Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

Mon. Not. R. Astron. Soc. **399,** 2231–2238 (2009) doi:10.1111/j.1365-2966.2009.15426.x

Classifying broad absorption line quasars: metrics, issues and a new catalogue constructed from SDSS DR5

S. Scaringi,[1]★ C. E. Cottis,[2] C. Knigge[1] and M. R. Goad[2]

[1]Department of Physics and Astronomy, University of Southampton, Highfield, Southampton SO17 1BJ

[2]Department of Physics and Astronomy, University of Leicester, University Road, Leicester LE1 7RH

Accepted 2009 July 19. Received 2009 July 13; in original form 2008 July 4

ABSTRACT

We apply a recently developed method for classifying broad absorption line quasars (BALQSOs) to the latest quasi-stellar object (QSO) catalogue constructed from Data Release 5 of the Sloan Digital Sky Survey. Our new hybrid classification scheme combines the power of simple metrics, supervised neural networks and visual inspection. In our view, the resulting BALQSO catalogue is both more complete and more robust than all previous BALQSO catalogues, containing 3552 sources selected from a parent sample of 28 421 QSOs in the redshift range $1.7 < z < 4.2$. This equates to a raw BALQSO fraction of 12.5 per cent.

In the process of constructing a robust catalogue, we shed light on the main problems encountered when dealing with BALQSO classification, many of which arise due to the lack of a proper physical definition of what constitutes a BAL. This introduces some subjectivity in what is meant by the term BALQSO, and because of this, we also provide all of the meta-data used in constructing our catalogue, for every object in the parent QSO sample. This makes it easy to quickly isolate and explore subsamples constructed with different metrics and techniques. By constructing composite QSO spectra from subsamples classified according to the meta-data, we show that no single existing metric produces clean and robust BALQSO classifications. Rather, we demonstrate that a variety of complementary metrics are required at the moment to accomplish this task. Along the way, we confirm the finding that BALQSOs are redder than non-BALQSOs and that the raw BALQSO fraction displays an apparent trend with signal-to-noise ratio steadily increasing from 9 per cent in low signal-to-noise ratio data up to 15 per cent.

Key words: catalogues – surveys – quasars: absorption lines.

1 INTRODUCTION

Broad absorption line quasars (BALQSOs) are a subclass of active galactic nuclei (AGN) exhibiting strong, broad and blue-shifted spectroscopic absorption features (Foltz et al. 1990; Weymann et al. 1991; Hewett & Foltz 2003; Reichard 2003b). These features are thought to be formed in fast ($0.1c$-$0.2c$) and powerful outflows from the accretion disc around the supermassive black hole at the heart of the AGN (Korista 1992). The vast majority of BALQSOs are radio-quiet (Stocke et al. 1992, but see Brotherton, de Breuck & Schaefer 2006 for some counter examples), and there are subtle differences between their continuum and emission-line properties and those of "normal" (non-BAL) quasi-stellar object (QSOs) (Reichard 2003b). However, despite these differences, BALQSOs and non-BALQSOs appear to be drawn from the same parent population (Reichard 2003b). Most BALQSOs belong to the subclass of the so-called HiBALs, only displaying absorption in certain high-ionization lines (e.g. N V 1240 Å, C IV 1549 Å, S IV 1397 Å). However, some, known as LoBALs, also show absorption in some low-ionization lines (most notably Mg II 2800 Å).

The most straightforward explanation for the differences between QSOs and BALQSOs is a simple orientation effect. Thus, *all* QSOs may undergo significant mass loss through winds (Ganguly & Brotherton 2007), but BALs are only observed if the central continuum and/or emission line source is viewed directly through the outflowing material. Viewed in this context, BALQSOs may be the only available tracers of a key physical process common to all AGN. Also, the powerful outflows we observe in BALQSOs are an important example of AGN feedback in action (Tremonti, Moustakas & Diamond-Stanic 2007). Such feedback is a key ingredient required in theoretical attempts to understand Galaxy 'downsizing' and may also be responsible for regulating the growth of supermassive black holes. Moreover, the fraction of QSOs

★E-mail: simo@astro.soton.ac.uk

displaying BAL features ($f_{\rm BALQSO}$) may provide a direct estimate of the opening angle of these outflows.

Historically, BALQSO samples have been selected on the basis of the so-called balnicity index (BI; Weymann et al. 1991) or similar metrics. These samples consistently yielded BALQSO fraction estimates in the range $f_{\rm BALQSO} \approx 0.10$–$0.15$ (Weymann et al. 1991; Tolea, Krolik & Tsvetanov 2002; Hewett & Foltz 2003; Reichard 2003a). In a previous paper (Knigge et al. 2008, hereafter Paper I), we showed that both the BI and a more recently defined metric, the absorption index (AI; Trump et al. 2006), are biased when selecting BALQSOs, the former being incomplete at the low-velocity end of the BALQSO distribution, and the latter suffering from significant contamination by objects with low-velocity absorption systems which may be unrelated to the higher velocity outflows.

Here, we use a combination of the classic BI metric, a simple neural network and visual inspection [the hybrid-learning vector quantization (LVQ) approach we developed in Paper I] to produce a BALQSO sample that is both more complete than purely BI-based ones and, importantly, significantly more robust than AI-based ones. We have applied our hybrid-LVQ algorithm to the QSO sample associated with Data Release 5 (DR5) of the Sloan Digital Sky Survey (SDSS; Adelman-McCarthy et al. 2007; Schneider et al. 2007) using the BIs calculated from Gibson et al. (2009). The resulting catalogue contains 3552 BALQSOs selected from a parent sample of 28 421 QSOs on the basis of absorption close to the C IV high-ionization emission line. This catalogue may be obtained from http://www.astro.soton.ac.uk/~simo. A preliminary version of the catalogue has already been presented in Scaringi et al. (2008). In addition, we also provide (at the same address) a catalogue of the meta-data, i.e. the data pertaining to the parent QSO sample and subsequently used in the compilation of our BALQSO catalogue, so that members of the scientific community wishing to compile their own BAL/non-BAL subsamples may readily do so.

2 DATA AND METHODS

2.1 The QSO parent population

The SDSS DR5 QSO catalogue contains over 77 000 objects in total (Schneider et al. 2007). However, for the purpose of constructing a uniform BALQSO catalogue, we only consider objects whose spectra fully cover the C IV 1550 Å resonance line and its associated absorption region (up to 29 000 km s^{-1} blueward of the C IV line centre), which displays a particularly deep and well-defined absorption trough in the spectra of most BALQSOs. Given the wavelength range covered by the SDSS spectra, this implies an effective redshift window of $1.7 < z < 4.2$ for our QSO parent sample. This redshift window yields spectra for a QSO parent sample of 28 421 objects. This will be the parent sample used in this study to compile our BALQSO catalogue.

2.2 Metrics and preconditioning

Our BALQSO classification method works on continuum normalized spectra covering the wavelength range 1401–1700 Å with 1 Å dispersion. It also uses the associated BIs for training the neural network and to flag borderline cases requiring visual inspection. The BI metric is defined as

$$\mathrm{BI} = -\int_{25\,000}^{3000} \left[1 - \frac{f(v)}{0.9}\right] C\,\mathrm{d}v. \qquad (1)$$

Here, the limits of the integral are in units of km s^{-1}, and $f(v)$ is the normalized flux as a function of velocity displacement from line centre. The constant $C = 0$ everywhere, unless the normalized flux has satisfied $f(v) < 0.9$ continuously for at least 2000 km s^{-1}, at which point it is switched to $C = 1$ until $f(v) > 0.9$ again. Based on this definition, objects are classified as BALQSOs if their BI > 0 km s^{-1}. The BI by definition excludes strong, low-velocity absorption systems; for example, any deep absorption of width 3000 km s^{-1} which starts less than 2000 km s^{-1} blueward of the rest wavelength of the C IV emission line will be assigned BI = 0 km s^{-1}. Thus, BALQSO catalogues constructed using the BI metric are likely to be significantly incomplete at the low-velocity end of the distribution.

For this reason Hall et al. (2002) introduced the so-called AI, in an attempt to recover those low-velocity absorption system objects that were missed by the BI. The AI is defined as

$$\mathrm{AI} = \int_{0}^{29\,000} [1 - f(v)]\, C\,\mathrm{d}v, \qquad (2)$$

here now $C = 1$ in all regions where $f(v) > 0.9$ continuously for at least 1000 km s^{-1} and $C = 0$ otherwise. The two key differences that allow objects with BI = 0 km s^{-1} to achieve AI > 0 km s^{-1} are (i) that the AI includes regions within 3000 km s^{-1} of line centre (and also regions beyond 25 000 km s^{-1}) and (ii) that the AI includes objects with much narrower absorption troughs than the BI. The remaining differences are associated with the presence (absence) of the factor 0.9 in equations (1) and (2). The less stringent constraints imposed by the AI allow one to recover the majority of the low-velocity absorption systems missed by the BI, more than doubling the number of objects classed as BALQSOs. However, as shown in Paper I, the log-AI distribution is bi-modal, with low-velocity outflows preferentially occupying one mode and high-velocity outflows occupying the other. While it is beyond doubt that at least some of the BALQSOs classified solely by the AI are bona fide BALQSOs in the traditional sense, particularly in the region where the two modes overlap, it remains uncertain whether the two modes are physically connected. Thus, we cannot exclude the possibility that the AI includes substantial numbers of objects whose low-velocity absorption systems are unrelated to the high-velocity flows traditionally associated with the BALQSO phenomenon. Specific examples of hard-to-classify BALQSO spectra selected using either the AI or BI may be found in Paper I.

The classification problems described above are illustrated in Fig. 1. The figure displays QSO geometric mean composites created using the DR3 subset from our DR5 parent population normalized at 1750 Å.[1] More specifically, it shows the average properties of QSO spectra on a grid in AI/BI space, allowing a close examination of the absorption through the dependence on the AI and the BI. For reference we have also included in each panel the same non-BALQSO composite created from QSOs with AI = 0 km s^{-1} (dashed green curve).

It is generally clear from Fig. 1 that both the AI and the BI tend to select redder QSOs and that the troughs not only do get wider with increasing AI/BI but also get deeper. Moreover, Fig. 1 shows how QSO samples selected from the low-velocity region of the AI do not display the 'traditional' BALQSO properties. This is best shown in the second composite from the left panel (top) with BI = 0 and 1 km s^{-1} AI < 500 km s^{-1} , which shows little,

[1] We have used the DR3 subset so that we can use the AIs provided by Trump et al. (2006).

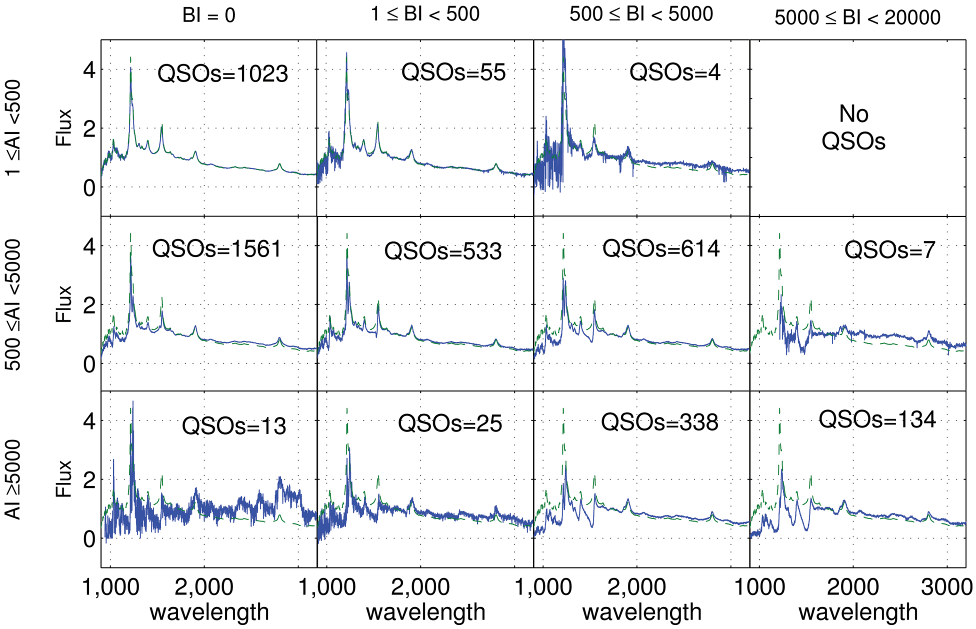

Figure 1. Composites in various AI–BI ranges (blue line), and composites created from AI = 0 km s^{-1} and BI = 0 km s^{-1} objects (dashed green line). The AI and BI bins on the side panels are in km s^{-1}. Reddening has not been taken into account.

if any, sign of absorption when compared to the non-BAL composite. The next panel down displays a composite created from 1561 QSOs (the largest subsample) which have 500 km s^{-1} < AI < 5000 km s^{-1} and BI = 0 km s^{-1}. Relative to the non-BAL composite, there is some evidence of absorption close to the C IV emission line. However, we caution that since the BAL composite is significantly redder than the non-BAL composite, identifying broad absorption lines in this spectrum is difficult, without first dereddening the spectrum. The remaining panels show spectra with increasingly prominent absorption in the vicinity of C IV, with the absorption strength (depth) and reddening increasing both with increasing AI (moving from top to bottom) and with increasing BI (left to right).

We conclude from examining Fig. 1 that using the AI > 0 km s^{-1} to select BALQSOs is unreliable, since QSOs with 0 km s^{-1} < AI < 5000 km s^{-1} and BI = 0 km s^{-1} have spectra that are not different from AI = 0 km s^{-1} non-BALQSOs. Moreover, most BALQSOs fall in the low-velocity region of both the AI and the BI continuum, which is also the region which turns out to be the hardest to classify. However, it is interesting to note that QSOs with AI > 5000 km s^{-1} and BI < 500 km s^{-1} do look like BALQSOs. We have decided to omit the AI metric from our hybrid classification method since about ≈50 per cent of the objects selected using this metric may not be genuine BALQSOs (see Paper I). Instead, our hybrid-LVQ method uses the BI, a simple neural network and visual inspection to select BALQSOs.

2.3 Hybrid-LVQ selection of BALQSOs

The method we use to classify BALQSOs has already been described in detail in Paper I, so we only provide an overview of the key points here. Briefly, our method is a hybrid of BI-based, neural network and visual classifications, and is designed to produce a more complete BALQSO sample than a pure BI selection, but without significantly increasing the number of false positives. Starting with a BI-based classification (as calculated by Gibson et al. 2009), we use a simple neural network-based machine learning algorithm called ‘learning vector quantization’ (Kohonen 2001) to identify objects that might have been misclassified by the BI. All such objects are then inspected and classified visually.

We caution that both the measured BI (and the AI for the reference) are very sensitive to our ability to perform an accurate fit to the underlying continuum. Overestimating the underlying continuum strength can yield a large positive AI and BI in the absence of any absorption. Conversely, if the continuum is underestimated, weak broad absorption features may go unrecognized. This is an issue which can also affect our hybrid-LVQ classification method. For this reason, we have decided to use the BI’s calculated from Gibson et al. (2009), since their continuum fitting algorithm is likely to be superior to the one we use for normalizing spectra for input into LVQ. This is mainly because they employ multiple composites in order to fit the underlying continuum (Trump et al. 2006).

For input into LVQ, we normalize all QSO spectra using the method described in Paper I and North, Knigge & Goad (2006), in which each spectrum is fitted with a modified DR3 QSO composite (constructed from objects with AI = 0 km s^{-1} as calculated from Trump et al. 2006) allowing for object-to-object differences in reddening and spectral index. We then bin each spectrum on to a uniform grid in wavelength and use the binned spectrum between 1401 and 1700 Å for our classification purposes.

The way we train our LVQ network to recognize BALQSOs has been described in detail in Paper I. In brief, we employ a training set composed of 400 BI > 0 km s^{-1} and 400 BI = 0 km s^{-1} QSOs and train our LVQ network to recognize BI > 0 km s^{-1} objects at first. We then visually inspect our neuron map for BALQSO misclassifications (locating BI > 0 km s^{-1} QSOs in BI = 0 km s^{-1} nodes and vice versa) and re-tag those objects. We then retrain our LVQ-network using the new BALQSO versus non-BALQSO tags to create a final neuron map. Note that redshift uncertainties are explicitly taken into account by our network and all the spectra have been de-reddened to match the non-BALQSO composite. Below, we will sometimes refer to the full hybrid method as “LVQ based”, but it is always worth keeping in mind that LVQ is only one part of a process also involving the BI and visual inspection.

2.4 The final BALQSO catalogue

Our LVQ-based DR5 BALQSO catalogue contains 3552 objects ($\approx$12.5 per cent of the parent QSO sample). Fig. 2 shows a flow diagram detailing the individual steps involved in creating this catalogue, along with the numbers of QSOs associated with each step. Overall, we find that 3205 QSOs (11.3 per cent of the parent sample) are classified as BALQSOs by the BI metric (i.e. BI $> 0\,\mathrm{km\,s^{-1}}$), and 3282 QSOs (11.5 per cent) are classified as BALQSOs by the LVQ network alone (without visual inspection). The subset of objects classified as BALQSOs by both methods comprises 2130 QSOs (7.5 per cent), and only these are added to the final catalogue without undergoing visual inspection. All of the QSOs for which the BI and LVQ classifications disagree are inspected and classified visually. This step contributes a further 1422 objects (5.0 per cent) to the catalogue.

That BALQSO classification can be difficult is highlighted when one considers the percentage of false identifications produced by each of the two automated methods (i.e. the BI and LVQ) in isolation. The 3552 BALQSOs in our final catalogue include 2840 of the 3205 objects classified as BALQSOs by the BI metric calculated by Gibson et al. (2009). Thus, the BI alone would have missed 20.0 per cent (712/3552) of the objects in our final catalogue and produced false positives at a rate of 11.4 per cent (365/3205). Similarly, LVQ alone would have missed 20.0 per cent (710/3552) of our BALQSOs and produced false positives at a rate of 13.4 per cent (440/3282). Both methods individually yield very comparable false identification rates, which highlights the large uncertainties associated with previous BALQSO classifications. Clearly, designing a fully automated, reliable and reasonably complete classification scheme for BALQSOs is a difficult task.

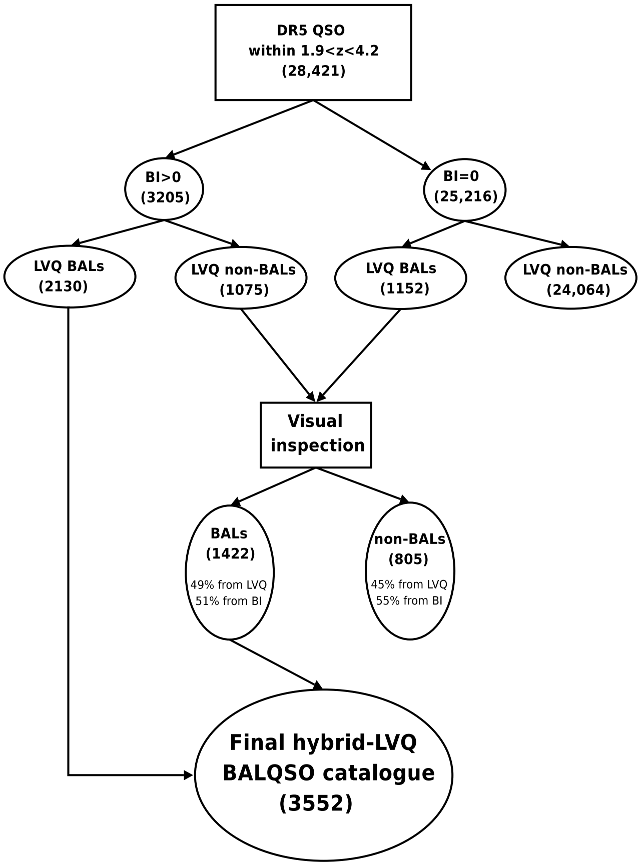

Figure 2. Flow diagram illustrating the steps involved in our hybrid-LVQ classification method.

In order to explore this issue further, we present Fig. 3, which shows four QSO spectra that highlight some of the subjectivity and difficulty associated with classifying BALQSOs. The two spectra on the left side both have positive BIs and, as a consequence, are included in the Gibson et al. (2009) BALQSO catalogue, but not in ours. These were objects which were tagged as non-BALQSOs by our LVQ network and were visually inspected for final classification, since they both had positive BI's. The spectrum in the bottom left (SDSS J010858.02+005114.6) provides a particularly useful insight. Here, the C IV line shows no sign of absorption, and thus this object was not classified as a BALQSO by us. However, there is some evidence that the Lyman α line *does* show reasonable broad, blue-shifted absorption. By contrast, the spectra on the right have BI = 0, despite the fact that they show signs of absorption (and are therefore included in our catalogue). These objects were recognized by the LVQ network and, due to the disagreement between the BI and LVQ verdicts, visually inspected for final classification.

One last note of caution concerns the rates of false positives and negatives among objects that were *not* visually inspected. While the sample of 2227 BALQSOs that *were* inspected visually may be considered to be fairly reliable, the samples of non-inspected objects are not as clean. In particular, since LVQ and BI alone produce false positives at rates of 11.4 per cent and 13.4 per cent, respectively, we may expect 1.5 per cent (0.114 × 0.134) of the 2130 BALQSOs on which they both agree to be false positives. This amounts to roughly 33 expected false positives in our BALQSO catalogue. Conversely, both methods miss approximately 20 per cent of BALQSOs, so they will erroneously agree on a non-BAL classification for 4 per cent (0.2 × 0.2) of true BALQSOs. This amounts to roughly 85 false negatives, i.e. 85 BALQSOs that are missing from our catalogue.

Because of the many problems encountered when trying to compile BALQSO catalogues, we have decided to produce for the scientific community a meta-catalogue which includes our whole DR5 parent sample used in this work instead of just a BALQSO catalogue.[2] The first 10 entries of this meta-catalogue are presented in Table 1. For each QSO in our parent sample we provide all tags that we have found to be useful in creating our own hybrid-LVQ BALQSO catalogue.

3 DISCUSSIONS

3.1 The classification of borderline cases

In this section, we highlight the difficulties in compiling BALQSO samples using composites derived from our QSO meta-catalogue as shown in Fig. 4. Each panel displays the non-BALQSO composite (shown in dashed green) normalized to 1750 Å and reddened to match the other composites shown in each panel (solid blue lines), which were created by selecting relevant QSO subsets culled from the meta-catalogue. In the top row we show composites from QSOs in our parent sample which were finally classified as BALQSOs by our hybrid-LVQ method, subdivided into objects with AI = $0\,\mathrm{km\ s^{-1}}$ (top-left panel), BI = $0\,\mathrm{km\ s^{-1}}$ (top middle panel) and LVQ non-BAL. All of these composites show clear signatures of absorption bluewards of C IV. We note that, for consistency, we have

[2] An electronic version can be found at http://www.astro.soton.ac.uk/∼simo and on the VizieR server, as well as in the electronic version of this article (see Supporting Information).

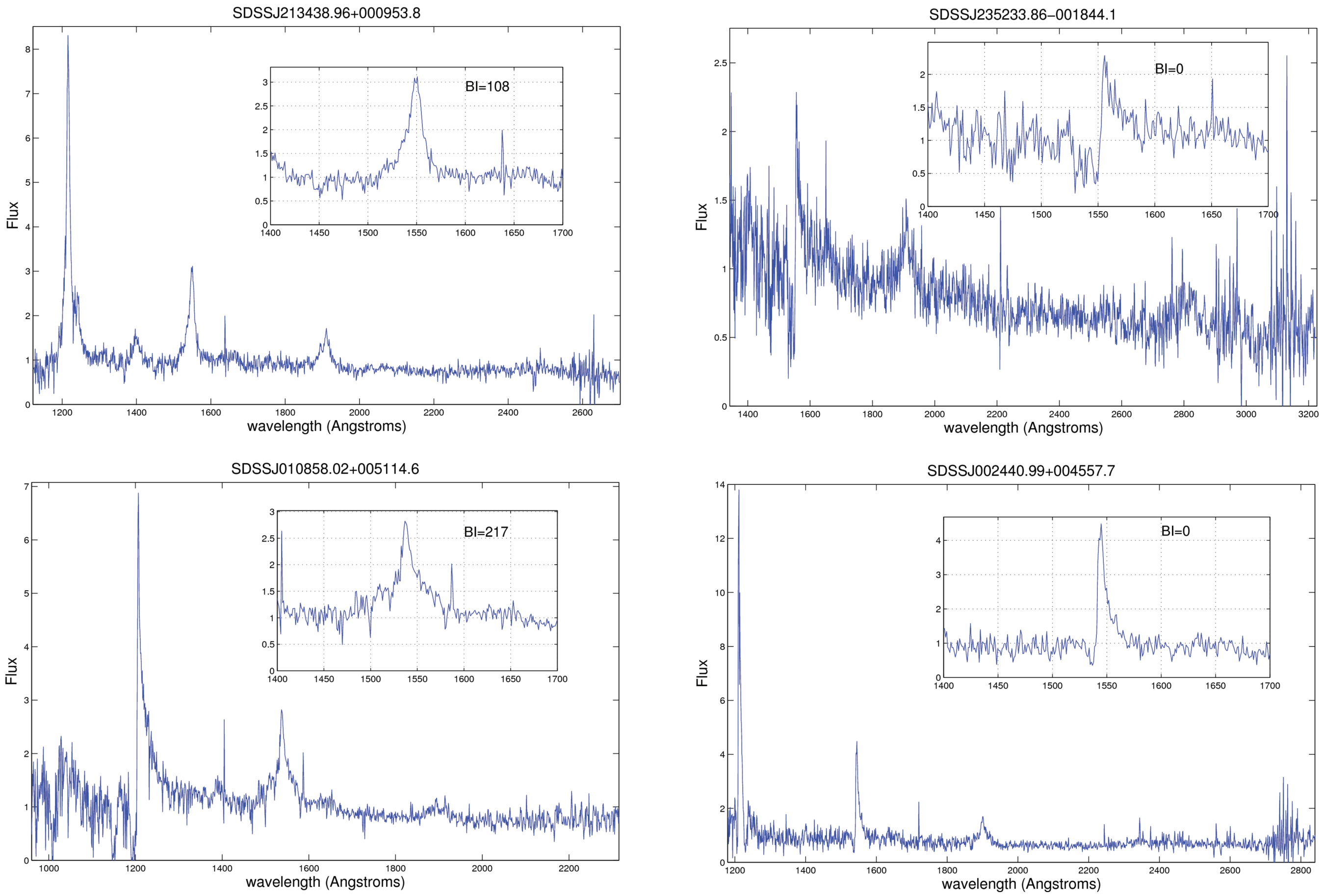

Figure 3. Four QSO spectra with different classification tags from the Gibson et al. (2009) catalogue and ours. The two spectra on the left have positive BIs but are not included in our BALQSO catalogue, whilst the two spectra on the right have BI = 0 km s^{-1} and are included in our catalogue.

Table 1. First 10 objects from our DR5 meta-catalogue. The column names are the same as those used by the SDSS team with the exception of the last two. LVQ_tag is set to 1 if the neural network regarded the QSO as a BALQSO, 0 if not. Final_tag is set to 1 if the QSO is considered as a BALQSO by our hybrid-LVQ method. The BIs have been taken from Gibson et al. (2009). The ts_t_qso and ts_t_hiz columns represent low-z quasar selection flag and high-z quasar selection flag, respectively, as defined by the SDSS team (Schneider et al. 2007). The catalogue can be found in electronic format from http://www.astro.soton.ac.uk/∼simo and the VizieR server, as well as in the electronic version of this article – see Supporting Information.

SDSS name	RA	Dec.	z	M_I	ts_t_qso	ts_t_hiz	BI	LVQ_tag	Final_tag
	(°)	(°)		(M_i)				(km s^{-1})	
000006.53+003055.2	0.027228	0.515349	1.8227	−25.100	0	0	0	0	0
000008.13+001634.6	0.033898	0.276304	1.8365	−25.738	0	0	0	0	0
000009.38+135618.4	0.039088	13.938447	2.24	−27.419	1	0	0	0	0
000009.42−102751.9	0.039269	−10.464428	1.8442	−26.459	1	0	0	0	0
000013.80−005446.8	0.057505	−0.913004	1.8361	−25.648	0	0	0	1	1
000014.82−011030.6	0.061778	−1.175193	1.8902	−26.149	0	0	0	0	0
000015.47+005246.8	0.064492	0.87968	1.8476	−26.017	0	0	0	0	0
000030.37−002732.4	0.126576	−0.459005	1.803	−25.368	0	0	0	0	0
000038.65+011426.3	0.161078	1.24064	1.8352	−25.171	0	0	2144	1	1
000038.99−001803.9	0.162498	−0.301102	2.1224	−26.673	1	0	0	0	0

only used objects already included in SDSS DR3 in constructing the composites shown in the left-hand panels, since only these have AI values calculated by Trump et al. (2006).

The composite in the upper left panel comprises objects with AI = 0 km s^{-1} (and therefore also BI = 0 km s^{-1}) that were classified as BALQSO by the LVQ network and subsequently confirmed as BALQSOs visually. Although only 126 objects were used for the creation of this composite, the absorption near C IV and the slightly truncated emission line are clear BALQSO signatures. These are mostly BALQSOs whose troughs are smaller than 1000 km s^{-1} (and hence with AI = 0 km s^{-1}).

The BALQSO composite in the top middle panel was created from objects with BI = 0 km s^{-1}, but subsequently identified as BALQSOs by the LVQ neural network. This composite shows the

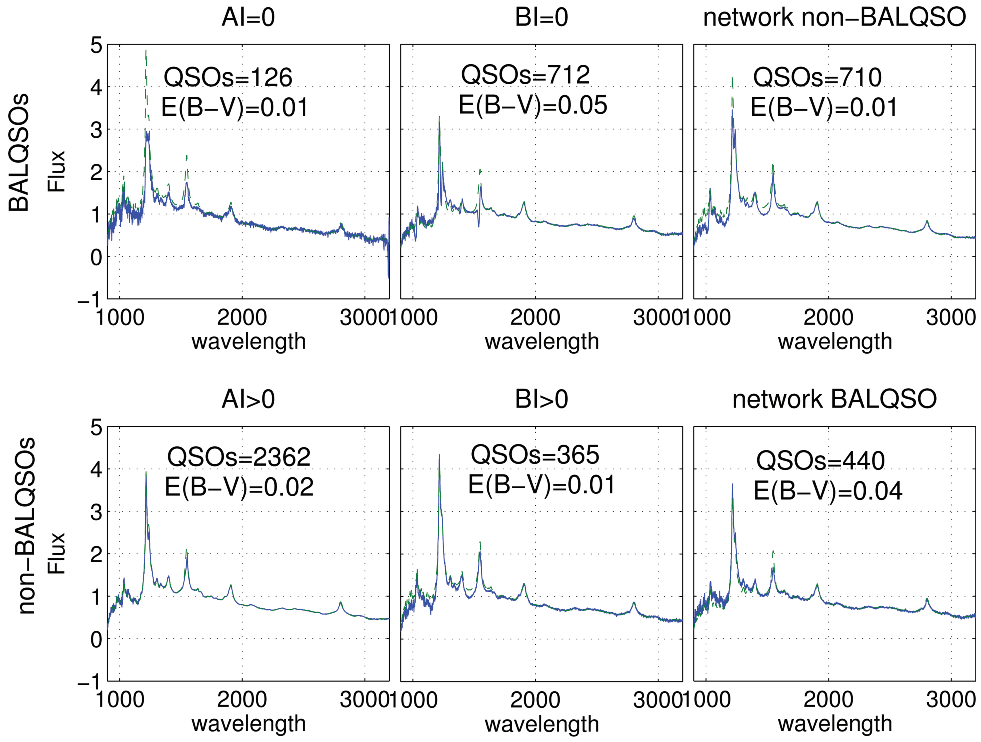

Figure 4. Various composites created in order to examine our BALQSO classification parameter space. The solid blue line displays composites created by selecting different QSOs in our classification parameter space. The dashed green line is a composite created with AI = 0 km s^{-1} QSOs after being normalized at 1750 Å and de-reddened to match the composites in each panel.

largest amount of reddening and a fairly narrow and deep absorption bluewards of C IV. Such narrow features will, by definition, be classified as non-BALs using the BI metric since they lie within 3000 km s^{-1} of the line centre. In the upper right panel, we show the composite formed from QSOs with BI > 0 km s^{-1}, identified as non-BALs by the LVQ network, but finally included in the catalogue on the basis of visual inspection. This composite shows a strong, broad, absorption line bluewards of C IV, but very little reddening. Furthermore, since the composites in both the top right and top middle panels contain similar numbers of QSOs, it is evident that neither method alone is as reliable in identifying BALQSOs as one might hope. All of the information needed to recreate these composites can be found in our meta-catalogue by querying on various tags.

The bottom row of Fig. 4 compares those composites comprising QSOs classified as BALQSO candidates using only a single metric that were then re-classified as non-BALs after visual inspection. In detail, the composite at bottom left is created from objects with AI > 0 km s^{-1}, but finally classified as non-BALs by our hybrid-LVQ method. There is virtually no evidence for absorption in this composite spectrum, despite all of the 2362 comprising this composite having AI > 0 km s^{-1}. This again points out the problematic nature of the AI metric for BALQSO selection purposes.

The middle bottom panel displays a composite created from objects having BI > 0 km s^{-1} (and therefore also AI > 0 km s^{-1}) and a non-BALQSO LVQ tag, finally classified by us as a non-BALQSO. This composite looks very similar to the one on the top right panel, which has already been discussed above. Here again, we see a fairly strong, smooth and broad (>3000 km s^{-1}) absorption feature associated with C IV. The similarity between these two composites is not entirely unexpected, since all of the objects forming them had the same automated classifications (positive BI and a non-LVQ tag), and thus differed only in the outcome of the visual inspection step. Since disagreement between BI and LVQ is most likely to happen for difficult borderline cases, we should certainly expect some misclassifications and thus overlap between the two subsets of QSOs represented by these composites. However, closer inspection does reveal some significant differences between the composites that point to the subtle but consistent absorption-line properties that were obviously picked by the visual classification step. For example, the peaks of the C IV and Lyman α lines are lower in the top right composite than in the bottom middle one, and only the top right one shows clear evidence of absorption eating into the blue wing of the C IV line (compared to the non-BAL composite). Moreover, even though both composites show some evidence for absorption affecting the bluest part of the spectrum – shortwards of Lyman α, and particularly around the Lyman β and O VI blend near 1030 Å – this absorption is stronger in the top right panel. Finally, the broad absorption trough associated with C IV in the bottom middle panel is suspiciously symmetric between 2000 and 20 000 km s^{-1}, the limits within which the BI is calculated. This may indicate that this trough is formed from the superposition of many narrow lines that may or may not be associated with the traditional BAL-flow.

All of these differences are consistent with the idea that the objects represented in the top right panel (which is included in our final BALQSO catalogue) are more likely to be genuine BALQSOs than those represented in the bottom middle panel (which are not included in our final catalogue). However, there is no escaping the fact that the differences are extremely subtle and that a definitive classification scheme for such borderline cases remains elusive. This conclusion is supported by a visual re-inspection of all of the objects contained in these two subsamples: while we generally remain happy with our classifications as 'best-bet estimates', it is clear that in many cases a definitive classification is impossible.

Since the 365 borderline cases represent 11 per cent of the Gibson et al. (2009) sample, we caution that there is a systematic uncertainty of ∼11 per cent on the BALQSO fraction suggested by even the best presently available classification schemes. This is one of the key reasons we have decided to provide the community with all of the meta-data we have used in constructing our own catalogue.

The last figure on the bottom right panel displays a composite created by selecting QSOs originally classified as BALQSOs by our neural network but re-classified as non-BALQSOs during the visual inspection phase (these objects all had BI = 0 km s^{-1} by definition or they would have not been inspected visually). The composite here is somewhat redder than the non-BAL [$E(B - V) = 0.04$], but no clear signatures of absorption are present. This highlights the importance of a visual inspection phase when constructing BALQSO catalogues.

To summarize, it is clear that no single metric (or visual intervention) is adequate in deriving both complete and clean samples of BALQSOs at the moment, so a variety of complementary metrics should instead be employed. Our own experience with unsupervised and supervised learning networks shows that, even though much of the classification work may indeed be automated, human intervention is not only useful, but also, often, a necessity when dealing with classification involving “not so clearly defined” training samples.

3.2 The effect of S/N

Fig. 5 shows f_{BALQSO} as a function of signal-to-noise ratio (S/N) for BI-selected QSOs, LVQ-selected QSOs and our final BALQSO fraction using our hybrid-LVQ method. The same trend as that found by Gibson et al. (2009) is evident for BI-selected BALQSOs: f_{BALQSO} steadily increases from ≈9 per cent in low S/N data up to 15 per cent in high S/N data. We suspect that this is because in low S/N data even relatively small random fluctuations in a shallow BAL trough can trigger the zero reset in the BI calculation and can thus result in BI = 0 km s^{-1}. We note that this would not necessarily be the case if BALs were identified using a more sophisticated metric than the BI to isolate the BAL. We cannot rule out, however, that the apparent trend in the BAL fraction with S/N has a more interesting cause, such as an underlying trend with redshift or luminosity (i.e. the BAL fraction may be higher among low-redshift and/or high-luminosity QSOs, which would also have higher S/N spectra, on average). However, the simpler and more mundane explanation – that the trend is primarily due to the difficulty in identifying BAL features in low-S/N spectra – seems far more likely. We have also visually inspected some of the objects with high BI that are not included in our final catalogue and conclude that these are cases where the BI must have been calculated incorrectly and should have been set to BI = 0 km s^{-1}.

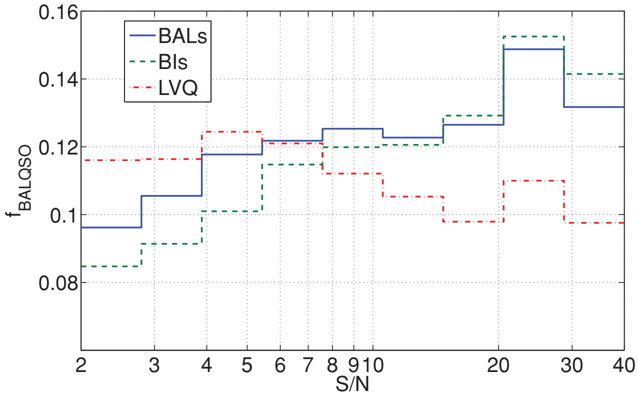

Figure 5. f_{BALQSO} as a function of S/N calculated in the same way as Gibson et al. (2009) for BI-selected QSOs, hybrid-LVQ selected QSOs and the final BALQSOs included in our final hybrid-LVQ catalogue.

By contrast, the BALQSO fraction produced by LVQ alone at high S/N levels is roughly constant and slightly lower than the fraction suggested by the BI or indicated by our final catalogue. Thus, the maximum efficiency of LVQ (when working on high-quality spectra) is comparable to, but slightly less than, that of the BI. Fig. 5 also shows that the LVQ-suggested BALQSO fraction actually *increases* towards the lowest S/N levels. Given that the number of low-S/N BALQSOs suggested by LVQ alone is actually higher than that in our final catalogue, and that every LVQ-selected BALQSO candidate was either included in the catalogue or rejected as a false positive via visual inspection, this implies that LVQ has a tendency to classify low-S/N spectra as BALQSOs, leading to a higher false positive rate in this limit. This is not entirely unexpected and actually means that LVQ and BI selections are highly complementary methods when applied across the full range of S/N levels.

4 CONCLUSIONS

We have used a recently developed technique for identifying broad absorption lines in quasar spectra to compile a more robust and complete BALQSOs catalogue. Our technique is based on a combination of the traditional ‘balnicity index’, a simple neural network and visual inspection of borderline cases and is designed to produce BALQSO samples that are more complete than purely BI-based ones, while still avoiding a high incidence of false positives. Our final catalogue covers the redshift range $1.7 < z < 4.2$ and contains 3552 BALQSOs, corresponding to a raw fraction of ≈12.5 per cent of the SDSS DR5 QSOs parent sample with a false positive rate of ∼11 per cent. In the process of constructing a robust BALQSO catalogue, we have explored in detail the classification parameter space for BALQSOs and highlighted the difficulties in BALQSO classification using single metrics. We have also constructed – and made available – a meta-catalogue that contains all of the information needed to recreate our BALQSO catalogue from its much larger QSO parent sample, or to create alternative BALQSO samples using different selection criteria. In addition, all of the composite spectra shown in this paper will be made publicly available. Meta-catalogues provide an elegant solution to problems encountered regarding subjectivity and transparency (Hogg & Lang 2008), in particular when dealing with ‘ill-defined’ astronomical objects such as BALQSOs.

ACKNOWLEDGMENTS

The authors wish to thank the anonymous referee for constructive and useful feedback. This work is supported at the University of Southampton and University of Leicester by the Science and Technology Facilities Council (STFC). Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society and the Higher Education Funding Council for England. The SDSS Web Site is http://www.sdss.org/. The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural

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

REFERENCES

Adelman-McCarthy J. K., 2007, ApJS, 172, 634

Brotherton M., de Breuck C., Schaefer J., 2006, MNRAS, 372, L90

Foltz C. B., Chaffee F. H., Hewett P. C., Weymann R. J., Morris S. L., 1990, BAAS, 22, 806

Ganguly R. et al., 2007, ApJ, 665, 990

Gibson R. R. et al., 2009, ApJ, 692, 758

Hall P. B. et al., 2002, ApJS, 141, 267

Hewett P., Foltz C., 2003, AJ, 125, 1784

Hogg D. W., Lang D., 2008, in Bailer-Jones C. A. L., ed., AIP Conf. Ser. Vol. 1082, Classification and Discovery in Large Astronomical Surveys. Am. Inst. Phys., New York, p. 331

Knigge C., Scaringi S., Goad M. R., Cottis C. E., 2008, MNRAS, 386, 1426 (Paper I)

Kohonen T., 2001, Self-organizing Maps, 3rd edn. Springer Series in Information Sciences. Springer, Berlin, p. 501

Korista K. et al., 1992, ApJ, 401, 529

North M., Knigge C., Goad M., 2006, MNRAS, 365, 1057

Reichard T. et al., 2003a, AJ, 125, 1711

Reichard T. et al., 2003b, AJ, 126, 2594

Scaringi S., Cottis C. E., Knigge C., Goad M. R., 2008, in Bailer-Jones C. A. L., ed., AIP Conf. Ser. Vol. 1082, Classification and Discovery in Large Astronomical Surveys. Am. Inst. Phys., New York, p. 191

Schneider D. P. et al., 2007, AJ, 134, 102

Stocke J., Morris S., Weymann R., Foltz C., 1992, ApJ, 396, 487

Tolea A., Krolik J. H., Tsvetanov Z., 2002, ApJ, 578, 31

Tremonti C. A., Moustakas J., Diamond-Stanic A. M., 2007, ApJ, 663, L77

Trump J. R. et al., 2006, ApJS, 165, 1

Weymann R., Morris S., Foltz C., Hewett P., 1991, ApJ, 373, 23

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Table 1. DR5 meta-catalogue.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

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