colour diagram, bringing in two different clustering-analysis techniques – it is claimed that these provide robust support for splitting the candidates into 'red' and 'blue' classes. The point of this is unclear, and not so strongly connected to what we should expect this particular choice of colour-colour diagram to tell us. The paper finishes with some cross-matching with SDSS and SIMBAD and a discussion of the different classes of emission line objects (which includes emission-line galaxies and low-redshift QSOs that the relatively red and broad H-alpha filter was no doubt designed for).

In its present form, the paper has the following problems: it is being presented before the publication of the data release it uses (the smallest major issue); the selection of emission-line candidates suffers from both a possible data problem and questionable errors; it fails to establish much value for the extensively-described clustering analysis (why the focus on just the r-z,g-r diagram?). It is also on the verbose side, for content. I conclude that rejection is the best course at this time, on grounds of a potentially flawed approach, and a lack of significance as presented. The authors should reconsider what they are trying to do here and produce a more focused, tightly-argued contribution. Below I give more detail on the three main points.

1. Publication before the DR3 release paper.
   This is a problem because the reader is left, in section 2.1, needing to wade through a bit of a forest of earlier works looking for enlightenment on the particular data this paper uses. I've tried, but don't feel I should have to try so much. From digging, the range in psf widths is from 0.8 to 2 arcsec (sampled with 0.55 arcsec pixels, right)? Are all four relevant filters exposed contemporaneously (i.e. consecutively on the same night)? All from the 'main survey'?

2. Data selection and errors.
   Figures 2 and 5 give cause for concern that the text does not follow up and resolve. This along with my next major point are the most important ones. The content of Figure 2 suggests to me that there is a problem with background subtraction (most likely in the narrowband) that the authors have ignored. In all but the faintest magnitude range, where the much larger errors mask the problem, there is evidently an unphysical spread of datapoints above and below the main stellar locus in what I suspect is the most densely populated part of the locus (in the region of $r – i \sim 0.3$). In the two intermediate magnitude ranges, it would seem that objects belonging to this plume above the main locus are liable to be selected as candidate emission line sources. It should be remembered that outliers in any photometric diagram are there because they are real or affected by untracked data problems. Would many of these objects also be in the parts of the sky shown in Figure 5 where the densities of candidates are (implausibly?) high?

This issue is potentially made worse by the authors choosing to use a 4-sigma cut in their selection, and by adopting a form for the error (in equation 1) that is likely to be too optimistic. I deduce that the authors are copying over the choice of a 4-sigma cut from Wevers et al (2017), for whom this was a very safe choice (see their figure 3 and notice how they cut, well clear of the main locus). The form of error misses the dependence on the slope of the regression which implicates the error in r – i, as well as in r – j0660 (see equation 10 in Monguio et al 2020, A&A, 638, A18 and the discussion about other terms that can enter). A question the authors should be asking of themselves, in respect of Figure 5 is why is b > 0 so well off for candidates and why is l ~ 40, b ~ -15 favoured also. A check that could be tried is to make a map (or histograms) of the fraction of all sources that are selected as candidates (see bottom panel in Figure 3 in Witham et al 2008, Kohoutek & Wehmeyer did something similar in their 1997 Hamburg catalogue).

3. Clustering analysis of the r-z,g-r diagram.
   Why is so much attention lavished on this particular colour-colour plane? Astrophysically, it is not that information-rich for the simple reason that, for stars, the reddening vector has the same sense of impact on position in the plane as intrinsic colour (see discussion of the analogous r-i,g-r diagram by Monguio et al – their Figure 14). The main deviation you expect comes from M stars that create the upward hook (at lower reddening for dwarfs than for reddened giants). The very crammed Figure 6 in this paper shows galaxies, BCGs, do much the same. Did the authors consider examining a higher dimensionality, using more filter information? One might expect machine learning techniques to make more powerful discriminations if deployed in this way. It is absolutely no surprise that there are 'blue' and 'red' emission line objects – all the stellar/Galactic surveys have seen this, and it is well-known that e.g. CVs are blue, nearby, little reddened objects, that QSOs have emission lines on top of (usually) 'blue' power-law spectra, and YSOs are red because they are reddened and commonly intrinsically red anyway. It is not shown, but presumably this colour plane (as presented in Figure 7) doesn't look too different if you were to plot unselected stars.
   Ultimately, the paper just doesn't manage to make a compelling case that the blue/red distinction being made in r-z,g-r is an important new classification. The bimodality that the conclusion section brings up as a finding is already apparent in Figure 7 – particularly in the g-r histogram. Machine learning is not needed to uncover this. Section 3 hardly makes any use of the HAC and HDBSCAN results – it is mainly about cross-matching with SDSS for spectra and (the sometimes flakey) SIMBAD for classification. There seems to be a nascent plan to write **another** paper on spectroscopic cross-match – in my view, it probably should be here and not delayed, given the low value of the clustering analysis.

None of the above is about verbosity, and indeed there are a number of smaller issues in e.g. Section 3 that could be itemized. On the latter, I'll just mention I was surprised by the need for a 2 arcsec cross-match distance wrt SDSS (would need a comment on the median distance and/or on what might be going on), and more on the bias that would likely carry through (SDSS spectroscopy database heavily biased to galaxies?). On verbosity, I would just note that the introduction looks like a primer on all the various types of emission line objects – in contrast, it is short on the motivation behind this study (why off the Galactic Plane is an important hunting ground for emission line objects). Other papers of this type have not found it necessary to be didactic at such length for the good reason that most readers of the paper will already have a knowledge of the basics. The amount of space given to the HAC and HDBSCAN algorithms is perhaps overlong too.