In G magnitude, the quasars misclassifications occur in bands around integer values of the magnitude. These correspond to the borderline regions between models trained at different magnitudes. Sources with similar magnitude to the training data are often classed as outliers because of minor mismatches in the noise. This problem can perhaps be addressed by requiring that a source is classified by a model trained on data fainter than the source itself by some margin.

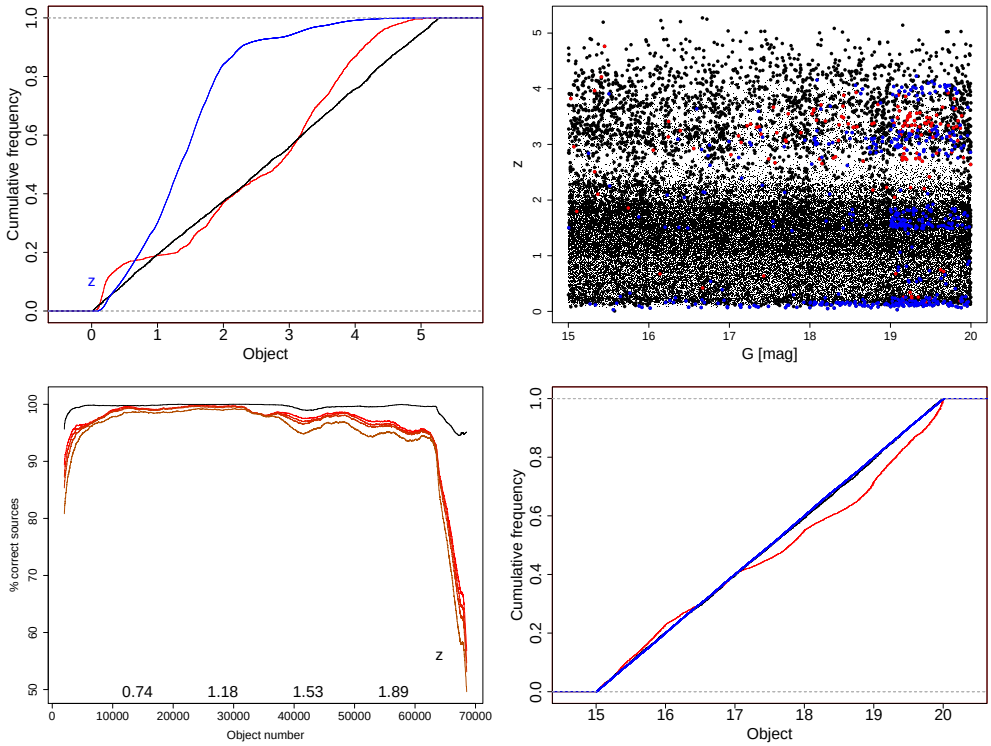

FIGURE 6: As Figure 5, but showing the Quasars in GMag versus redshift, z. Misclassified sources at upper right are coloured red for stars, blue for galaxies, black for unknown.

For the ultra-cool dwarfs, we show the effective temperature and logg. The library shows many misclassifications around the edges of the parameter distribution. This may indicate that the training set did not adequately sample the full parameter space.

In Figure 8 we show a similar plot for APec stars, concentrating on G magnitude and Teff. The results for APec stars are much sparser than for the previous libraries, but it is still clear from the

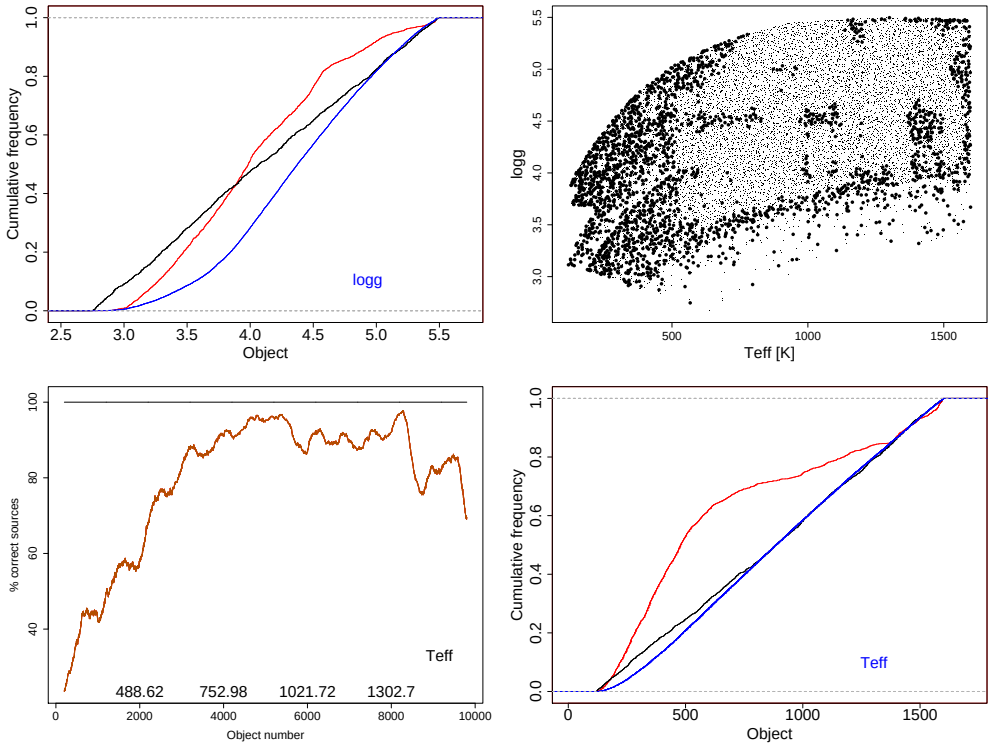

FIGURE 7: As Figure 5 and Figure 6, but showing the ultra-cool dwarfs in logg versus Teff space. The misclassified sources are almost all classed as unknown, and are clustered around the edges of the parameter distribution. This indicates that the library has been undersampled when constructing the models, and this leads to sources being rejected by the one-class SVM.

panel at lower left that the misclassifications occur preferentially amongst the high temperature objects. Most misclassifications are into the Quasars class.

Figure 9 shows an analysis of the results from the classification of physical binaries in cycle 7. Physical binaries are an particularly interesting class because, as the brightness ratio increases, they essentially blend into the single stars class with no definite boundary.

We present the classification as a function of three parameters, instead of only two as in the previous cases. These parameters are the brightness ratio (BR), effective temperature of the primary Teff1, and the G magnitude. The top two panels of Figure 9 show the distribution of misclassified sources in GMag-Teff1 and GMag-Brightness ratio space.

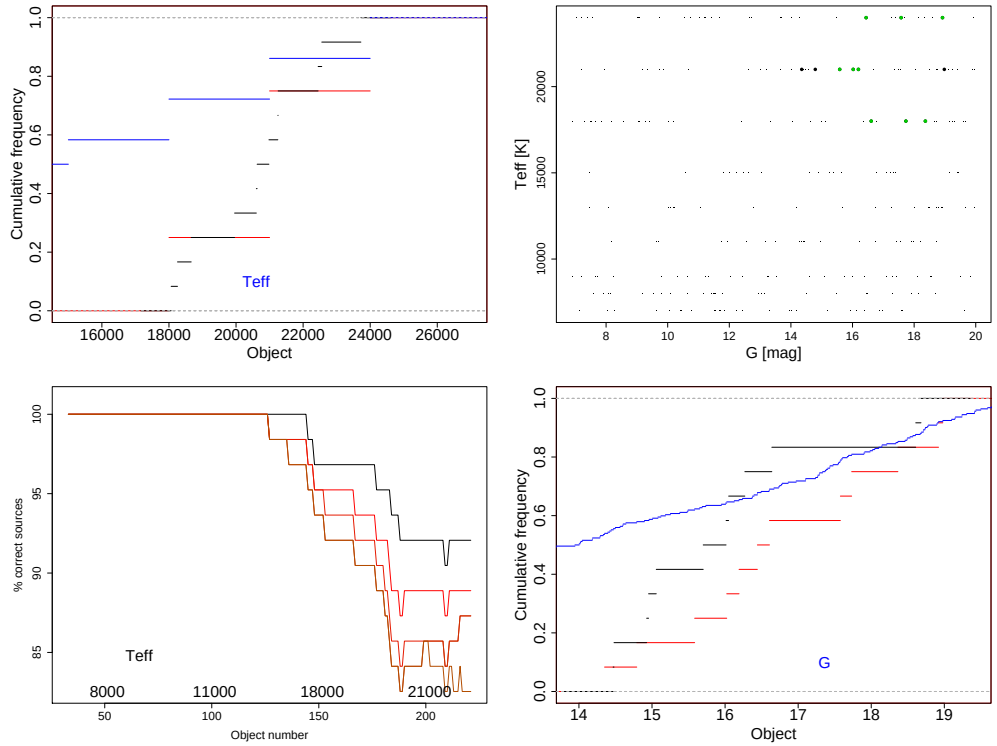

FIGURE 8: As Figures 5 to 7, but showing the Apec stars in Teff versus G magnitude space. The library is more sparse than the preious examples. The misclassified sources are split between Unknown and quasars, and are concentrated at the high temperature end of the distribution.

4.2 Class fractions

We have discussed briefly several times the fact that the class fractions encountered by Gaia will be highly unbalanced. In the version of the code discussed here, the class imbalance is built into the positon-Gmag classifier. In the future we intend to include it as a separate prior.

The test sets we have used are composed exclusively of sources from a single input grid. In this way ,we can anlyse which grids are better classified, and which types of objects within each grid are still problematic. If we want to assess the overall performance of the classifier, however, we have to take the class fractions into account, not only in the classification itself, but also in the importance of various contaminating populations.

when assessing the result, we can define the *completeness* in a particular class as

$$\text{completeness}_j = \frac{n_{i=j,j}}{N_i}, \tag{10}$$

where $n_{i,j}$ is the number of objects of true class i classified as output class j and N_i is the total number of input sources of class i. Input sources can be lost from the output class due to

misclassification into another class, or by remaining unclassified due to an insufficiently high classification confidence. The *contamination* of the output sample can be defined as the number of falsely classified sources of that class divided by the number of sources classified into that class, whether correctly or incorrectly,

$$\mathrm{contamination}_j = \frac{\sum_{i \neq j} n_{i,j}}{\sum_i n_{i,j}}. \tag{11}$$

The second of these has strong implications for assessing the performance of the classifier in the case of strongly unbalanced class fractions. If we consider the case of quasars contaminating stars, quasars are comparatively rare, so if we have equal numbers of each in the test sets, we would have to adjust the quasars down by a factor of 100 or more to get the real expected contamination. If we consider normal stars contaminating the output quasar sample, the opposit is the case. We would have to multiply the number of contaminating stars up by an appropriate factor to get the true contamination.

If we consider the case of SDSS stars in Table 4, we see that only 0.08% of the stars are misclassified as quasars. If the stars are 100 times as numerous as quasars, however, we would expect the true fraction of contaminants in the output quasar sample to be of order 8%. The factor of 100 is probably conservative. for the Phoenix random test set, the situation is worse, with 0.34% of the input sources being misclassified as quasars, which would translate to 34% of the output quasar sample with a ratio of 100:1.

We investigate the effects on the completeness and contamination of the output quasar sample caused by varying the assumed class fraction and also the probability threshold for classification. For this experiment we use only the photometric classifier. We start with a 1:1 ratio of quasars to stars and reduce the number of quasars. This has two effects. First, the class fraction prior in the classification is adjusted so that quasars are *a priori* less likely. This effect reduces the posterior probability of a source being a quasar (left-hand plot of Figure 10). If the posterior probability for a source falls below the selected threshold, the source drops out of the output quasar sample. A threshold of 0.67 is indicated by the horizontal line in the left-hand plot of Figure 10.

This tends to reduce the completeness (green curves in Figure 10, right hand side). It also increases the contamination, since the contaminating stars are 'upweighted' proportionally to their relative class fraction. However, the increasing prior probability against quasars in the classification eventually excludes the contaminating stars, causing the sharp falls in the contamination seen in Figure 10. The contamination will tend to fall in the long term if the contaminating sources tend t obe less probably quasars than the true quasars.

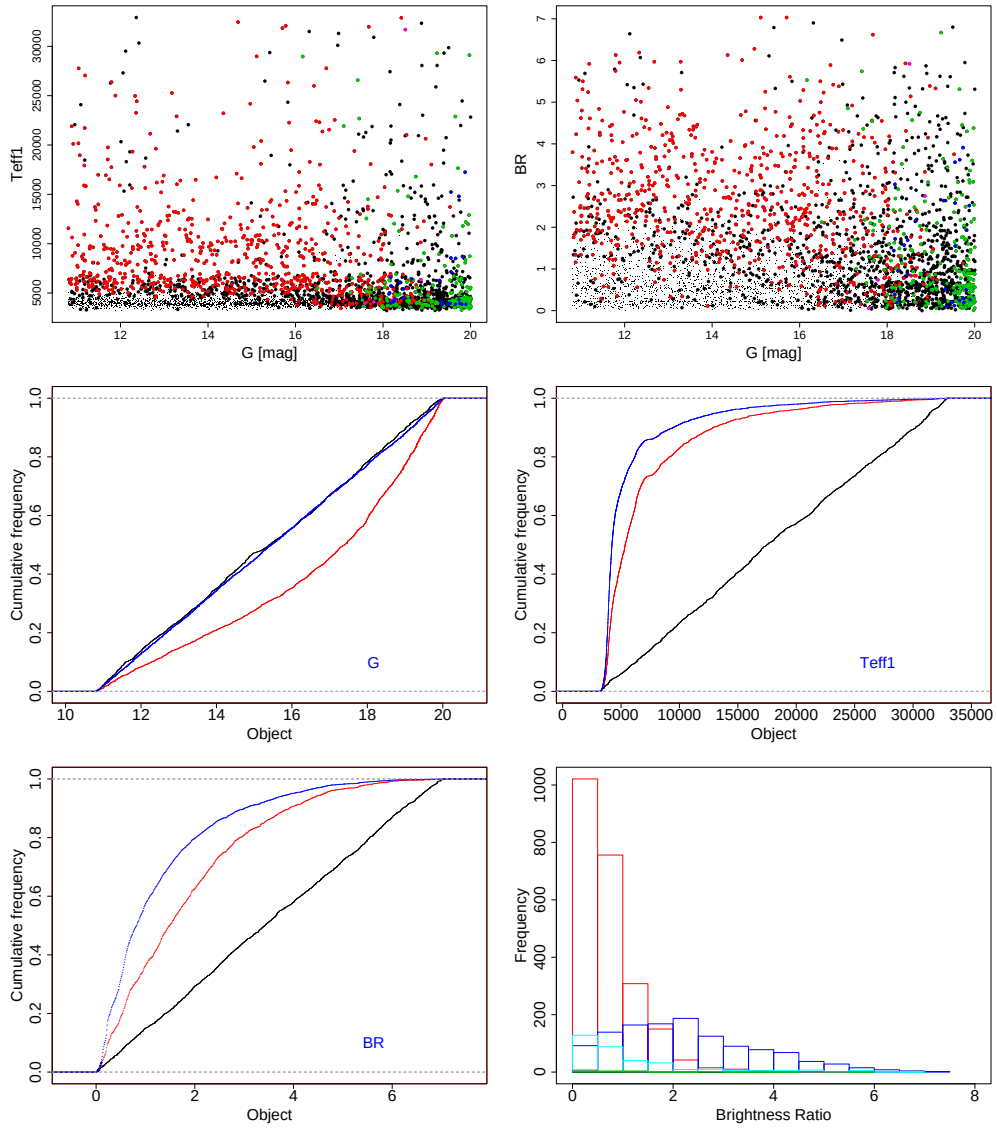

FIGURE 9: Similar to Figures 5 to 8, but with a slight change in format. These plots illustrate the performance on hpysical binaries in the cycle 7 results. We investigate three parameters, the G magntude, the effective temperature of the primary ,Teff1, and the brightness ratio, BR, which is in fact $log_{10}L_1/L_2$, the log of the bolometric luminosity ratio. On the top row we show plots of the Teff1 against GMag, and BR against GMag for the input sources. Misclassified sources are plotted with large symbols. Colour coding for misclassified sources is: black=UNKNOWN, red=STAR, scarlet=WD, green=QUASAR, blue=GALAXY. The plots in the middle row and the left-hand plot on the lower row show the cumulative distributions of all input sources (blue) and misclassified sources (red), as well as the distribution of a uniform sample with the same size as the number of misclassified sources (black). At lower right we show a histogram of the classifications distributed by brightness ratio (classification threshold: $P(\mathrm{class}) = 0.5$) The red histogram shows sources classified as binaries, blue indicates stars, cyan indicates Unknown, green indicates Quasars, and black indicates Galaxies.

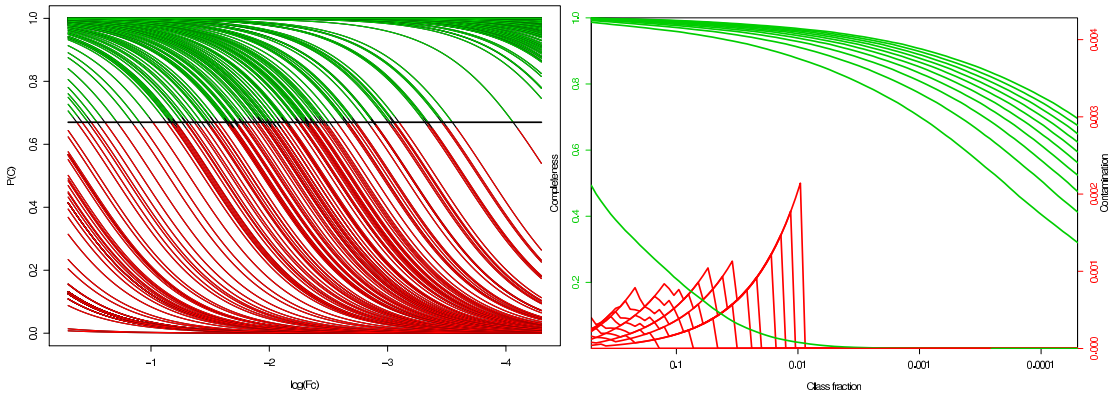

FIGURE 10: At left, the result on the output probabilities from the photometric classifier of varying the fraction of the input class (in this case Quasars). The log of the fraction of sources is shown on the x-axis. The threshold for classification is indicated with a horizontal line at P=0.67. probabilities still accepted as quasars are plotted in green and rejected regions are plotted in red. On the right is shown the resulting completeness (in green), and the contamination (red). Both these are plotted for various classification thresholds between 0.5 and 0.95.

4.3 Test on high radial velocity stars

The DSC was tested on stellar data with varying Radial velocity. The dataset is the cycle 5 VRAD grid, with 40 objects. The radial velocities for objects in this grid ranged from zero to five hundred km/s. Only the photometric classifier was used for this test.

4.3.1 Results of radial velocity test

Figure 11 shows the probabilities from the photometric subclassifier versus the four varying parameters Teff, logg, Fe/H and Rv. Table 5 shows the number of correct and incorrect classifications broken down by parameter.

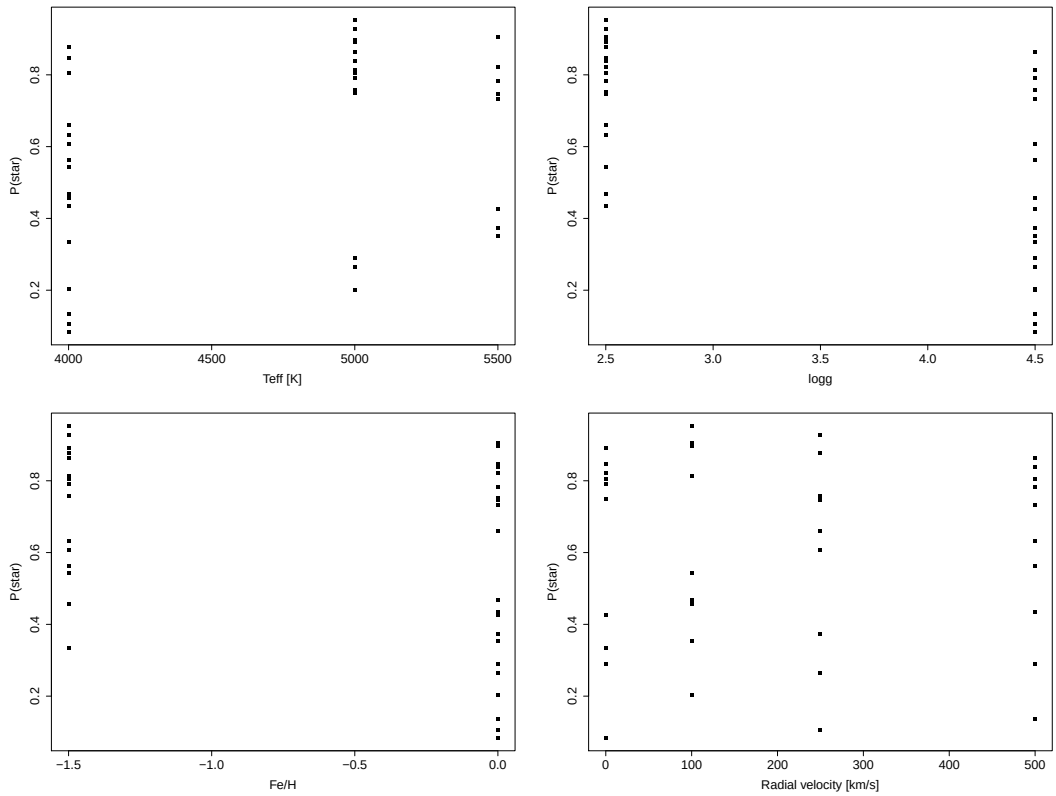

FIGURE 11: Plots of P(Star) versus various parameters for the radial velocity test. Clockwise from top left: Teff, logg, Rv and Fe/H.

The results for this are broadly consistent with the results for the main stellar libraries of cycle 5 data, which had a correct classification rate of 70% for both Marcs and Basel (Table 4). The correct classication rate might be skewed by the presence of more low temperature stars compared to the cycle 7 test sets - the low temperature stars are very badly classified.

The main conclusion is that there is no clear evidence of any effect of the varying RV on the

	$P > 0.67$	$P \leq 0.67$	% correct
Teff=4000	3	13	18.75
Teff=4500	12	4	75.0
Teff=5000	5	3	62.5
logg=2.5	15	5	75.0
logg=4.5	15	5	75.0
Fe/H=0.	10	14	41.67
Fe/H=-1.5	10	6	62.5
RV=0.	6	4	60.0
RV=100	4	6	40.0
RV=250	5	5	50.0
RV=500	5	5	50.0
Overall	20	20	50.0

TABLE 5: Correct and incorrect classifications, with a P=0.67 threshold, for the various parameter values in the radial velocity test.

classification performance.

4.4 Overlapping stellar libraries

4.4.1 Overview

We test the performance of the DSC on the overlapping regions of the cycle 5 data stellar libraries. For this test, only the photometric classifier was used.

4.4.2 The libraries

The Basel library in cycle 5 includes stars with $3000 < Teff < 15000$K. The Marcs library has $4000 < Teff < 8000$K, while the A stars library covers $8000 < Teff < 15000$. Thus the region from $Teff = 4000$ to $Teff = 8000$ is covered by both Basel and Marcs, whilst the region from $Teff = 8000$ to $Teff = 15,000$K is covered by both Basel and A libraries. There is no overlap between Marcs and A. We prepared data from all three random libraries with G=15.

Figure 12 shows the distributions of stellar parameters for the overlapping regions only. The distributions of Teff and A_v are broadly similar. There are differences between the libraries in logg and metallicity.

Figure 13 shows the median spectra for the stars in the overlapping regions. The Basel and Marcs median spectra between 4000K and 8000K are almost identical - the main visible difference is a notch at the top of the RP spectrum. The A stars median spectrum is apparently somewhat bluer than the Basel median spectrum between 8000K and 15000K. This may reflect differences in the metallicity distribution.

4.4.3 Results of the overlap test

Figure 14 shows classification results for the objects in the overlapping regions for each library. These results were obtained with DSC V7.1 with Astrometric classfier and PostionGMag classifier turned off, i.e. photometric classification only.

Table 6 shows the numbers of objects correctly classified with $P(star) > 0.5$ and $P(star) > 0.67$ for the overlapping libraries, and also the number of objects for which $P(star) < 0.5$. The most significant difference seems to be the much lower misclassification rate for the A library compared to the Basel library in the same region.

Figure 15 shows the cumulative distribution functions for P(Star) for the three libraries over their whole range (i.e. the whole range of the BaSeL library). The cumulative distribution for BaSeL simply shows cumulative P(Star) versus number of objects, with the objects sorted according to increasing Teff. A steep slope in this graph indicates generally a good classification performance, whereas a shallow slope indicates poor performance. Dashed lines are plotted for

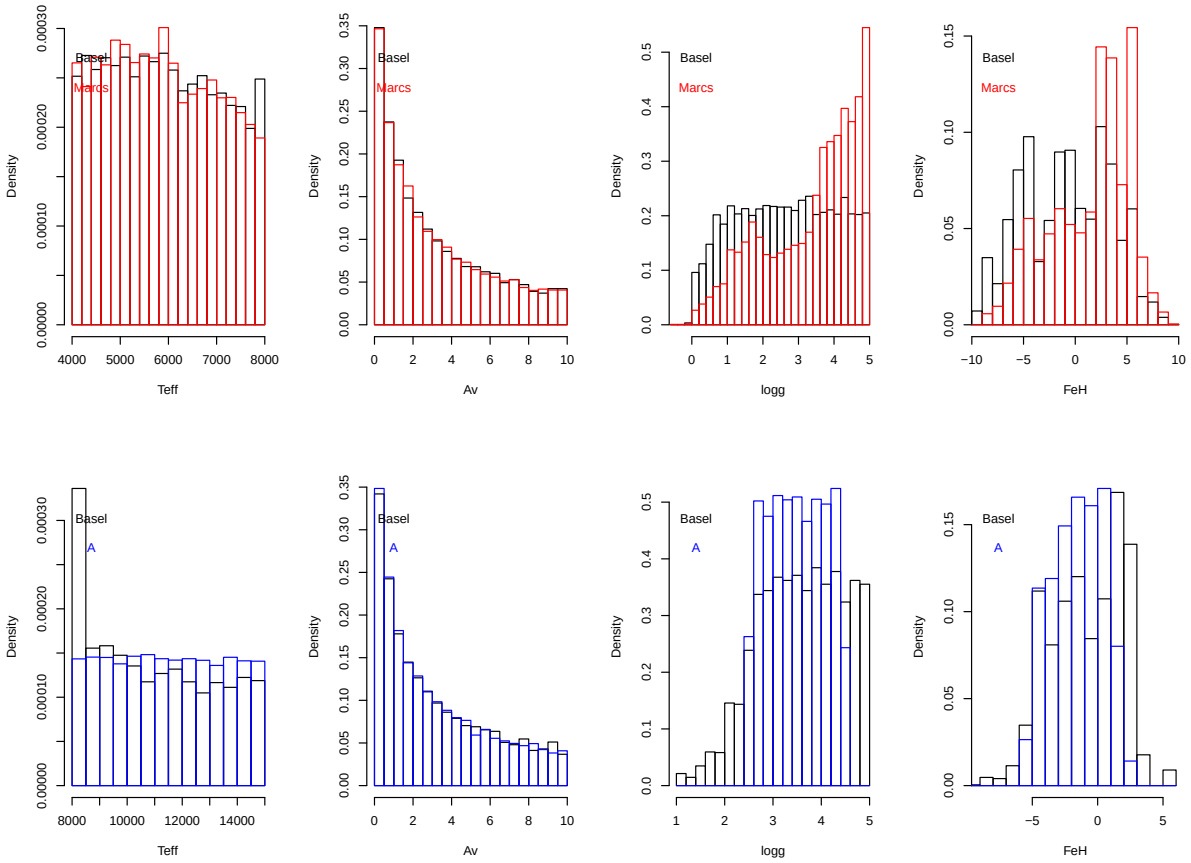

FIGURE 12: The distributions of parameters for stars in the overlapping regions. The parameters shown are Teff, Av, logg and Fe/H. On the top row are shown the distributions for both Basel and Marcs in the region between $Teff = 4000$ and $Teff = 8000$K. The Basel distribution is plotted in black and the Marcs in red. On the bottom row are the distributions of Basel and A stars between $Teff = 8000$K and $Teff = 15000$K. The Basel distribution is again in black whilst the A stars distribution is plotted in red.

	P(Star)¿0.67	P(Star)¿0.5	
Basel 4000-8000K	8,205 (93.2%)	8,500 (96.5%)	302 (3.4%)
Marcs 4000-8000K	13,469 (89.9%)	14,231 (94.9%)	759 (5.1%)
Basel 8000-15000K	4,210 (94.4%)	4,392 (98.4%)	70 (1.6%)
A 8000-15000K	9,983 (99.9%)	9,951 (99.6%)	5 (0.05%)

TABLE 6: Numbers and percentages of stars classified correctly at P=0.5 anmd P=0.67 thresholds, and numbers of objects falling short of P(star)=0.5, for the overlapping libraries.

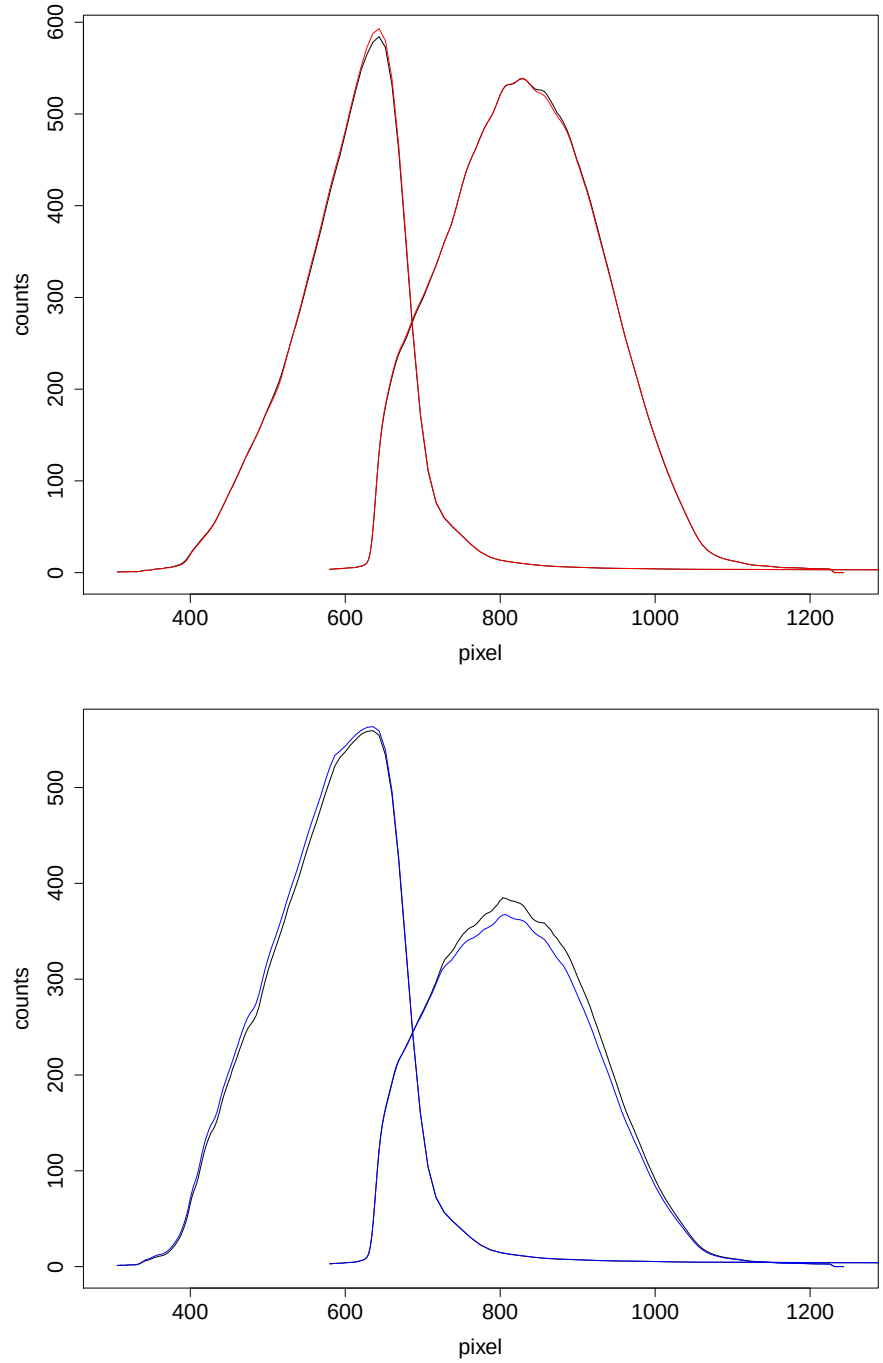

FIGURE 13: In the top panel, the median spectra for the overlapping region between Basel and Marcs (4000K to 8000K). Basel is plotted in black and Marcs in red. The spectra are almost identical. In the bottom panel, the median spectra for Basel and A in the overlapping region (8000K-15000K). Basel is shown in black and A in blue. Here, the A stars spectrum is bluer than the Basel.

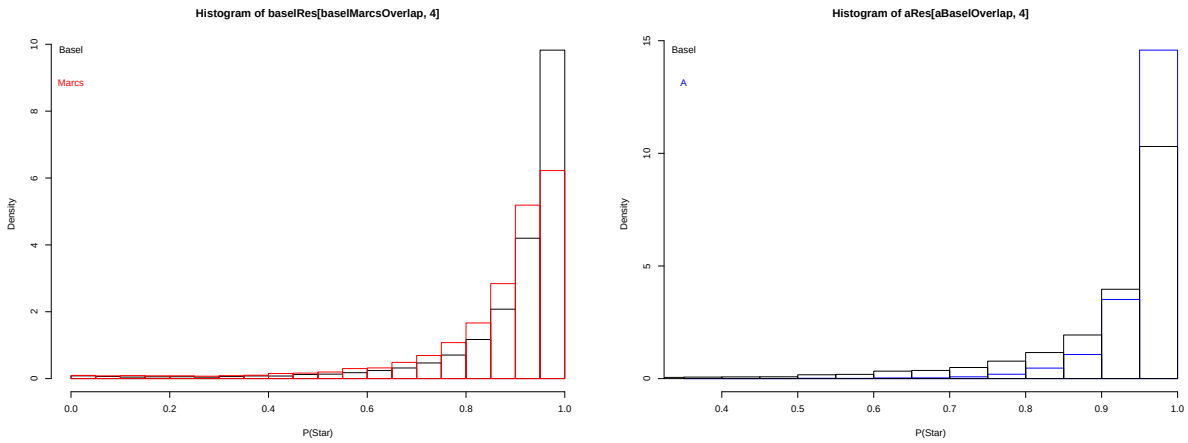

FIGURE 14: Classification results. On the left, the results of Basel and Marcs classification between 4000K and 8000K. Marcs results are shown in red and Basel in black. The y-axis shows the object density, but the histogram bins are equal so the two plots are directly comparable. The Basel histogram has a higher proportion of objects in the top probability bin and so the classification is more successful. On the right, the same comparison for Basel and A stars libraries in the region between 8000K and 15000K. The A stars are in blue. The A stars classification places a greater proportion of objects into the top bin, and so is more successful by that measure.

the cases P(star)=0.5 and P(star)=0.67. Also shown on the same axis are similar curves for Marcs (red) and A stars (blue). These are shifted so that they start at the same point as the first Basel star with the minimum temperature of the library - the Marcs curve starts at the first Basel point where Teff=4000K, the A star curve starts at the first Basel point with Teff=8000K. The curves are also scaled to the same number of objects as there are in the overlapping section of the Basel grid. This means they also end at the points of equivalent Teff on the Basel curve. Both the x and y values are scaled by the same factor, so the slopes are directly comparable.

From Figure 15 it can be seen that the Basel grid is quite poorly classified between 3000 and 4000K, where the Marcs grid starts. The average P(star) here is barely over 0.5. The performance improves after this and is reasonably consistent for the rest of the temperature domain. The Marcs slope is slightly shallower than the Basel one and the A stars slightly steeper, which supports the result seen in Figure 14.

Figure 16 is similar to Figure 15, but shows instead the false negative rate, or more accurately the number of objects for which $P(star) <= 0.5$, with the objects sorted by Teff. As before, the Marcs and A stars curves have been shifted and scaled so that they occupy the same domain as the overlapping Basel points. The change in performance for Basel is very clear. Of the 1151 total misclassifications in the Basel library, 778 occur for sources with $Teff < 4000$K and 989 for sources with $Teff < 5000$. Only 162 occur for sources with $Teff > 5000K$. The Marcs misclassifications are also clustered at low temperatures, with 596 out of a total of 759

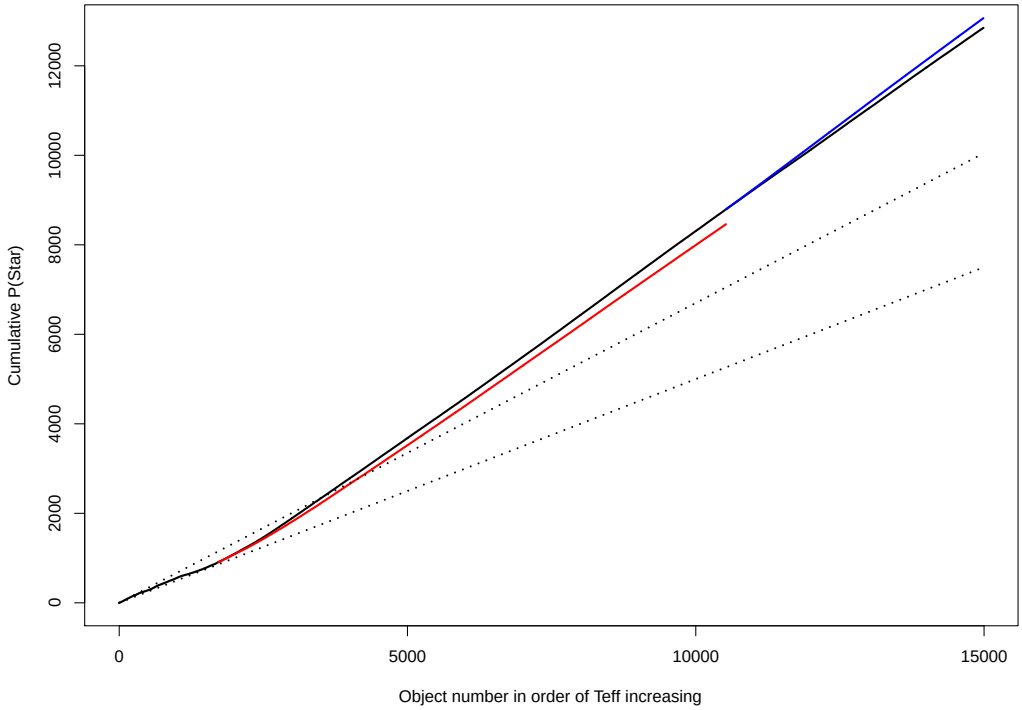

FIGURE 15: The cumulative true positive probabilities (i.e. P(star)) for Basel (black line) Marcs (red) and A stars (blue). For the Basel plot, the x-axis indicates the number of objects sorted on Teff, from lowest to highest. The y-axis is the cumulative P(star). The two dashed lines show the rate of increase of cumulative P(star) if all values were 0.5 (lower dashed line) or 0.67 (upper dashed line). The Marcs and A stars curves also show cumulative P(star) versus number of objects in increasing Teff order, but their start points have been moved to the equivalent point on the Basel curve (i.e. where $\text{Teff}_{Marcs/A}$=Teff_{Basel}) and both the x and y values have been scaled by the ratio of the number of Basel sources in the overlap region to the total number of objects in the Marcs or A star library. This means the segment corresponding to, for example, Marcs, occupies the same domain as the Basel stars of equivalent temperature, and the slopes are comparable since both dx and dy are scaled by the same factor.

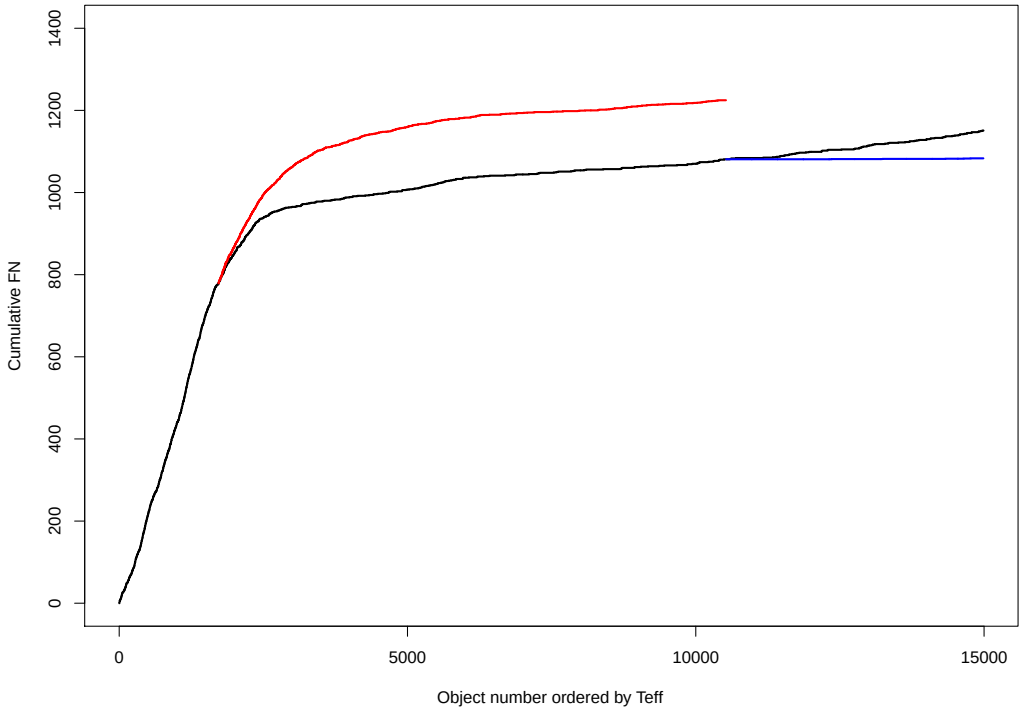

FIGURE 16: The cumulative false negative rate, with objects sorted for increasing Teff. The Basel curve is shown as-is, the Marcs (red) and A stars (blue) are shifted so that their start point coincides with the corresponding point in the Basel distribution, and the scale of these curves is multiplied by the ratio of the number of Basel points in the overlapping region to the number of Marcs or A stars points, so that these curves occupy the same domain as the overlapping Teff region of the Basel curve.

misclassifications occuring at $Teff < 5000$K.

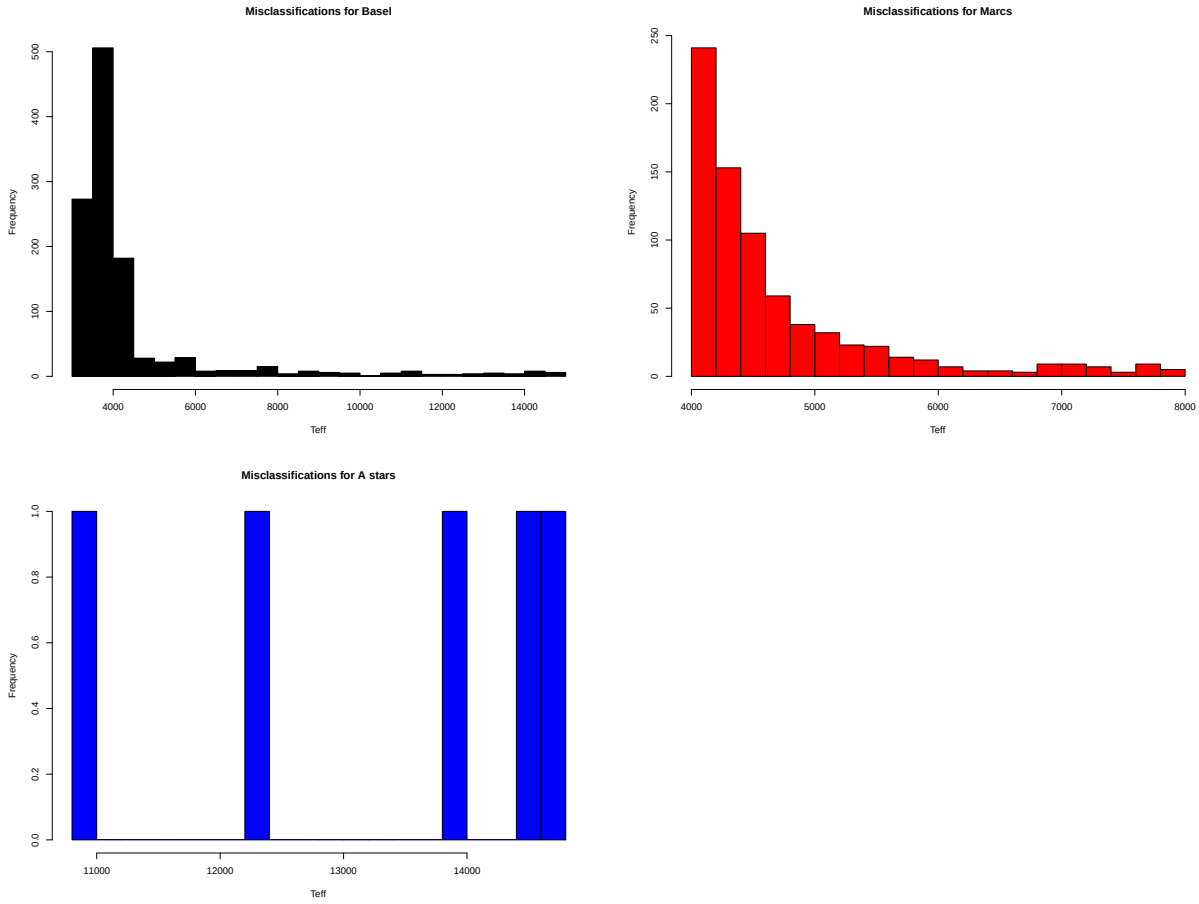

FIGURE 17: Distribution of misclassified sources (false negative, $P(star) <= 0.5$) for Basel, Marcs (red) and A stars (blue).

Figure 17 shows that the Basel and Marcs misclassifications are dominated by low-temperature sources.

5 Comparison of different subclassifiers

We examine the correlations in the performances of the different subclassifiers.

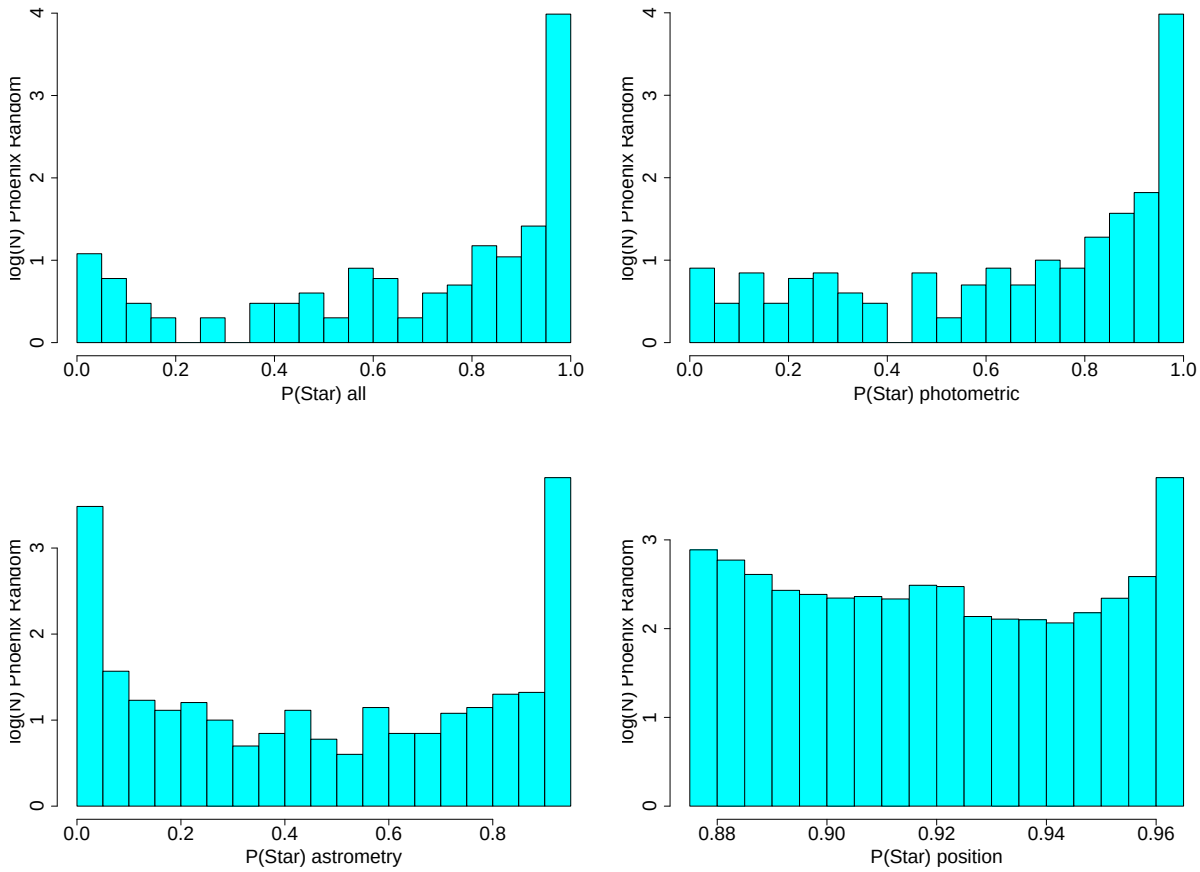

FIGURE 18: Distribution of output probabilities for the Phoenix random grid from various subclassifiers. Clockwise from top left: combined probability, photometric, position-Gmag and astrometric. The frequencies are plotted on a log scale because most of the bins have few counts. Where the frequency is zero, the value of log(N) has been set to zero. Note that there are no negative values of log(N) as there are only integer numbers of objects in each bin.

Figures 18, 19 and 20 show the histograms of the various probability outputs for the Phoenix random, SDSS stars and SDSS quasars tests respectively. The histograms are plotted on a log scale because of the large contrast between counts close to zero or one and counts in the middle of the range.

From these plots, one can see that the photometric classifier provides the strongest positive evidence for the correct classification. The Astrometric classifier for the phoenix and SDSS stars has many sources with $P \sim 0$, which are then misclassifications. This is not true of the quasars, which have a prominent spike at $P \sim 0.45$ (the probability from the astrometric classifier is split equally between the quasars and galaxies).

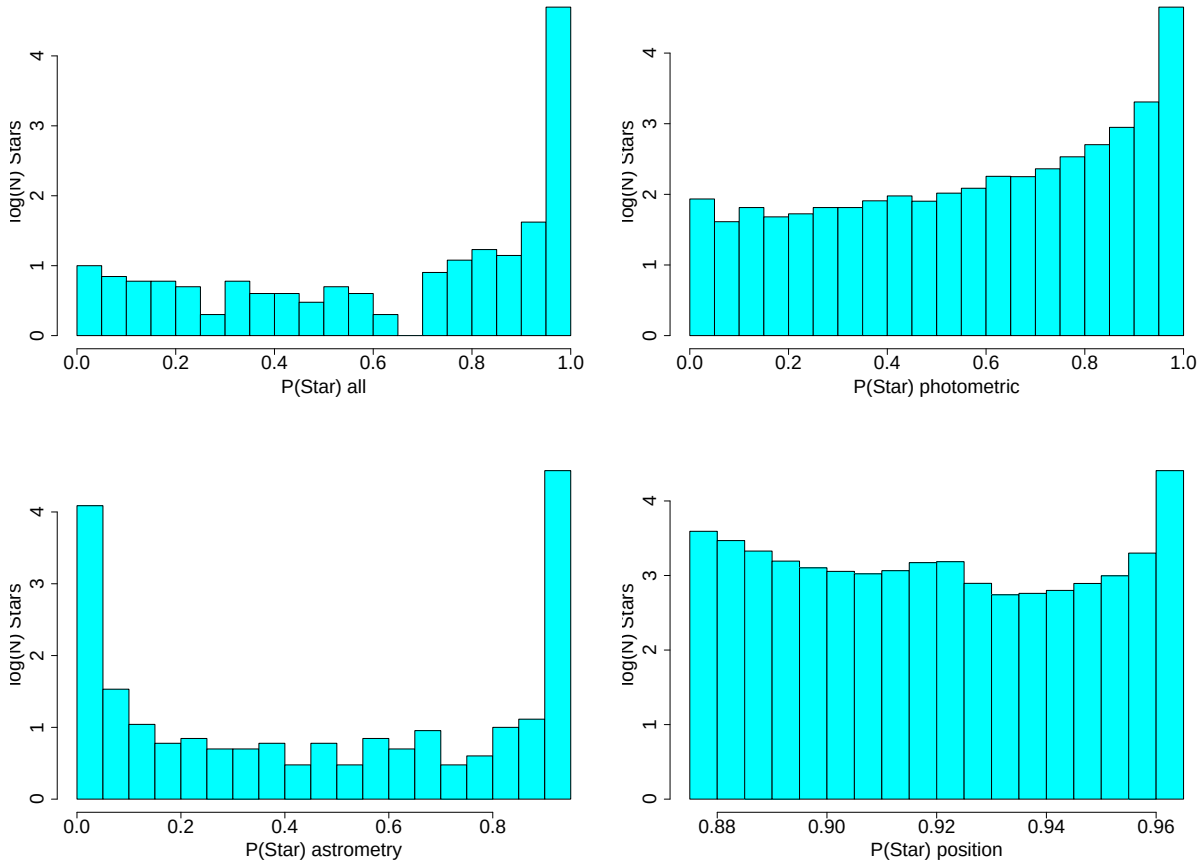

FIGURE 19: As Figure 18, but for the SDSS stars.

In Tables 7, 8 and 9 we show a breakdown of the subclassifier results. These tables are each subdivided into sixteen cells, to show the correct or incorrect classification according to all three subclassifiers, plus the overall result ($2^4 = 16$).

The eight cells in the left hand half of each table show the number of sources correctly classified overall, whilst the eight cells in the right half show the number incorrectly classified.

The top two rows of the table show the sources correctly classified by the photometric classifier, the lower two rows the misclassified sources. We use a threshold of $P(\mathrm{correct}) = 0.5$ as the decision boundary. Note that in the DSC full results, the decision boundary for correct classification is at $P(\mathrm{correct}) = 0.67$.

The left hand side of each half of the table (i.e. columns 1 and 3) shows the number of sources correctly classified by the position-G magnitude subclassifier. For the purposes of this table, we have removed the effect of the class fraction prior from the position-Gmag probabilities by dividing through by an estimated prior, replacing with an equal prior, and then renormalizing.

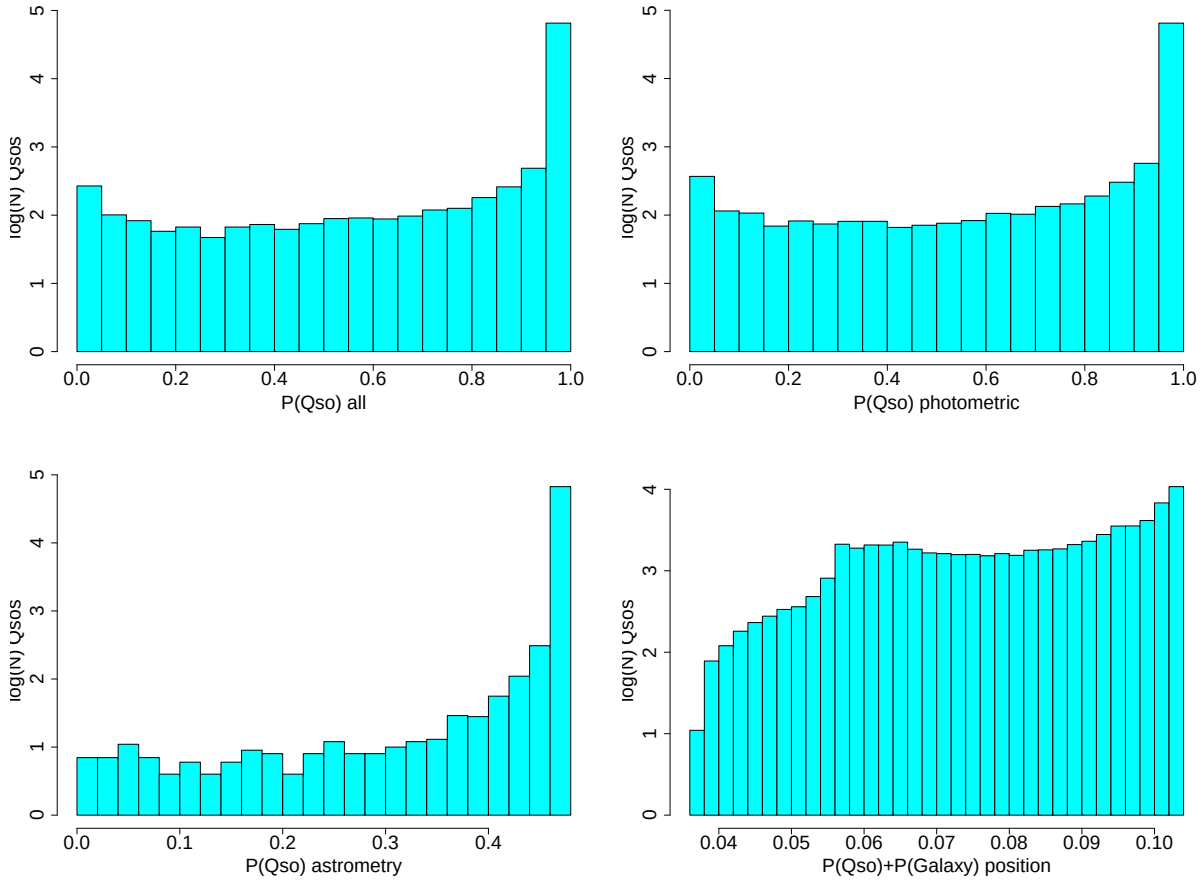

FIGURE 20: As Figure 18, but for the quasars.

The prior divided out was chosen so that roughly 50% of the stars were misclassified by the position G mag classifier. A factor of 25 increase in the quasar prior was found to achieve this. The threshold for correct quasar classification is set to 0.25%, because the probability for an extragalactic object is split equally between the quasar and galaxy classes. Even with the class fraction prior removed, very few quasars were misclassified by the position-Gmag classifier. One reason for this is that the test quasars all have $G > 15$, whilst the stars are evenly distributed over a wide range of magnitudes ($6 < G < 20$).

Finally, rows 1 and 3 show the number of sources classified correctly by the astrometric classifier, and rows 2 and 4 show the number of sources misclassified by the astrometric subclassifier. For the quasars, the threshold is again $P(\mathrm{Qso}) = 0.25$, because the probabilities are split between quasars and galaxies.

From Tables 7, 8 and 9 we can note particularly the meaning of the following elements;

		$P_{all} > 0.5$		$P_{all} < 0.5$	
		$P_{pos} > 0.5$	$P_{pos} < 0.5$	$P_{pos} > 0.5$	$P_{pos} < 0.5$
$P_{phot} > 0.5$	$P_{AC} > 0.5$	3404	3246	0	0
	$P_{AC} < 0.5$	2408	724	0	17
$P_{phot} < 0.5$	$P_{AC} > 0.5$	3	25	0	0
	$P_{AC} < 0.5$	1	0	0	19

TABLE 7: Breakdown of results by subclassifier for Phoenix random test set. A probability threshold of 0.5 is used in this classification, in contrast to the main results, because it is easier to understand. 2 516 sources with $P(\mathrm{UNKNOWN}) = 1$, $P(\mathrm{UNDEFINED}) = 1$ or $P(\mathrm{UNCLASSIFIED}) = 1$ for any subclassifier were omitted, leaving 7 484 test sources in the sample.

		$P_{all} > 0.5$		$P_{all} < 0.5$	
		$P_{pos} > 0.25$	$P_{pos} < 0.25$	$P_{pos} > 0.25$	$P_{pos} < 0.25$
$P_{phot} > 0.5$	$P_{AC} > 0.25$	66 480	4	1	0
	$P_{AC} < 0.25$	83	0	1	0
$P_{phot} < 0.5$	$P_{AC} > 0.25$	216	0	898	0
	$P_{AC} < 0.25$	0	0	1	0

TABLE 8: Breakdown of results by subclassifier for quasars. A probability threshold of 0.5 is used for the overall classification and the photometric classification, but a probability threshold of 0.25 is adopted for the position-Gmag and astrometric classifiers, because in these cases Quasars and Galaxies are indistinguishable and so the probability tends to be split between them (in the case of the astrometric classifier, it is formally impossible for $P(\mathrm{Qso})$ to rise above 0.5 because of the split with galaxies). 2 872 sources with $P(\mathrm{UNKNOWN}) = 1$, $P(\mathrm{UNDEFINED}) = 1$ or $P(\mathrm{UNCLASSIFIED}) = 1$ for any classifier were omitted, leaving 67 684 test sources in the sample.

		$P_{all} > 0.5$		$P_{all} < 0.5$	
		$P_{pos} > 0.5$	$P_{pos} < 0.5$	$P_{pos} > 0.5$	$P_{pos} < 0.5$
$P_{phot} > 0.5$	$P_{AC} > 0.5$	17 282	19 797	0	0
	$P_{AC} < 0.5$	12 038	176	0	17
$P_{phot} < 0.5$	$P_{AC} > 0.5$	92	499	2	7
	$P_{AC} < 0.5$	50	2	7	20

TABLE 9: Breakdown of results by subclassifier for SDSS stars. A probability threshold of 0.5 is used in this classification. 12 104 sources with $P(\mathrm{UNKNOWN}) = 1$, $P(\mathrm{UNDEFINED}) = 1$ or $P(\mathrm{UNCLASSIFIED}) = 1$ for any classifier were omitted, leaving 37 896 test sources in the sample.

Row 1 Col 1: These are the sources classified correctly by all subclassifiers. For all the classes, many of the objects end up in this bin.

Rows 3 and 4, Cols 1 and 2 These sources are misclassified by the photometric classifier, but are correctly classified overall based on the results of either the astrometric subclassifier or the position Gmag subclassifier or both.

Row 4 Col 2 These sources are incorrectly classified by all the subclassifiers, yet end up overall in the correct category. This applies to only one source. This result may seem counterintuitive, but in fact if all the subclassifiers return a moderate probability less than a half for a particular class, but can't agree amongst themselves on an alternative class, one can see that this can occur (see discussion in section 2.1.2.

Rows 1 and 2, columns 3 and 4 These sources are correctly classified by the photometric classifier, yet end up misclassified because of the results of one or both of the other two subclassifiers.

5.0.4 Results breakdown and discussion

For the Phoenix stars, a large majority of objects are classified correctly. A total of 29 objects are misclassified by the photometric subclassifier, but 'saved' by the astrometric and position Gmag subclassifiers, and a total of 17 objects correctly classified by the photometric subclassifier are ultimately misclassified due to the two other subclassifiers.

For the quasars, the largest category of objects are classified correctly overall and by all the subclassifiers. 216 objects are misclassified by the photometric classifier but correctly classified by the position-Gmag nad astrometric classifiers and end up correctly classified. A total of two objects are correctly classified by the photometric classifier, but end up misclassified due to a combination of the position Gmag and astrometric classifiers, and the class fraction prior. 899 objects are misclassified, despite correctly classified by the position-Gmag classifier.

The SDSS stars results resemble those of the Phoenix stars. The majority of the objects are spread between the three bins at the top left. A total of 443 sources are misclassified by the photometric classifier, but end up correctly classified due to the other subclassifiers and the class fraction prior. A total of 17 sources are correctly classified by the photometric classifier, but are eventually misclassified due to the other subclassifiers and the prior.

6 Robustness against damaged data

A subclassifier is not run if the input data are missing, or if NaN's or saturated values are present. The performance of the photometric subclassifier was investigated in the case of various types of other damage or imperfection to the BP or RP spectra, or errors in the overall flux or wavelength

calibration.

6.1 Data sets

The damaged data models are very simple, since the exact processing method is not yet established and the likely data problems are not yet known. We investigate four types of compromised data. They are:

- Hot pixels, caused by cosmic ray hits or possibly other events.
- Cool pixels. Cause unknown.
- Bad flux calibration, causing the G magnitude to vary from its true value.
- Bad wavelength calibration, causing a global shift to the spectrum.

We prepared simple versions of these types of data from the cycle 5 simulations, which include various types of stars, quasars, galaxies, binaries and white dwarfs. We selected objects which were (reasonably) well classified in their unaltered form, and applied a progressive degradation to the data, to find out at what level the classification begins to be compromised. We carried out tests for normal stars (MARCS library), galaxies and QSOs.

For each class of objects, one hundred reasonably well classified examples were first selected. By 'reasonably well classified', we mean that the true positive probability was greater than 0.5 for the undamaged spectrum.

The simulated spectra are provided with 180 resolution elements, corresponding to a factor of three oversampling with respect to the BPRP pixels. For this test, we resampled the spectra to the approximate pixel sampling of the BPRP chips (60 elements in each of BP and RP), before clipping the low signal elements at the edges. The models were trained on the remaining 86 resolution elements from both BP and RP.

For the 'hot' and 'cool' pixel datasets, the data degradation was carried out on each pixel in turn. Fifty different 'degrees' of damage were applied for each pixel.

For the wavelength calibration and flux calibration tests, the whole spectrum was affected (there is no pixel-by-pixel test). For the flux calibration, the value of the G magnitude in the calPhotSource was altered progressively. For the wavelength calibration, the entire spectrum was resampled with pixel bins shifted by up to 1 pixel redward and blueward of the true spectrum centre. A renormalization was carried out to ensure that the flux was conserved.

6.1.1 Hot pixels

A three dimensional grid of data was built for this experiment, of one hundred objects by 86 pixels by 50 different hot pixel 'strengths'. Each of the 86 pixels, in BP and RP, used by the DSC cycle 7 models was degraded in turn. The hot pixel is generated by multiplying the original flux by a factor

$$f = 1. + i/10. \quad i = 1, ...50 \tag{12}$$

(13)

so there are fifty hot pixels of different 'strengths' for each affected pixel.

The damaged data grid is run through DSC with the normal models available (see cycle 7 documentation for DSC). The output probabilities (from the BPRP subclassifier) are represented in Figure 21 for the stellar spectra, Figure 22 for the galaxies and Figure 23 for quasars. These are the average probabilities over 100 objects.

These figures indicate that the classifier performance is quite strongly affected by hot pixels across the whole range of pixels. for the most sensitive pixels, misclassification can arise for hot pixels of a factor of order 1.5 in flux. In one or two cases, the threshold is even lower.

In these Figures, misclassifications have been colour coded according to the class assigned. The assigned class is the largest probability in the BPRP probability vector. It should be borne in mind that the class assignments are based on the averaged results, so may not reflect the true statistical distribution of misclassifications in the data.

For the MARCS (stars) data, the presence of blue and red points in the misclassification area indicates a tendency for the stellar spectra to be misclassified as binaries or quasars. As the hot pixels become extreme, there is a greater and greater tendency for the sources to be classed as UNKNOWN (black points in the Figure). This is encouraging as it indicates that the outlier detector is excluding strange objects from the classification proper.

For the galaxies plot, most misclassifications are into the quasars class (red). Again, badly damaged spectra are classified as UNKNOWN.

The quasars plot is interesting as it indicates that no misclassifications occur into other astrophysical classes, but rather that all damaged spectra are classified as UNKNOWN. We stress again here that the output classes have been assigned based on averaged probabilities over all one hundred sample objects, not on the basis of individual misclassifications.

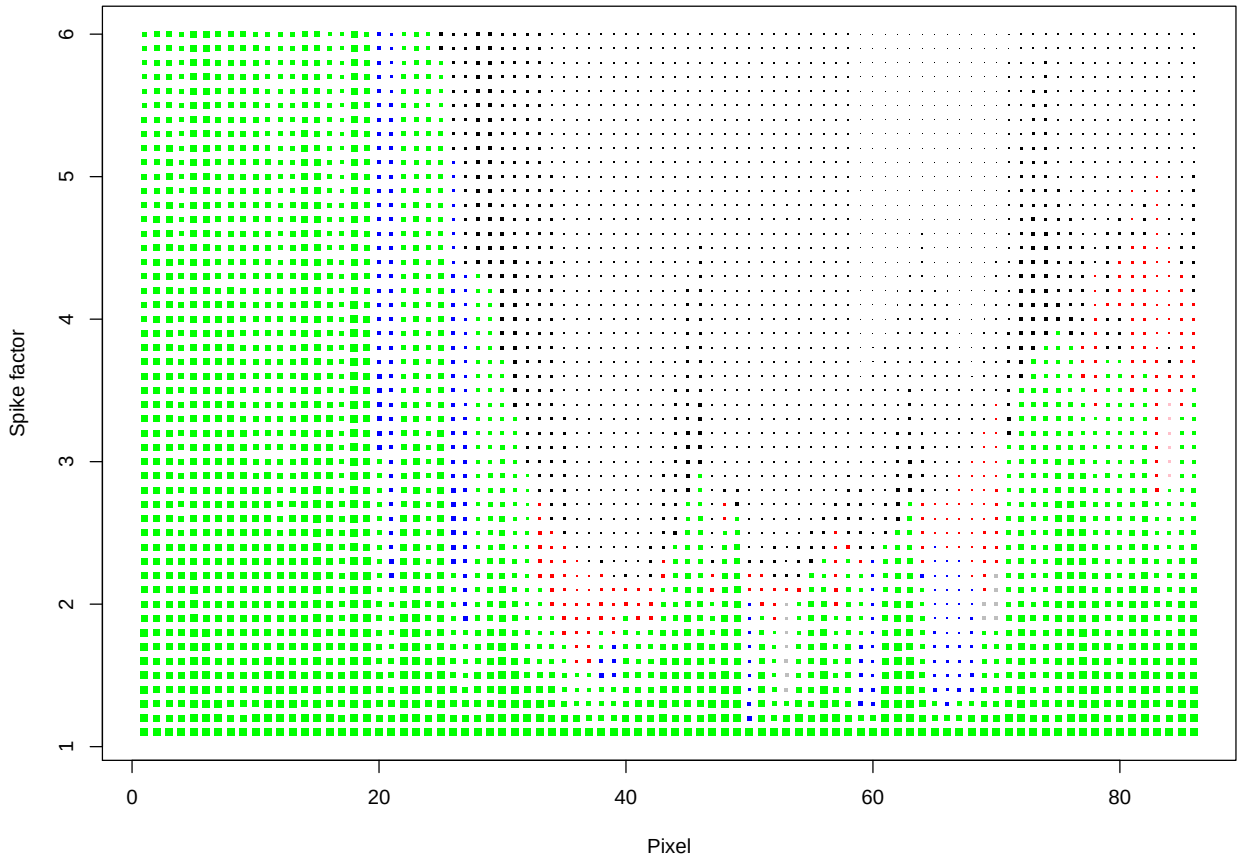

FIGURE 21: The performance of the classifier on stellar spectra damaged by the addition of spurious extra flux to one pixel (a hot pixel). The classifications are applied based on the average probability over all one hundred sources. The x-axis shows the pixel number affected. The combined BP and RP spectra cover 86 pixels, after accounting for resampling and edge clipping. The y-axis is the factor applied to the original flux in the affected pixel. The size of the plotting symbols represents the probability returned that the object is a star (which it is). Larger symbols represent larger values of $P(star)$. Additionally, symbols are colour coded according to the most probable source type. This would correspond to the classification in the event that there is no probability threshold applied. Green symbols are stars, red symbols quasars, blue symbols binaries, grey symbols white dwarfs, pink symbols are galaxies and black symbols are unknown.

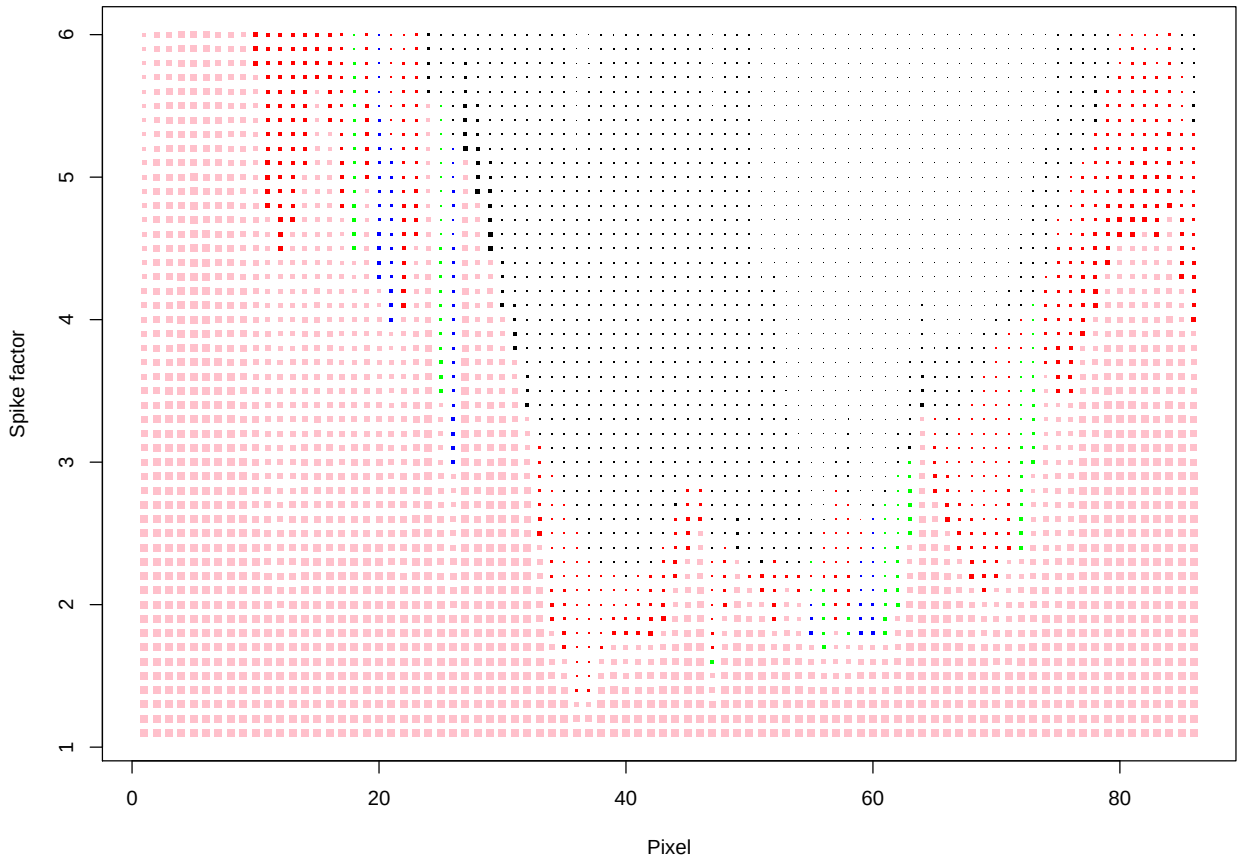

FIGURE 22: The performance of the classifier on one hundred galaxy spectra damaged by the addition of spurious extra flux to one pixel. Axes, symbols and colours are similar to Figure 21, except that now the size of the symbols indicates the returned probability that the source is a Galaxy, rather than a star.

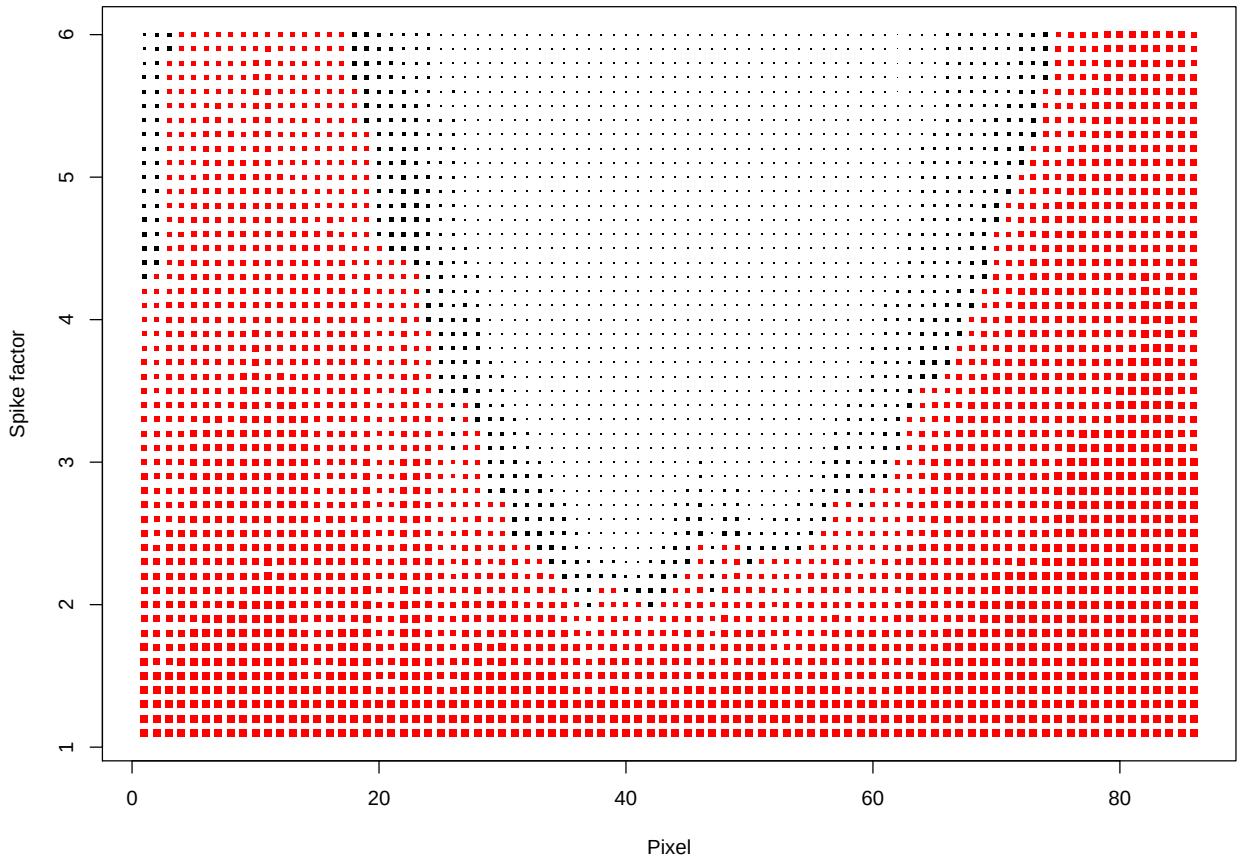

FIGURE 23: The performance of the classifier on one hundred quasar spectra damaged by the addition of spurious extra flux to one pixel. Axes, symbols and colours are similar to Figure 21, except that now the size of the symbols indicates the returned probability that the source is a quasar, rather than a star.

6.1.2 Cool pixels

This is similar to the hot pixels data. All 86 BP and RP pixels are tested in turn with fifty different levels of flux loss. Pixel fluxes are multiplied by the factor

$$f = 1./(1. + 0.1 * i); i = 1, ...50, \tag{14}$$

to introduce the 'cool' pixel. The results for one hundred stars, one hundred galaxies and one hundred quasars are plotted in Figures 24 to 26

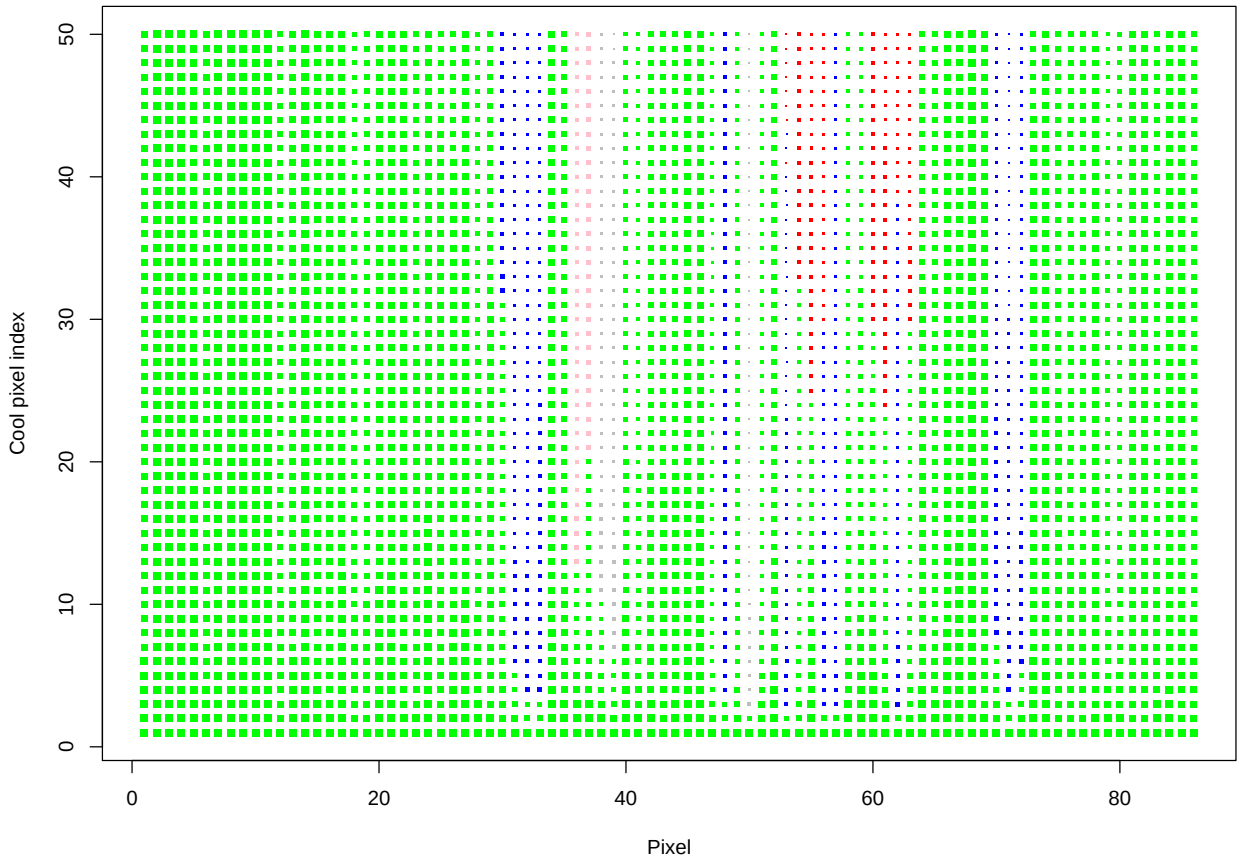

FIGURE 24: The performance of the classifier on one hundred star spectra damaged by the presence of a cool pixel. Axes, symbols and colours as for Figure 21, except now the y-axis is the index used to generate the 'cooling factor' in equation 14. Larger y-axis values therefore represent worse damage to the original spectra.

These figures show again that, for sensitive elements, even modest alterations of the flux value can cause misclassification of the source. Values of the 'cooling factor', by which is meant

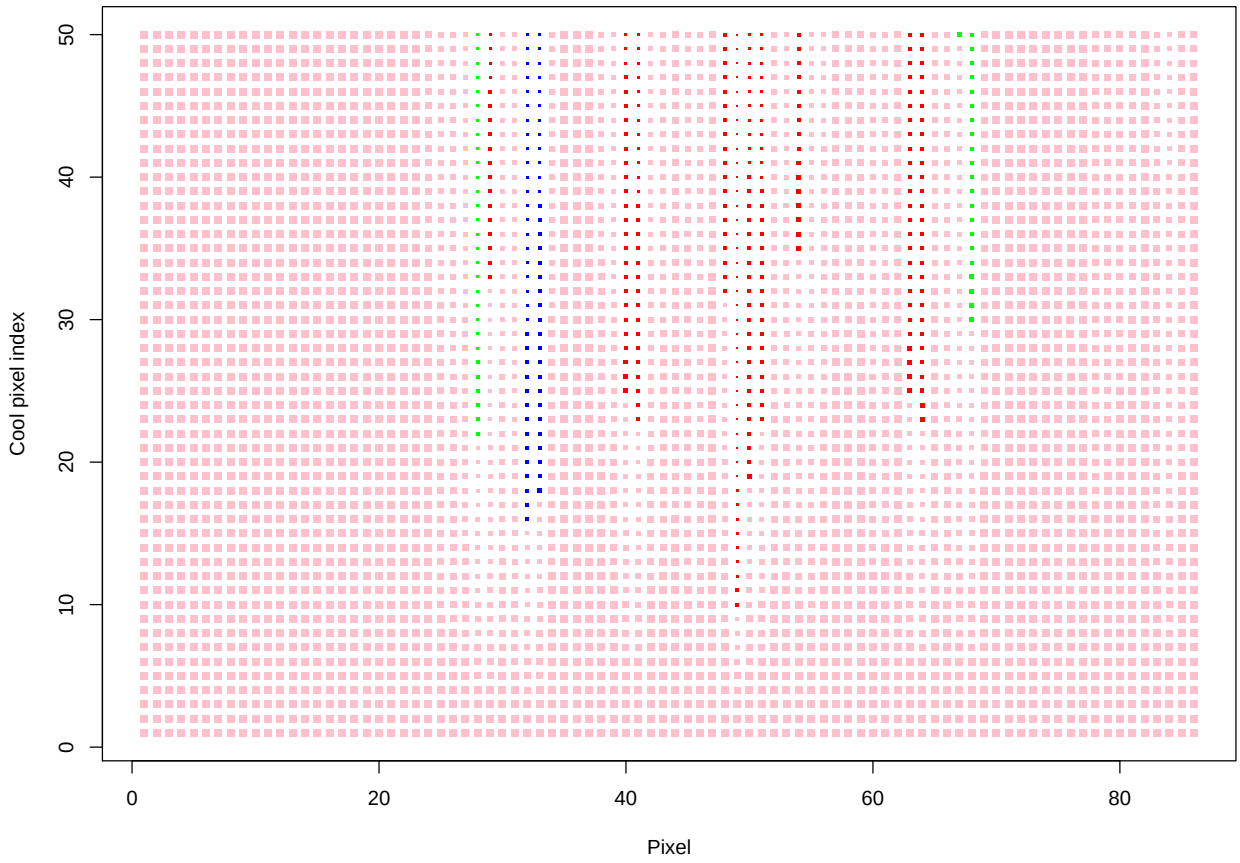

FIGURE 25: The performance of the classifier on one hundred galaxy spectra damaged by a cool pixel. Axes, symbols and colours are similar to Figure 24, except that now the size of the symbols indicates the averaged probability that the sources are Galaxies, rather than stars.

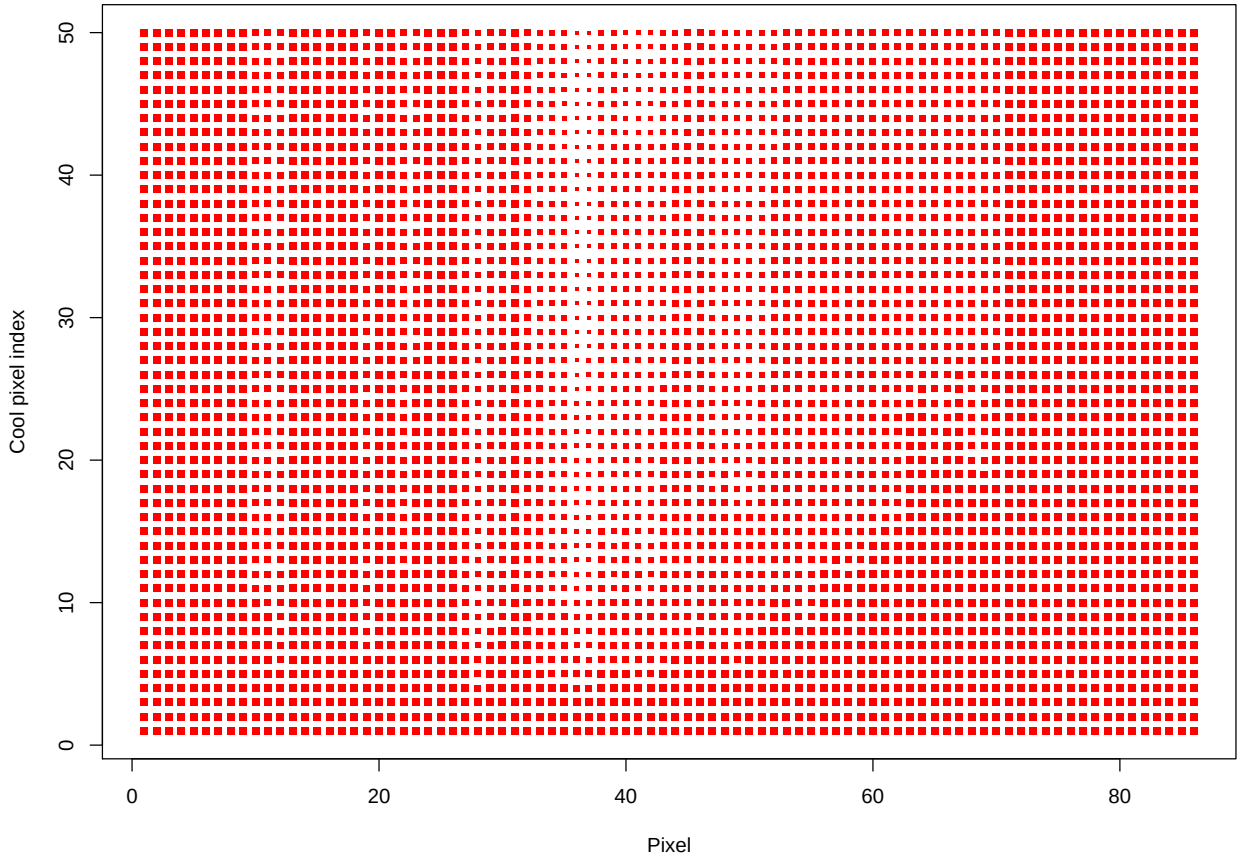

FIGURE 26: The performance of the classifier on one hundred quasar spectra damaged by the addition of a cool pixel. Axes, symbols and colours are similar to Figure 24, except that now the size of the symbols indicates the averaged probability that the sources are quasars, rather than stars.

the index i in Equation 14, of about 5 correspond to a pixel with a factor 0.67 lower in flux than the undamaged spectrum. Factors of approximately this size in the most sensitive pixels are sufficient to cause misclassification of stellar spectra. The number of sensitive pixels and the level of damage needed to cause misclassification is generally much lower for all three test classes than it was for the hot pixel data in the previous section. The quasars show no misclassifications at all due to cool pixels. The maximum factor by which the pixels can be reduced is 1/6, or 0.167, so the flux reduction is quite large and it is not clear that considering stronger reduction factors would be helpful.

Misclassifications of stars tend to be into the quasar class, and misclassifications of galaxies can be into a variety of other classes, at least including stars, binaries and quasars.

6.1.3 Flux calibration

The G magnitude is shifted from its original value, G_0, by one hundred different values between -0.5 and 0.5 magnitudes, in steps of 0.01 magnitudes. This affects the normalization applied to the data, and therefore rescales the whole spectrum when compared to the SVM models.

As with the hot pixels and cool pixels, one hundred of each main class of sources were classified and the results averaged to produce the output shown in Figures 27, 28 and 29. The plots show the averaged output probability for each main class as a function of the data degradation.

The plots for stars and galaxies show that the performance falls off dramatically with a scale of between 0.1 to perhaps 0.3 magnitudes. This fall-off is not necessarily entirely symmetric, and the performance seems to fall away more steeply on the 'bright' side than the 'faint' side. To understand this, we review the procedure for dealing with different magnitudes in DSC.

An input spectrum for DSC is assigned to the next faintest of the preprepared SVM models in the model grid. The spectrum is then normalised to the magnitude of the training data used to prepare that model. The main source of error in this process in normal circumstances is that the input spectrum will have slightly different noise characteristics to the training data used to build the model. For input spectra brighter than the training data, this does not make a crucial difference. For input spectra fainter than the model, the performance declines on a scale of about 0.5 to 1 magnitudes (Figure 2 of [11]).

If there is an error in the G magnitude, the spectrum will be wrongly normalised and will have too much or too little flux compared to the support vectors in the model. The SVM standardization will not correct for this problem, and we can expect the results to rapidly deteriorate. The tests on the DSC magnitude handling in [11] indicate that wrong flux normalization will cause problems for discrepancies larger than about 0.1 magnitudes, and this is what we see in Figures 27 to 29. The stars data apparently are misclassified after an offset of 0.1 to 0.2 magnitudes. The galaxies are similar to the stars and the quasars are misclassified after a factor of 0.2 to 0.4. As with the

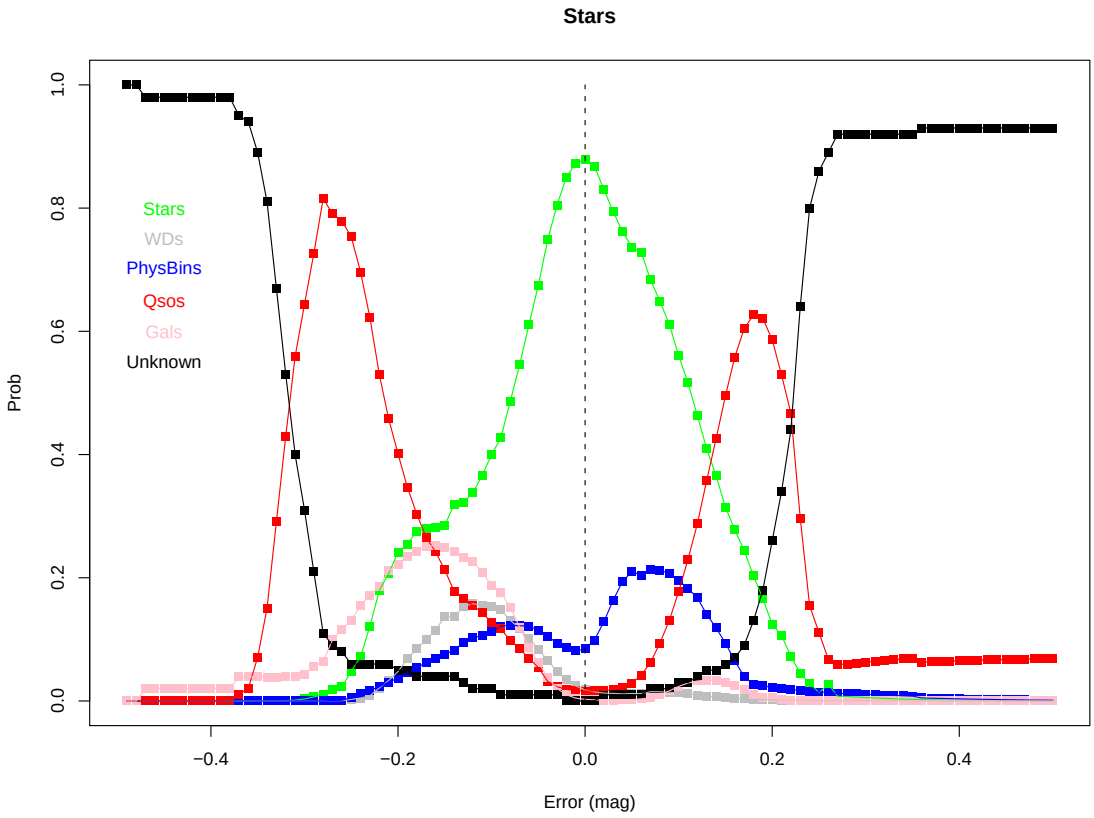

FIGURE 27: Variation of output probabilities with variation of the flux in the CalPhotSource table for one hundred MARCS stellar sources. The flux is varied by up to one magnitude in one hundred 0.01 magnitude steps around its true value. The probabilities for the different output classes are colour coded (see key in plot).

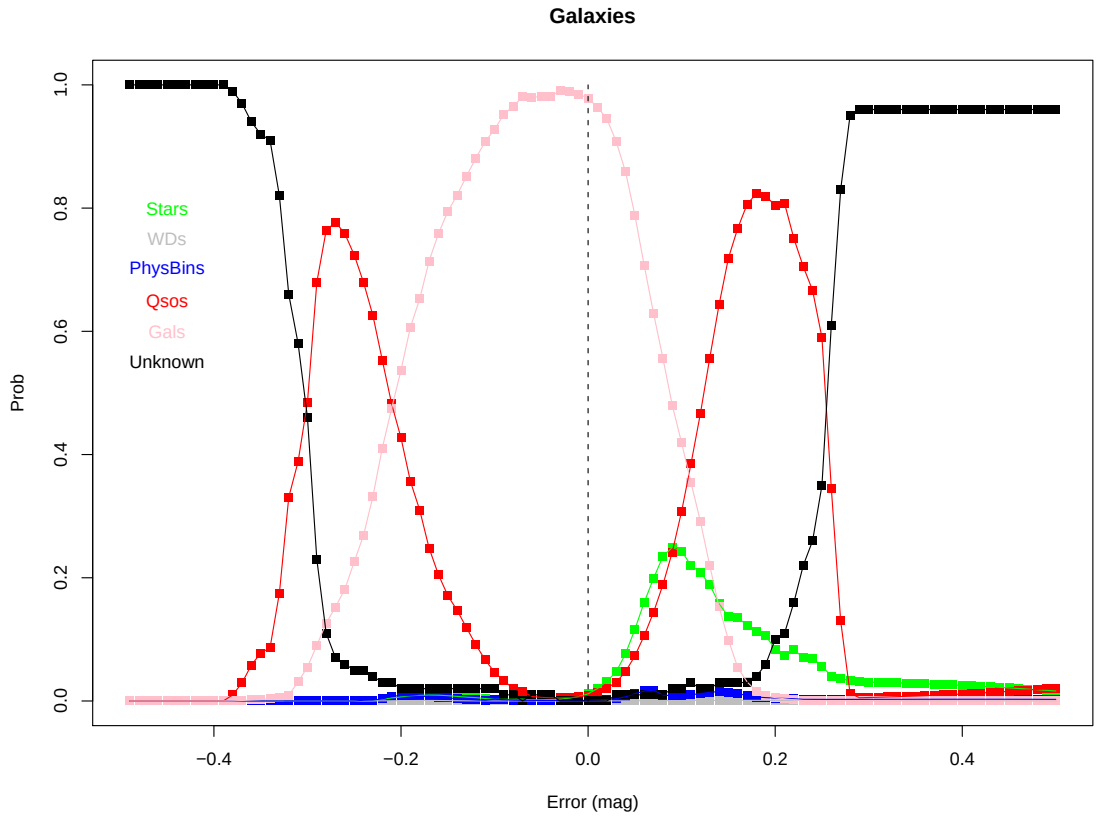

FIGURE 28: Variation of output probabilities with variation of the flux in the CalPhotSource table for one hundred Galaxies. The flux is varied by 1 magnitude in one hundred 0.01 magnitude steps around its true value. The probabilities for the different output classes are colour coded (see key in plot).

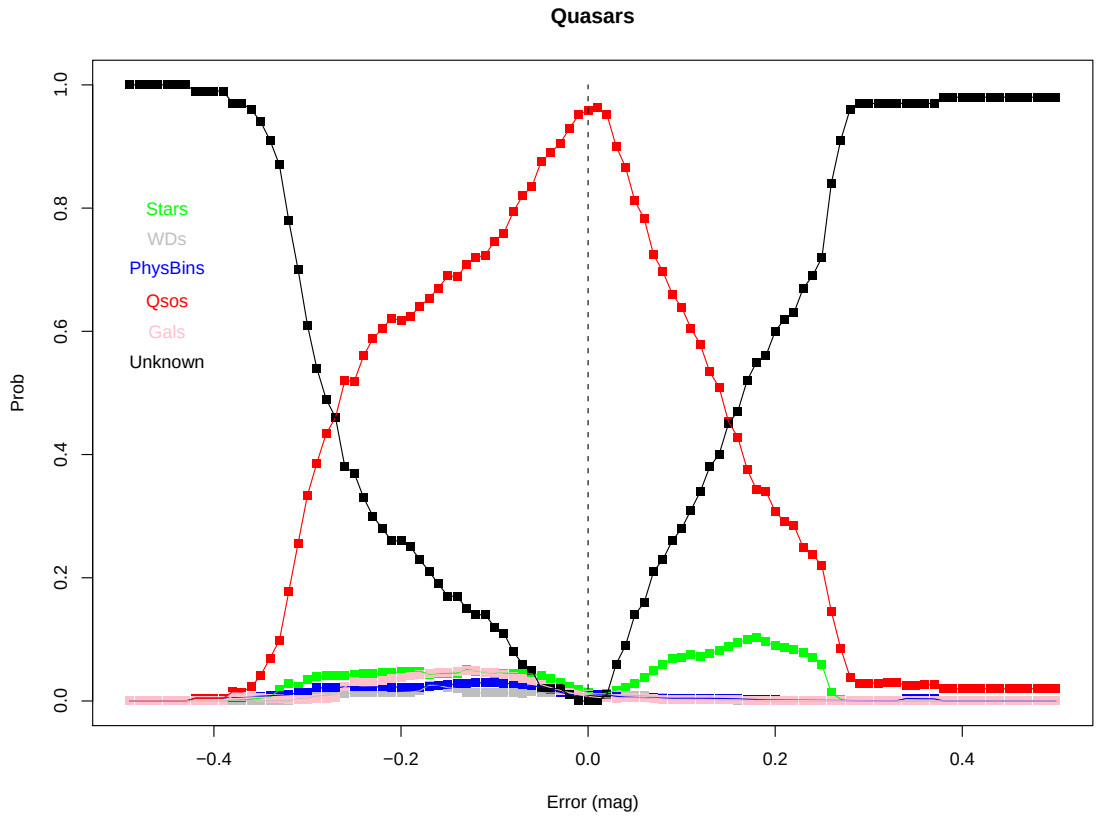

FIGURE 29: Variation of output probabilities with variation of the flux in the CalPhotSource table for one hundred Quasars. The flux is varied by 1 magnitude in one hundred 0.01 magnitude steps around its true value. The probabilities for the different output classes are colour coded (see key in plot).

hot and cool pixel data, we must bear in mind that the probabilities plotted are averages over one hundred sources, and the performance for individual objects could vary.

From the stars and galaxies plots, it seems that misclassification as a quasar is likely for modest magnitude offsets (0.1 - 0.3) whilst classification as UNKNOWN sets in for worse offsets. for the quasars, misclassifications seem to be overwhelmingly into the UNKNOWN category. This is consistent with the results from the hot pixel data.

6.1.4 Wavelength calibration

The spectrum is shifted from -1. resolution element (bluewards) to +1 resolution element (redwards) relative to the original. One hundred steps of 0.02 pixels are used. The shift is done by calculating new pixel fluxes from linear combinations of original fluxes from neighbouring pixels. A flux normalization is performed after the resampling to ensure flux conservation.

The results are shown in Figures 30, 31 and 32. These plots are similar to the flux calibration plots, and show the output probabilities for each class averaged over all one hundred input sources as a function of the spectrum shift. In all cases, strong effects on the output probabilities from DSC are seen with these $\sim$0.1 pixel shifts. The stars classification exhibits a strong peak with a scale of 0.1 pixels or so. The Quasar and galaxy classifications are both slightly more robust. Both seem also to be more robust to the positive shift side, which corresponds to the spectrum being shifted to the red. This may be due to the fact that the Galaxy and Quasar training sets include redshifted objects.

The stars spectra, when shifted bluewards (negative shift), show a tendency to be misclassified as quasars. When shifted redwards, there is a tendency to misclassify as white dwarfs. The galaxies are most likely misclassified as quasars for modest pixel shifts in both directions, and the quasars are classified as UNKNOWN rather than into any other astrophysical class. For large shifts (order 1 pixel) in either direction, the stars and galaxies are also classed as UNKNOWN.

6.1.5 Summary of the robustness test

The tests indicate that the DSC performance can be quite sensitive to even relatively small variations in the pixel-to-pixel response. Variations of 0.1 magnitudes in flux calibration, or shifts of about 0.1 pixels in the dispersion solution between the training data and the evaluation data, can also lead to misclassifications. We are working on some ways to mitigate some of these effects.

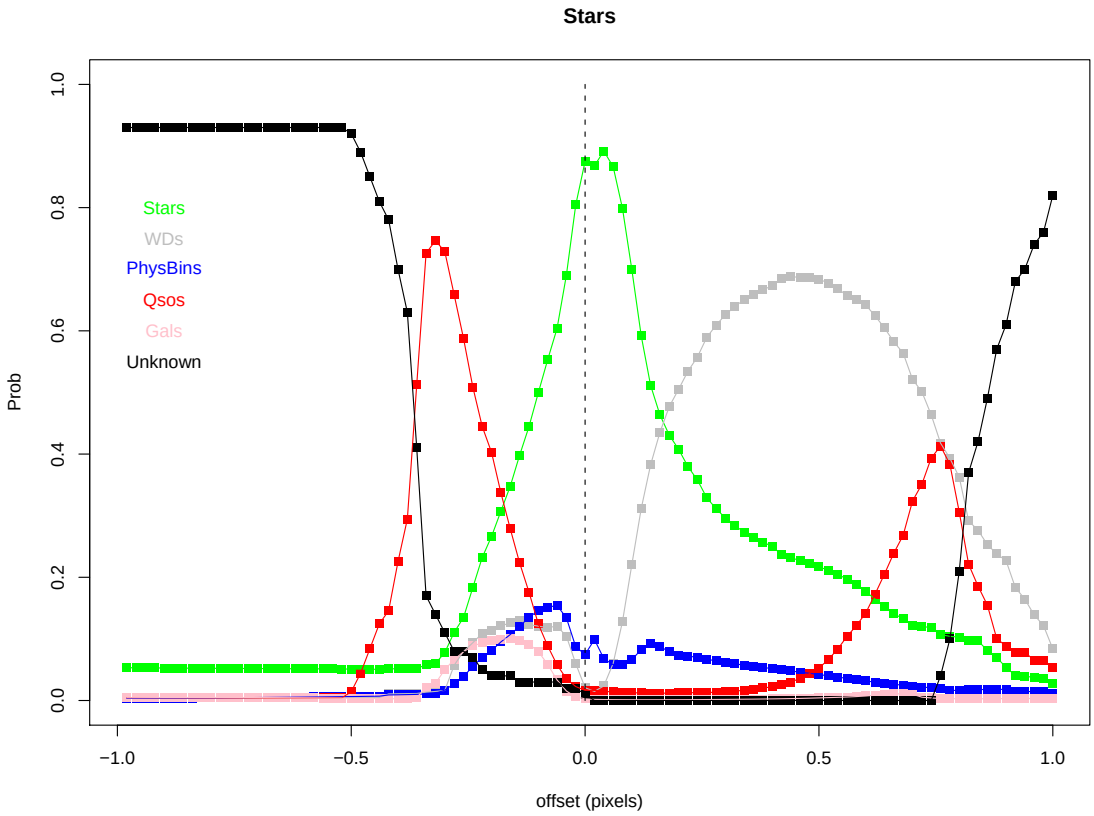

FIGURE 30: The performance of the BPRP subclassifier for one hundred stellar spectra with the pixels shifted by between -1 and +1 resolution elements in steps of 0.02 (100 steps). Green symbols show $P(Star)$ (the true class), red $\rightarrow$ $P(Quasar)$, blue $\rightarrow$ $P(PhysBinary)$, grey $\rightarrow$ $P(Whitedwarf)$, black $\rightarrow$ $P(Unknown)$, pink $\rightarrow$ $P(Galaxy)$.

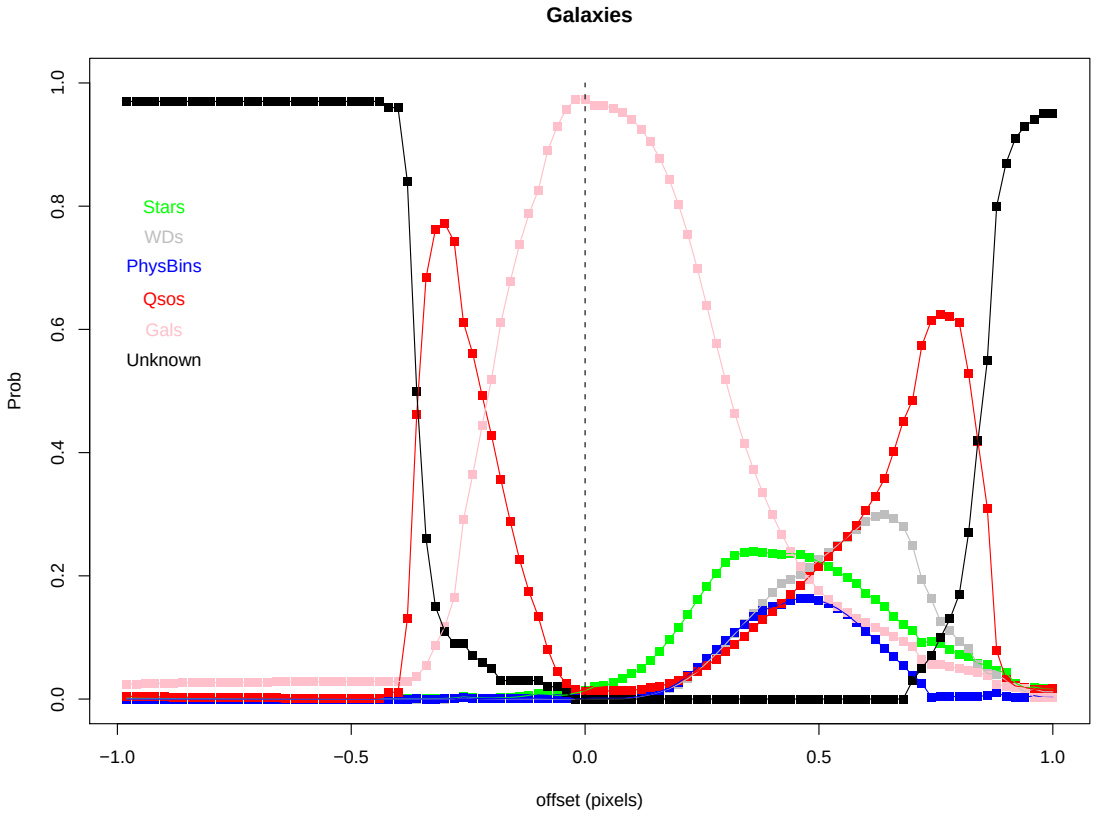

FIGURE 31: The performance of the BPRP subclassifier for one hundred galaxy spectra with the pixels shifted by between -1 and +1 pixels in steps of 0.02 (100 steps). Colours have the same meaning as for Figure 30.

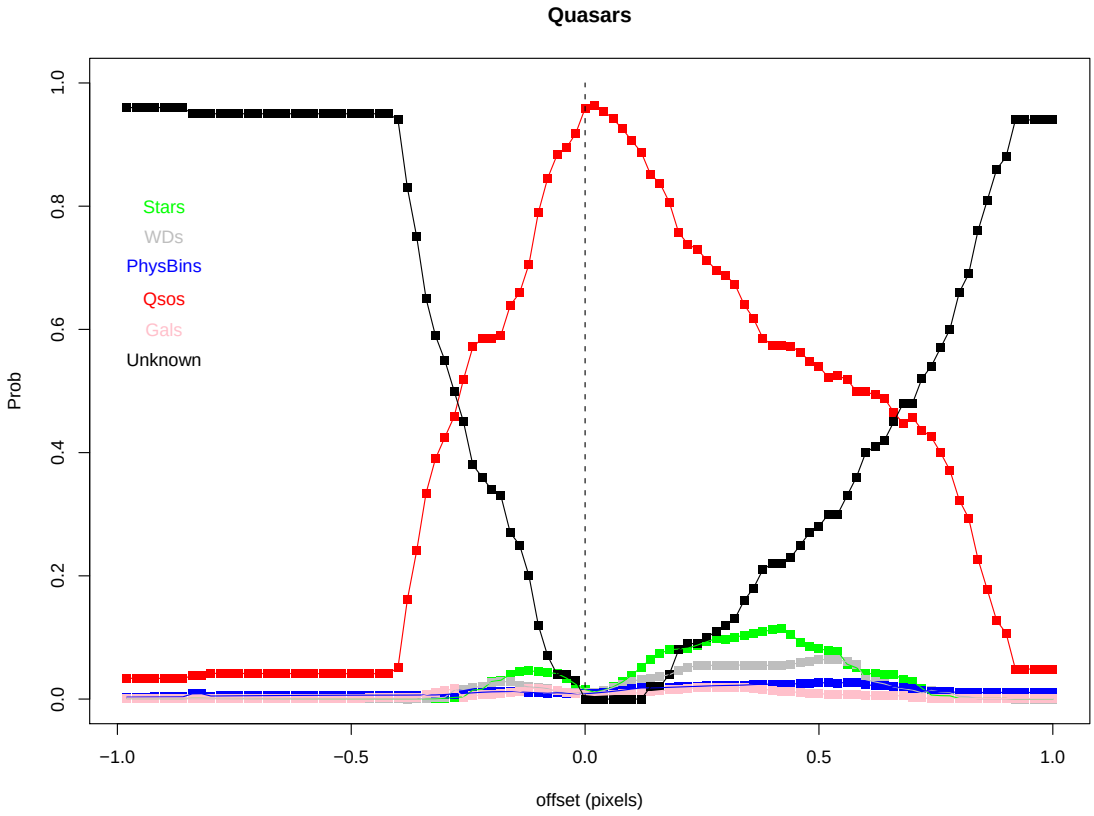

FIGURE 32: The performance of the BPRP subclassifier for one hundred quasar spectra with the pixels shifted by between -1 and +1 pixels in steps of 0.02 (100 steps). Colours have the same meaning as for Figure 30.

7 Summary, conclusions and future work

The DSC currently consists of three working subclassifiers. The photometric subclassifier returns a probability based on the appearance of the BP and RP spectra, and is based on Support Vector Machines. The Astrometric Classifier returns a probability based on the proper motions and parallax, and is based on a Gaussian mixture model. The Position-Gmag classifier returns a probability based on the sky position and G magnitude of the source, and in the version of the code presented here, is based on a simple parametric model. The probabilities from all these subclassifiers are combined to produce a combined probability vector for all the possible output classes.

Tests with the current classifier reveal that the completeness for most libraries exceeds 90%. Contamination into incorrect astrophysical classes is below 1% except for the APec stars, which have a $\sim 4\%$ contamination into the quasars class. The ultracool dwarfs and WR stars libraries have relatively low completeness. In the case of the ultracool dwarfs, this is consistent with a trend to less accurate classification seen for the main cool stars libraries at lower temperatures (Section 4.4). Tests show that, for the ultracool dwarfs at least, there could be a problem with the sparseness of the training set used. We will investigate using active learning or a similar technique to improve the training sets and attempt to improve the completeness in these grids.

In previous cycles, classes such as the white dwarfs and binary stars have given results with significantly worse completeness than the other libraries. It is hoped that the combination of position-Gmag and, particularly, astrometric information can improve these results. We will test this in the next development cycle, when we will once again have simulated data covering these libraries.

In summary, the immediate improvements which will be attempted for DSC in the coming year are;

- Run tests with a complete set of test cases including Binaries and White dwarfs
- Run tests with semi-empirical data (from SDSS) tested on models trained on synthetic data (e.g. Phoenix for the stars). In the next data cycle we will have both semi-empirical and synthetic libraries for normal stars, quasars and galaxies.
- Improve the parametric model for the position-Gmag classifier, and calibrate it against known Quasar and Galaxy populations.
- Investigate using active learning or similar techniques to build the training data sets, and whether this will help particularly with rare objects such as rare types of stars.

8 References

[1] A.-M. Janotto, C.A.L. Bailer-Jones, S. Chastel, et al. CU8 Software Design Description. Scientific Chains, February 2010.

[2] C.A.L. Bailer-Jones and K. Smith. Combining probabilities, July 2011.

[3] C. Liu, A.-M. Janotto, C.A.L. Bailer-Jones, et al. CU8 Scientific Algorithms Software Design Description, February 2011.

[4] C.A.L. Bailer-Jones. Developing further the Discrete Source Classifier, October 2008.

[5] B. Schoelkopf, J. Platt, J. Shawe-Taylor, A.J. Smola, and R.C. Williamson. *Neural Computation*, 13:1443, 2001.

[6] J.C. Platt. In A.J. Smola, P. Bartlett, and D.S. Schoelkopf, editors, *Advances in large margin classifiers*. MIT press, 1999.

[7] T.-F. Wu, C.-J. Lin, and R.C. Weng. *Journal of Machine Learning Research*, 5:975, 2004.

[8] K.W. Smith. A Nelder-Mead tuner for Svm, March 2009.

[9] R. Sordo and A. Vallenari. Description of the CU8 Cycle 5 simulated data, October 2009.

[10] R Sordo and A. Vallenari. Description of the CU8 Cycle 6 simulated data, May 2010.

[11] K. W. Smith. DSC Software test report, February 2011.