Publications of the Astronomical Society of Australia (PASA), Vol. 33, e015, 30 pages (2016).

doi:10.1017/pasa.2016.13

Comparison of Three Statistical Classification Techniques for Maser Identification

Ellen M. Manning[1], Barbara R. Holland[1], Simon P. Ellingsen[1,5], Shari L. Breen[2], Xi Chen[3,4] and Melissa Humphries[1]

[1]School of Physical Sciences, University of Tasmania, Private Bag 37, Hobart, Tasmania 7001, Australia
[2]CSIRO Asronomy and Space Science, PO Box 76, Epping, NSW 1710, Australia
[3]Key Laboratory for Research in Galaxies and Cosmology, Shanghai Astronomical Observatory, Chinese Academy of Sciences, Shanghai 200030, China
[4]Key Laboratory of Radio Astronomy, Chinese Academy of Sciences, Beijing 100012, China
[5]Email: Simon.Ellingsen@utas.edu.au

(Received December 23, 2015; Accepted March 11, 2016)

Abstract

We applied three statistical classification techniques—linear discriminant analysis (LDA), logistic regression, and random forests—to three astronomical datasets associated with searches for interstellar masers. We compared the performance of these methods in identifying whether specific mid-infrared or millimetre continuum sources are likely to have associated interstellar masers. We also discuss the interpretability of the results of each classification technique. Non-parametric methods have the potential to make accurate predictions when there are complex relationships between critical parameters. We found that for the small datasets the parametric methods logistic regression and LDA performed best, for the largest dataset the non-parametric method of random forests performed with comparable accuracy to parametric techniques, rather than any significant improvement. This suggests that at least for the specific examples investigated here accuracy of the predictions obtained is not being limited by the use of parametric models. We also found that for LDA, transformation of the data to match a normal distribution led to a significant improvement in accuracy. The different classification techniques had significant overlap in their predictions; further astronomical observations will enable the accuracy of these predictions to be tested.

Keywords: masers – methods: classification – stars: formation

1 INTRODUCTION

In recent years, astronomical instrumentation across a range of wavelength bands has improved to the point where high-resolution, sensitive surveys of large areas of the sky are becoming much more common (e.g. Benjamin et al. 2003; Johnston et al. 2007). The higher data rates from new instrumentation and large surveys give the opportunity to collect detailed information on very large numbers of sources and undertake more sophisticated statistical investigations of their properties. This will enable both more reliable identification of sub-groups within the broader population, and identification of rare or unusual objects. However, these new instruments also present the astronomical community with a challenge of how best to extract the maximum utility from large volumes of data.

The desire to accurately and efficiently classify astronomical sources identified in large surveys into different groups is an increasingly common one. Attempts to develop efficient criteria for targeted searches for interstellar masers, is one specific example of an application of survey source classification. A number of studies have found that star-formation regions with an associated interstellar maser differ significantly in their infrared or millimetre continuum properties from the majority of the population (e.g. Ellingsen 2005, 2006; Chen et al. 2011). In developing criteria for targeting future searches, it is desirable to identify a large fraction of the population of interest while including only a small number of sources which do not yield detections. In the terminology of classification, it is important to minimise both the number of false-negatives and false-positives. A related issue is in understanding the characteristics through which the classification has been achieved. For example, if you are able to develop efficient criteria for targeting a search for interstellar masers on the basis of infrared or millimetre continuum properties, what is the physical meaning of those characteristics—do they correspond to a particular mass range, or evolutionary phase of the associated high-mass star-formation region?

Maser emission occurs naturally in a range of astrophysical environments, including the molecular gas close to newly forming stars, the envelopes of late-type stars, and close to the nuclei of some active galaxies. Masers have proven to be a reliable signpost of the very early stages of high-mass star formation (e.g. Ellingsen 2006); with recent improvements in the availability of sensitive large-area surveys at mid-infrared through millimetre wavelengths, they are increasingly being used as tools to study high-mass star formation (e.g. Titmarsh et al. 2013). Masers can provide information on the dynamics of the star-formation region through observations of their kinematics (e.g. Goddi, Moscadelli, & Sanna 2011), on the magnetic field from observations of the polarisation (e.g. Surcis et al. 2012), and potentially the presence and absence of different transitions can provide an evolutionary timeline (e.g. Ellingsen et al. 2007; Breen et al. 2010). If it is possible to use classification techniques to reliably identify which regions host different types of maser transition, then an understanding of the physical properties of those regions in combination with the maser-based evolutionary timeline could provide important insights into the formation of high-mass stars.

These types of classification problems are commonly encountered in a wide range of scientific disciplines, and from the broader literature we have been able to identify a number of commonly used classification techniques. When considering different classification methods, Breiman (2001b) has suggested that there is a trade off between parametric techniques that are easy to interpret but not always as accurate, and non-parametric methods that are more difficult to interpret, but deliver a higher level of accuracy. Here, we use three different classification techniques to investigate their strengths and weaknesses when applied to the specific problem of efficiently identifying target sources for searches for interstellar masers. The three methods we have chosen for our investigation are linear discriminant analysis (LDA), logistic regression, and random forests. These three methods were chosen because they have proven effective across a wide range of problem domains, they are relatively easy to implement, and they include two parametric methods (LDA and logistic regression) and one non-parametric method (random forests).

LDA uses similar calculations and techniques to principal component analysis (PCA) which is quite widely used in astronomy (e.g. Lo et al. 2009; Einasto et al. 2011). Kobel et al. (2009) used LDA in their classification of different photospheric magnetic elements on the Sun. They found that the predictions they were able to make on the basis of LDA showed good agreement with the results from previous studies. This can in part be credited to the semi-artificial segregation between the classes of photospheric magnetic elements, as the variables chosen were those with the most significant differences in brightness values. Logistic regression has been less commonly applied in astronomy than PCA, although it has previously been used to successfully identify which star-formation regions are more likely to host different types of interstellar masers (e.g. Breen et al. 2007; Ellingsen et al. 2010). Yuan et al. (2010) and Song et al. (2009) have both shown that logistic regression can be an effective means of predicting solar flares. Random forests are a relatively new, non-parametric classification technique which has proven to be very effective in other fields, such as ecology. Cutler et al. (2007) compared the results of classifying ecological data, using the same classification methods as are used here and found that random forests had the highest accuracy. Within astronomy, random forests have been used by Bailey et al. (2007) to improve the reliability of finding supernovae from images, while Carliles et al. (2010) used them to assign photometric redshifts. Recently, they have also been used as the basis of processes for automated rapid classification and decision making. Morgan et al. (2012) used random forests as part of a method for making time-efficient recommendations as to which gamma-ray burst events are likely to be high-redshift in order to prioritise whether a specific event deserves additional observing time. They found that by observing the top 20% of recommended events, it was possible to identify 56% of the high-redshift bursts, while using the top 40% of recommendations allows identification of 84% of high-redshift events. Mirabal et al. (2012) used random forests to accurately classify whether unidentified objects detected in Gamma-rays by the *Fermi* satellite were likely to be Active Galactic Nuclei (AGN) or pulsars (they achieved accuracies of 97.7 and 96.5% for AGN and pulsar identification, respectively).

To better understand the strengths and limitations of these different classification techniques, both in terms of their efficiency and the degree to which the outcomes of the classification process can be related to the properties of the astronomical sources, we compared their performance on three published datasets (Breen et al. 2007; Ellingsen et al. 2010; Chen et al. 2012). For each of these three sets of data, we applied the three classification techniques to make predictions as to which infrared (or millimetre) sources are likely to also be associated with masers. In Section 2, we describe in more detail each of the classification techniques used. The properties of each of the datasets are outlined in Section 3 where we examine the results of applying the different classification techniques in each case.

2 CLASSIFICATION TECHNIQUES

In the context of the current work, our data typically consists of astronomical sources for which a range of parameters (e.g. the intensity in a particular wavelength range) have been measured, along with parameters which are related to the quality or uncertainty in the measurement and others which identify the particular astronomical object (e.g. the source number or coordinates). These parameters are all potential inputs to the different classification techniques and we refer to these as predictor variables. In the field of machine learning, these are often referred to as features, however, as that term frequently has a different meaning in astronomical literature

we do not use that terminology here. For some (sometimes all) of the sources in the data set, we also have information as to whether or not that source has an associated maser emission from a specific molecular transition. Hence, we are seeking to accurately classify our astronomical sources into two classes, those with an associated interstellar maser and those without.

2.1. Linear discriminant analysis

LDA finds the linear combination of predictor variables which maximises the separation of the different classes and minimises the variation within classes (Feigelson & Babu 2012). LDA can be visualised geometrically as projection from a high-dimensional space onto a line. When given a new source to classify, LDA uses this linear combination to convert the high-dimensional data to a real number, and the classification of the sample is determined by comparing this number to a threshold value. The technique is relatively simple and so is unsuitable if there are complex, non-linear interactions between the variables. LDA is a technique of dimensionality reduction similar to PCA, which is more commonly used in astronomy. Both LDA and PCA attempt to model the data with linear combinations of the predictor variables; the difference is that PCA does not use classification information in producing the model, whereas LDA does (Feigelson & Babu 2012).

The assumptions of LDA are that the data follows a multivariate normal distribution for each class, classes may have different means but are assumed to have the same variance structure. This makes LDA a parametric method in the sense that it assumes a particular model of the data. Most astronomical data are not normally distributed, so transformations of the variables are usually required. For each of the three datasets we studied, LDA was applied to both the original data and to the transformed data as a comparison. Data Set 2 required an inverse function to normalise the data (each predictor variable was transformed via a $\frac{1}{x}$ function). In the case of Data Sets 1 and 3 where the samples were naturally clustered, an inverse transformation would have destroyed the bimodality present. For this reason, a log transformation was selected as it improved the normality of the data while still being easy to interpret.

LDA models were fitted using the `lda` function in *R* (R Core Team 2013, part of the `MASS` package), we left the `prior` input parameter at its default setting which is to assume that probability of being in a particular class is equal to the relative frequency of the class in the training data.

2.2. Logistic regression

Logistic regression is a form of generalised linear modelling (GLM) that is used to predict the probability of an event occurring; in this case, whether or not an astrophysical source has an associated interstellar maser. The probability of occurrence P is calculated from

$$P = \frac{1}{1 + e^{-z}},$$

where $z = b_0 + b_1.x_1 + \cdots + b_n.x_n$, the b values are regression coefficients and the x_i values are the predictor variables. P is then compared to a cut-off threshold of 0.5 (50% likelihood) to determine whether an object is predicted to have an associated maser, or not. Like LDA, logistic regression is also a parametric method.

Linear regression assumes that the response variable is normally distributed, in contrast logistic regression assumes that the response variable follows a binomial distribution (which is applicable in our case of two classes). This means that the method of least squares (used in linear regression), cannot be applied to logistic regression (Hosmer & Lemeshow 2000). Instead, maximum likelihood (formulated by Fisher 1922) is used to estimate the parameters of the model. The likelihood function is calculated using the product of contributions to the model from each of the predictor variables (Hosmer & Lemeshow 2000).

Logistic regression was implemented using the function `glm` which is part of the base *R* package (R Core Team 2013). To perform a logistic regression, the family option in `glm` is set to *binomial* and the link function is set to *logit*. It was not feasible to alter any other input parameters in the function to produce our models.

2.3. Random forests

Classification trees are a non-parametric technique of classification (in contrast to both logistic regression and LDA), which means that they do not assume an underlying model of the data (Cutler et al. 2007; Carliles et al. 2010). Classification trees can be more accurate than parametric approaches when complex interactions occur between the predictor variables. This could be the expected case for maser association with infrared or millimetre sources, as well as a broad range of astronomical classification problems. Individual classification trees may not be very accurate, especially when there are more than a few predictor variables, however, a collection of trees grown independently on randomly perturbed versions of the data greatly increases the accuracy of predictions (Breiman 2001c; Cutler et al. 2007). Random forests work by producing large numbers of classification trees and then determining the classification of a particular sample (in our case, an astronomical source) by allowing each of these trees to 'vote' and then taking the majority rule (Breiman 2001a). This voting system is also how the probability of a sample being classified into a certain group is calculated; by dividing the number of trees voting for a certain classification by the total number of trees.

To produce individual classification trees in a random forest, a bootstrap sample is selected for each tree. For a data set with N entries, N samples are taken. Because sampling is

done with replacement, approximately two-thirds of the original data occurs at least once in each bootstrap sample (Efron & Tibshirani 1994). Hastie, Tibshirani, & Friedman (2001) showed that bootstrap sampling causes the variance of the estimated class to converge to a lower limit when more trees are added to the forest, and so rarely overfit (Breiman 2001c). A classification tree is grown from each bootstrap sample using recursive binary partitioning. The branching points of the trees are called nodes. In standard trees, the predictor variable at each node is chosen based on the best split, which is determined by the Gini index (a measure of statistical dispersion, see Hastie et al. 2001, pg 271). In a random forest, the variable providing the best split is chosen from a random subset of predictor variables (Liaw & Wiener 2002). The predictor variable and the subset of predictor variables from which it is chosen is independent of any other nodes' variable choices. This approach decreases the dependence between individual trees. The splitting process continues until further subdivision no longer decreases the Gini index. The final classification given by each tree depends on the terminal node the source has been allocated to.

A nice feature of random forests is that they have an in-built way of estimating the classification error because of the use of bootstrapping to select slightly different data for each tree. Data not included in the bootstrap sample (approximately one-third of observations) for a particular tree are referred to as out-of-bag (oob) values. The tree grown from each bootstrap sample is used to predict the classification for each of the oob values, giving an estimate of the classification error as well as a means to compare the importance of each variable in the classification process (Breiman & Cutler 2013). The importance of a variable is expressed by the difference between the probability of predicting the class correctly in shuffled oob data (the sample order is rearranged to eliminate systematic errors) compared to the unshuffled oob data (Cutler et al. 2007).

Random forests also give a natural metric for determining the similarity of two different astronomical sources (or other groups of samples). Proximities between two sources are calculated in the random forest process. If a pair of sources end up in the same terminal node, their proximity is increased by one. Similar source pairs end up in the same terminal node more often than dissimilar ones. The proximities are then normalised (divided by the total number of trees) and the proximity of a point and itself is set to be one. The proximities are then expressed as a symmetric matrix, where the diagonal entries all have the value one. The proximity matrix can be used as input for multi-dimensional scaling (MDS), as a way of visualising the classification results (displayed in Section 3).

A potential drawback of random forests is that they cannot be used to directly test hypotheses (Cutler et al. 2007). They also do not give a clear representation of the actual classification process. However, although the internal calculations are difficult to interpret, they produce useful properties such as relative variable importance and an estimate of the classification error without extra external calculations (Breiman & Cutler 2013).

To create the random forests used in the modelling and classification, we used the *R* function `randomForest` (in the `randomForest` package). For an introduction to the usage and features of `randomForest` functions in the *R* environment, see Liaw & Wiener (2002). There are a number of parameters that can be varied when growing the random forest in order to optimise its classification and predictive accuracy. These include the number of trees in the forest, the number of variables randomly sampled as candidates at each split, and the maximum number of terminal nodes in the trees. The minimum size of the terminal nodes can also be varied, where a larger number leads to smaller trees which take less time to grow. Setting the node size to k means that no node with fewer than k cases will be split (Breiman & Cutler 2013). A terminal node size of 1 is therefore the most accurate, but in cases with large datasets, memory constraints may require this to be higher. We found that altering these parameters did not consistently increase the sensitivity or specificity significantly, so the default values for the parameters were used: 500 trees grown in the forest, a node size of 1 (default for regression is 5), and the maximum possible number of terminal nodes. The default number of variables chosen at each split is $\sqrt{p}$ for classification and $p/3$ for regression (rounded to the nearest integer), where p is the total number of predictor variables in the data set. Other factors that can be varied are whether or not the cases are sampled with replacement (the default, which we used, is with replacement), and the prior probability of each class occurring can also be set with the default being to assume equal class probabilities.

For both Data Set 2 and 3 (where predictions were done), random forests were grown using 3 000 trees rather than the default 500. Since each tree is grown independently, this is equivalent to combining the results of multiple smaller forests. 3 000 trees was chosen for both data sets because this produced the most accurate results in the cross validation. Generally, random forests is robust against over-fitting (see Breiman 2001c), however, in the case of Data Set 2, due to the very small training set compared to its number of predictor variables, more than 3 000 trees decreased the classification accuracy. In the case of Data Set 3, using more than 3 000 trees had no effect.

2.4. Accuracy of classification techniques

There are four possible outcomes of the classification of each astronomical source. The two desired outcomes are that the classification technique can correctly identify a source which does have an associated maser (a 'true positive'), or it can correctly identify a source as not having a maser (a 'true negative'). A perfect classification would have all samples with one or the other of these outcomes. There are however, two ways in which the classification scheme can give an incorrect outcome and depending on the circumstances these

Table 1. The relationship between the four possible classification results and the calculated values of the sensitivity and the specificity.

	Know negatives	Known positives
Classified as negative	True negatives	False negatives
Classified as positive	False positive	True positive
	Specificity (true negative rate)	Sensitivity (true positive rate)

are not necessarily of equal importance. A 'false positive' outcome is where a source which does not have an associated maser is classified as being associated with one, while a 'false negative' occurs when a source which does have an associated maser is classified as not having one associated (see the Confusion Matrix in Table 1).

For each classification method, we calculated both the sensitivity (known as recall in machine learning) and specificity. In this context, the sensitivity, or true positive rate (TPR), is the percentage of maser associations correctly predicted by the model, and specificity is the percentage of maser non-associations correctly predicted, or the true negative rate (TNR).

$$\text{Sensitivity (TPR)} = \frac{\text{True Positives}}{\text{True Postives + False Negatives}}$$

$$\text{Specificity (TNR)} = \frac{\text{True Negatives}}{\text{True Negatives + False Positives}}$$

2.4.1. *Predictor variable importance*

Logistic regression performed using *R* has two techniques for determining the importance of the variables included in the model. The first is a set of *p*-values provided when the logistic regression is performed. The second is the in-built `stepAIC` function, which includes all possible predictor variables in the starting model and iteratively removes variables which do not significantly contribute to the model to yield the most parsimonious model with the greatest predictive power. To determine which variables to include in the logistic regression models, we used a combination of the `stepAIC` function and manual variable selection. Variables that did not increase the accuracy of the model were excluded (see Section 3).

For LDA, variable selection was done manually. We used the logistic regression's selection as a starting point, and then included additional predictors if they improved the prediction accuracy.

Random forests includes an internal calculation of the mean decrease in accuracy for each of the variables utilised, which is a measure of how poorly the model performs when that variable is not included. Thus, the higher the value is, the more the predictor variable contributes to the accuracy of the model. Negative values decrease the accuracy and values close to zero offer little or no effect. It is worth noting that random forests is potentially robust enough to deal with all available variables and so including them all in the model does not generally decrease the accuracy significantly (Feigelson & Babu 2012).

2.4.2. *Cross validation*

The aim of classification is to build models that will generalise well to new data. When constructing models, there is a danger in over-fitting to the training data. In order to determine the accuracy of each of the classification methods on the three data sets, we used a 10-fold cross validation technique. Using a fitted model that has been trained on a randomly chosen 90% of the data, the classification of the remaining tenth is predicted. This procedure of training and prediction is then repeated 1 000 times in order to obtain an estimate of the classification error. Repeating the cross validation ensures that a high number of the possible combinations of the data are used, reducing sampling bias associated with randomly folding the data. Repeated 10-fold cross validation of this kind is especially useful when modelling a random forest as the over-fitting associated with regression tree techniques is compensated for by the generous error estimation of the cross validation (Borra & Di Ciaccio 2010).

In repeated 10-fold cross validation, the results from the multiple runs are averaged. In this case, the averaged cross validation produced a mean probability of being associated with a maser for each sample. A source was classified as a maser if the probability was 50% or above. The percentage of predicted classifications were then compared to the actual classifications (maser source or non-maser source) to determine the accuracy for each model for each of the three data sets. Adjusting the cut-off threshold for maser classification from 50% was also investigated to explore the trade-off between sensitivity and specificity of the model. This is useful information to have available when it is important to obtain all the positive classifications, even when it means many false positives are given, and alternatively the model can be adjusted so that there is only a very small chance of a false positive, at the expense of false negative classifications. The receiver operating characteristic (ROC) curves (explained in Section 2.4.3), display the results of this analysis.

2.4.3. *Receiver operating characteristic curves*

An ROC curve plots the TPR (sensitivity) against the false positive rate (1 − specificity), effectively showing the trade-off in prediction power for accuracy in a given classification model. The diagonal line $y = x$ represents randomly classifying the samples, with half predicted as positive and half as negative. Anywhere in the space above, this line means that the model is better than random classification, with the best possible system showing 100% sensitivity with no false predictions, resulting in a point in the top left-hand corner. ROC curves were plotted here to compare each classification method for each data set in Figures 2, 5, and 8.

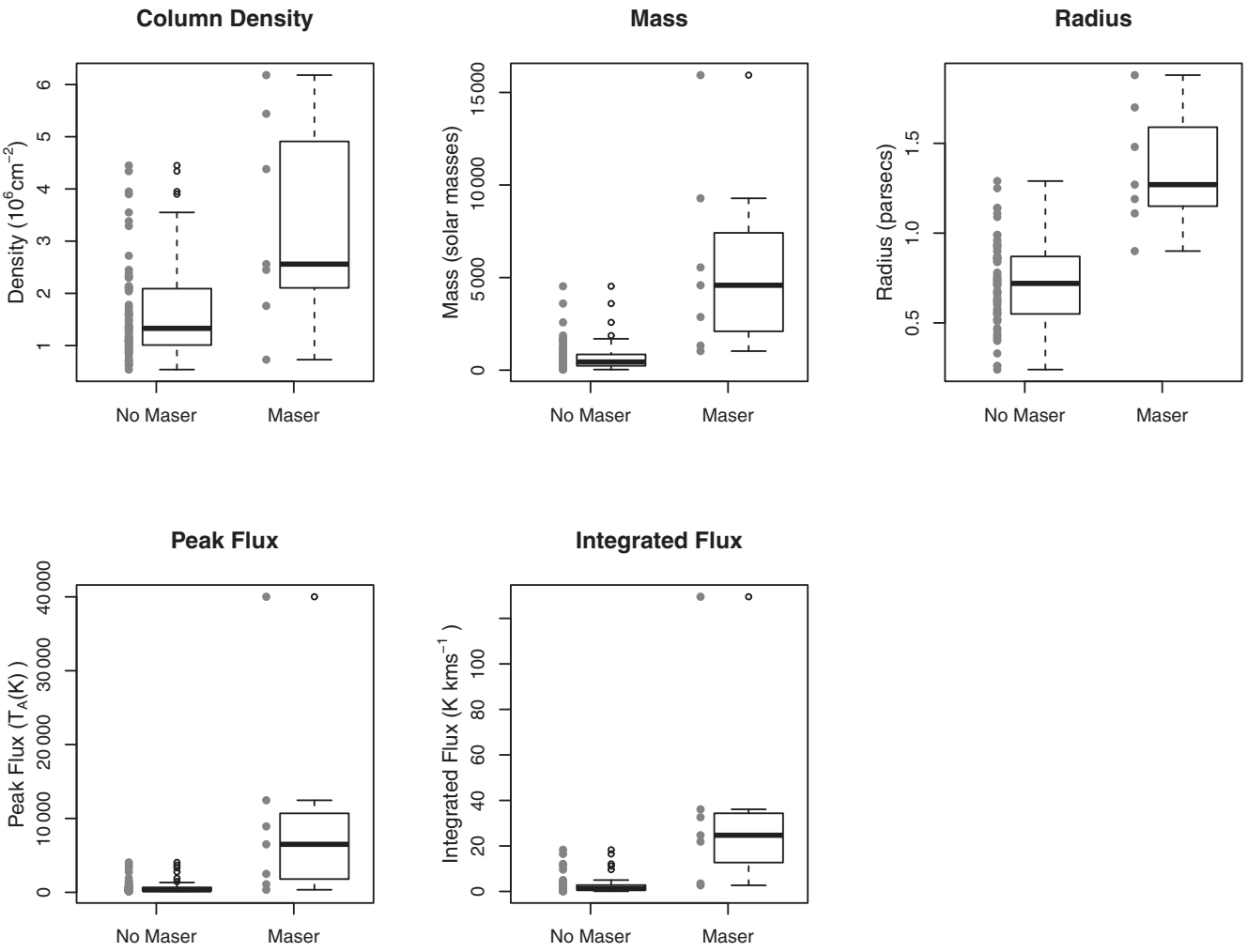

Figure 1. Boxplots comparing the variables from Data Set 1 between the sources with an associated water maser and those without. The outline in each of the boxplots represents the range between the first and third quartiles, with the median being the solid line horizontally through the box. The vertical lines outside the box extend to the minimum and maximum values, with any outliers (values separated from the quartiles by more than one and a half times the interquartile range) shown separately as dots. In this case, due to the very small number of samples being associated with a maser in this data set, the individual sample points are also plotted.

3 RESULTS

3.1. Water masers associated with star-formation regions in the RCW106 giant molecular cloud

Breen et al. (2007) undertook a complete search for 22 GHz water masers within the giant molecular cloud RCW 106. This search detected nine 22 GHz water masers and the region searched included 73 1.2-mm dust clumps observed and characterised by Mookerjea et al. (2004). Seven of the dust clumps were found to be associated with masers (Breen et al. 2007). Breen et al. used a form of logistic regression called binomial GLM to investigate the properties of the astronomical sources (in this case dust clumps) with and without water masers in RCW106. They found that water masers were associated with those sources which are denser, more massive, and have higher luminosity.

There are clear differences in the values of all the predictor variables between those sources with an associated water maser and those without, as is demonstrated by the boxplots shown in Figure 1. However, it should be noted that there are varying degrees of overlap in the ranges observed for the maser associated sources and those which are not. The obvious difference in the distributions for all the predictor variables means that we might expect that they should all contribute to the classification and that the relative importance might also be similar. The variable importance ratings returned by the random forest classification are a measure of the degree to which the classification trees utilised each predictor variable. The five predictor variables available as inputs for the classification process were : peak flux density, source radius, total integrated flux density, dust mass (calculated assuming a temperature of 40 K and optically thin dust emission), and column density. Using only source

Table 2. The predictor variables that increased the classification accuracy of the various methods for Data Set 1. Random forests provides an internal calculation of the mean decrease in accuracy (the higher the value, the more important the variable), logistic regression provides p -values (the lower the value, the more significant the variable's contribution to the model), and LDA provides no internal measurement of the importance of each variable, so it is just noted which variables were used (see Section 2.4.1).

	Random forests	Logistic reg.	LDA	Norm. LDA
Radius	10.68	0.1388	Y	Y
Int. flux	16.35	0.2485	Y	Y
Density				Y

Table 3. The results of cross-validating random forests, logistic regression, and LDA (without and with transformation of the predictor variables) classification and prediction for Data Set 1.

	Random forests	Logistic reg.	LDA	Norm. LDA
True neg.	66	65	66	65
False pos.	0	1	0	1
False neg.	2	2	3	2
True Pos.	5	5	4	5
Specificity (%)	100	98.5	100	98.5
Sensitivity (%)	100	71.4	57.1	71.4

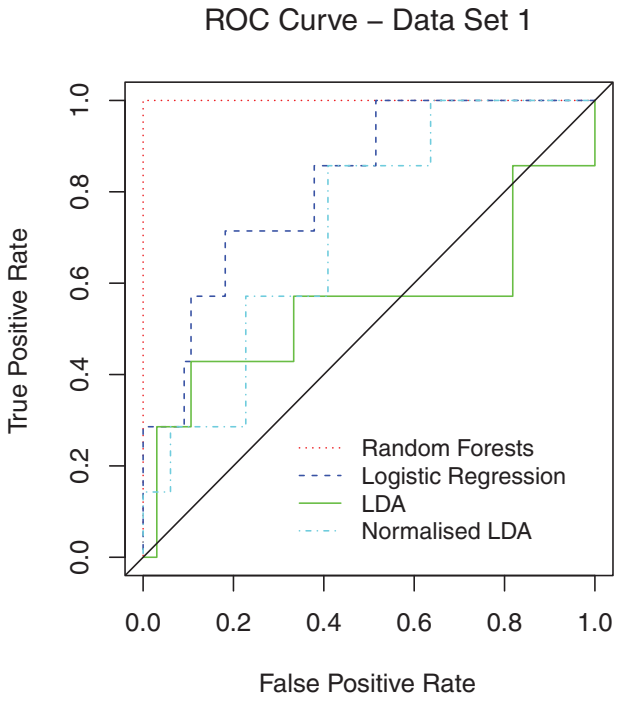

Figure 2. Receiver operating characteristic curves showing the results of the cross validation for Data Set 1. The diagonal line $y = x$ represents randomly classifying the samples, with half predicted as positive and half as negative. For definitions of classification results, see Section 2.4.

radius and the total integrated flux density provided the highest accuracy for random forests, logistic regression and LDA, while LDA using the 'normalised' data (transformed using a log function, see Section 2.1) was able to utilise the column density too. Table 2 shows the comparison of which of the predictor variables were included in the models based on their contributions to an increase in classification accuracy. Breen et al. (2007) showed that their sample of water masers preferred denser, more massive and more luminous sources. Our models indicated that the radius, luminosity and in the case of LDA on the normalised data, the density were important variables in predicting whether the sources were associated with a maser or not. Our results are in agreement with Breen et al. (2007), except that our models were not improved by inclusion of mass as a predictor variable.

Table 3 summarises the results we obtained from cross validation of the three different classification techniques under consideration (for details, see Section 2.4.2). Specificity values were high, due to the fact that the majority of the sources were not associated with masers, with the sensitivity values being lower in each case. For Data Set 1, random forests performed the best considering both sensitivity and specificity. Notably, there are very few false positive classifications over all the models, which is most likely due to the data being unbalanced in that the majority of the samples were not associated with masers. Another clear result is that performing LDA on the log transformed data increases the model's sensitivity, making it comparable to logistic regression in this case. The advantage of transforming the data is also obvious in the ROC shown below in Figure 2 (for an explanation on ROC curves, see Section 2.4.3).

Figure 2 shows that LDA under-performs for Data Set 1, however, when LDA is applied to the transformed data it is more accurate than logistic regression. The relatively small data set causes the apparent steps in the plot and this is also evident for Data Set 2 in Figure 5. The ROC curves for Data Set 3 (Figure 8) are much smoother because there are 214 samples rather than 73, or 32. Despite the apparent steps in the ROC curves, the plot very clearly shows the most accurate classification technique for this data set (the non-parametric method of random forests) and the least accurate (the parametric method of LDA using untransformed data).

Figure 3 shows a MDS plot for the full data set. MDS plots give a visual representation of the distances between proximities identified in the random forest implementation; sources that the random forest process identifies as being similar are clustered within the MDS plot. The distance values are arbitrary, they are simply relative magnitudes, plotted here as Dimension[1] and Dimension[2]. Figure 3 shows the four correctly identified maser sources in a group at the top-left, separated from the non-maser sources. 'Border-line' classifications were samples with a predicted maser association between 45 and 55%, with the last correctly classified maser shown just below the others as such. The model was not sensitive enough to detect the differences in the predictor

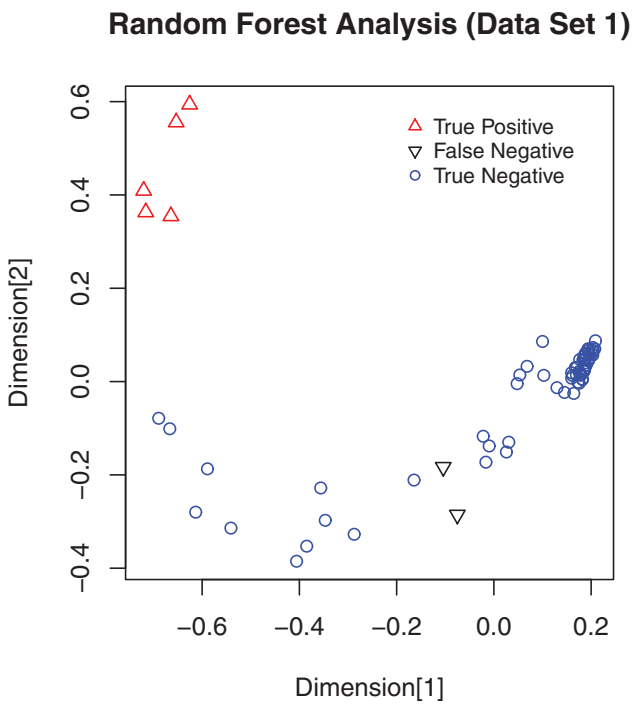

Figure 3. MDS plot of the proximity values produced by the random forest classification (Data Set 1). The values on the axes are arbitrary, the graph just compares relative magnitudes. The closer two points are on the plot, the more similar their properties as determined by the random forest classification. 'Border-line' classifications were samples with a predicted maser association between 45 and 55%.

variables for the other maser-associated sources (which is why they were classified as not having an associated water maser). This is probably due to the small number of sources in this data set.

3.2. The properties of water maser-associated YSOs in the LMC

Gruendl & Chu (2009) used the *Spitzer Space Telescope* Surveying Agents of Galaxy Evolution (SAGE) Legacy programme data (Meixner et al. 2006), along with other public data sets to identify high- and intermediate-mass young stellar objects (YSOs) in the Large Magellanic Cloud (LMC). Gruendl & Chu identified 855 definite YSOs in the LMC and compiled near- and mid-infrared photometric measurements for the sample. Ellingsen et al. (2010) made Australia Telescope Compact Array (ATCA) observations for the 22 GHz transition of water towards all known star-formation maser sites in the LMC, resulting in a total of 13 water masers in the LMC for which positions are known to arcsecond accuracy. The fields observed for the water maser observations included a total of 32 sources from the Gruendl & Chu (2009) YSO catalogue. Of the 13 water masers, 11 are within 2 arcsec of a Gruendl & Chu YSO, meaning that from a total catalogue of 855 sources there are 11 which are known to have an associated water maser and 22 which are known not to. The 33 sources for which there is information on whether or not they have an associated water maser can be used as a training set for classification/prediction.

Ellingsen et al. (2010) used the infrared data from Gruendl & Chu (2009) to construct the spectral energy distribution (SED) of each of the YSOs using the online SED-fitter of Robitaille et al. (2007) and this forms Data Set 2. For some wavelength ranges, the infrared data for the Gruendl & Chu (2009) sample is incomplete, hence there is missing data. However, the results of the SED modelling contain no missing data (although there is likely to be greater uncertainty in the fitted SED parameters for those sources which have less infrared photometric measurements contributing to the fitting process). All available information about a source is incorporated into the SED model. According to Ellingsen et al. (2010), there is very little variation in the amount of information available for each SED fit, with between seven and nine infrared intensities available for each source and in the majority of cases the chi-squared values for the resulting SED fits are reasonable. Due to the large number of sources modelled, we made no attempt to remove the sources where this was not the case, with the exception of one maser-associated source with more missing data than the others (making our training sample 32 with 10 known masers, and the total data set 854).

Fifteen predictor variables were extracted from the SED fitting results; distance to the source, age, radius, mass, and temperature of the central source, envelope accretion, or infall rate, outer and inner radius of the envelope, cavity opening angle, disc mass, ambient density, inclination of source to line of sight (LoS), average integrated flux density (from the outside of the YSO to the stellar surface, along the LoS), total luminosity, and mass of the envelope. Table 4 shows which variables were used in each model and how they contributed to that model. Across the different methods, the most important predictor variables appeared to be the mass of the central source, the outer envelope radius, the inclination towards the LoS, and the mass of the envelope. In comparison, Ellingsen et al. (2010) found that the majority of YSOs with an associated water maser have high luminosities, central masses, and ambient densities. They also tend to have redder infrared colours than those YSOs which are not associated with a maser. The distributions of the high-importance variables are shown in Figure 4.

Unlike Data Set 1, these data include some sources where the maser association is known (32) and some where it is unknown (822). This means predictions can be made on the unknown sources. To test how well the various methods will generalise to data where maser association is unknown, a cross-validation was applied. For a full description of the technique used, see Section 2.4.2. The predictions were then compared with the actual maser association. The results are given in Table 5.

The SED Data Set 2 had a fairly small number of entries with known maser status (32), but a large number of possible predictor variables (15). The results for the cross validation

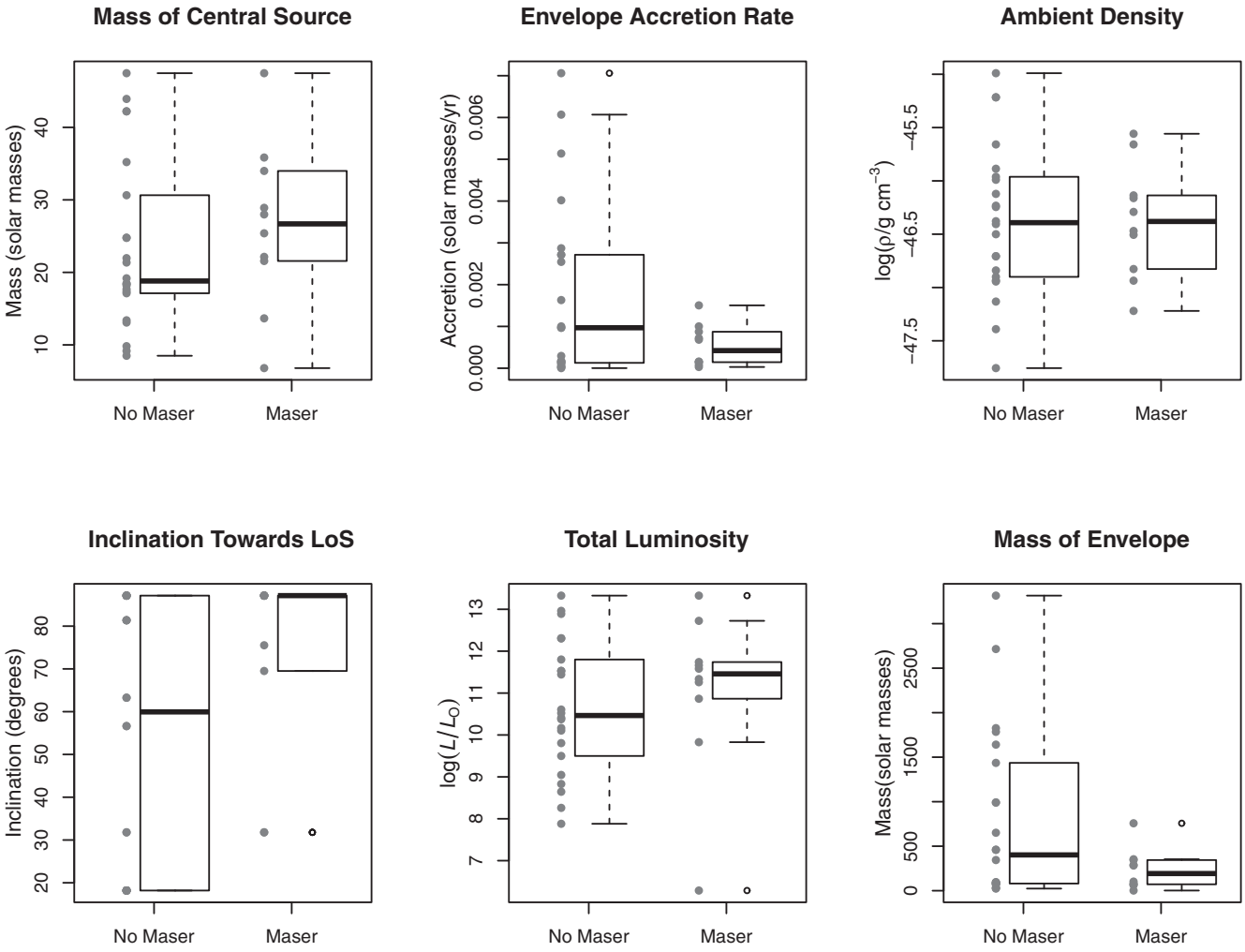

Figure 4. Only the predictor variables from Data Set 2 showing noticeable differences between those YSO with and without an associated water maser are shown. Due to the very small data set, the individual sample points are also plotted. Some of the variables are on logarithmic scales to better illustrate the differences. For an explanation of boxplots, see Figure 1.

are shown in Table 5 and in the form of ROC curves in Figure 5.

The cross validation results show that overall, the sensitivity values were quite poor, with LDA performed on the normalised data being the most accurate method. This was possibly due to the small data set, methods could not construct an accurate model using only 29 sources and then predicting on the remaining 3. These results can be visualised in a ROC curve shown in Figure 5, where in a number of cases the models fall below the $y = x$ diagonal line, meaning that they perform *worse* that simple random classification with a 50% chance of a source being associated with a maser. Due to these poor results, we decided not to use this training data to make predictions on the remaining 822 sources with unknown maser status.

It is evident from the MDS plot in Figure 6 why the random forest model had a low sensitivity; the data is not clustered in groups to the same extent as Data Set 1 (Figure 3). The known maser sources are represented by the black and red triangles, while the blue circles and green diamonds represent the known non-maser sources. This could be due to the model's inability to condense the 15 predictor variables (equivalent to 15 dimensions) into a two-dimensional plot, or another bi-product of the small sample size.

In summary, the variables that had the most influence over the various classification models were mass of the central star, the outer envelope radius, the inclination towards the LoS, and the mass of the envelope (see Table 4). The likelihood of a YSO being associated with a water maser source did not appear to depend heavily on variables such as the age of the source, mass of the disc, ambient density, or average integrated flux. Ellingsen et al. (2010) applied Mann–Whitney tests to the different variables to find the difference in the medians of the distributions of those associated with masers and those not associated. Statistically significant differences were found in the data for the mass of the central star, the outer radius of the envelope, ambient density, inclination towards the LoS, and the total luminosity; results that agree

Table 4. The predictor variables that increased the classification accuracy of the various methods for Data Set 2. The value given for random forests is the mean decrease in accuracy, while logistic regression provides p -values. The most important variables in logistic regression and random forest models are shown in bold. For further explanation, see Table 2.

	Random forests	Logistic reg.	LDA	Norm. LDA
Distance				Y
Age			Y	
Mass	**3.943**	**0.0805**	Y	
Radius	2.193	0.395	Y	Y
Temperature	2.091		Y	
Accretion	2.257		Y	Y
Outer env.		0.1609	Y	
Inner env.	1.099		Y	Y
Cavity angle	**3.891**		Y	Y
Disc mass			Y	Y
Amb. density		0.1935		
Inclination	2.677		Y	Y
Av. int. flux			Y	
Total lum.	1.047			
Env. mass	**4.202**	0.2765	Y	Y

Table 5. The results of cross-validating random forests, logistic regression, and LDA classification and prediction for Data Set 2 (association of water masers with infrared YSO in the LMC) using the full sample of 32 sources with known water maser association status as the training sample. For definitions of classification results, see Section 2.4.

	Random forests	Logistic reg.	LDA	Norm. LDA
True neg.	20	17	15	20
False pos.	2	5	7	2
False neg.	6	8	5	4
True pos.	4	2	5	6
Specificity (%)	90.9	77.3	68.2	90.9
Sensitivity (%)	40.0	20.0	50.0	60.0

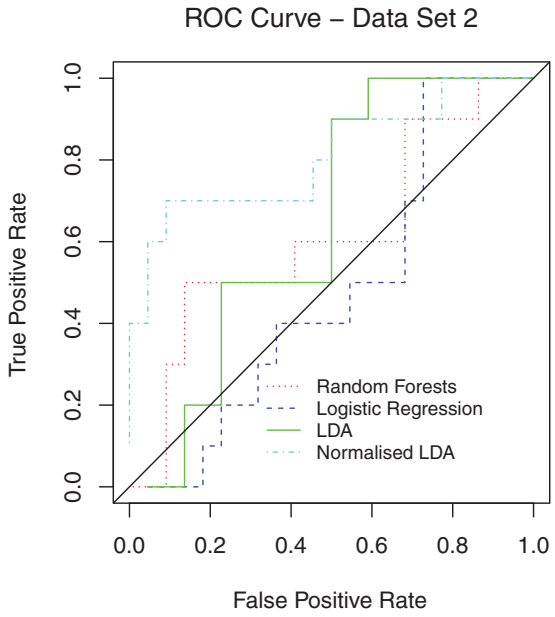

Figure 5. Receiver operating characteristic curves showing the results of the cross validation for Data Set 2. The diagonal line $y = x$ represents randomly classifying the samples, with half predicted as positive and half as negative. For a full description of a ROC curve, see Section 2.4.3.

Figure 6. The MDS plot for the random forest model used to predict potential YSOs with an associated water maser in the LMC (Data Set 2). 'Border-line' predictions were samples with a predicted maser association between 45 and 55%. For details on multidimensional plots in random forest analysis, see Figure 3.

with our analysis. It was previously suggested that the inclination angle is one of the most influential predictors in determining the SED for YSOs (Robitaille, et al. 2006). As a result of the orientation of the cavity, the inclination angle dictates the contribution from the inner, hotter regions of the envelope to the SED. Hence, our classification results here agree with those from previous studies, indicating that the physical variables mentioned above are likely to dictate water maser-association with certain YSOs in the LMC.

3.3. The properties of dust continuum emission associated with class I methanol masers

The final data set we investigated (hereafter, Data Set 3) was a search for 95 GHz class I methanol masers targeted towards regions selected on the basis of both their emission at mid-infrared and millimetre wavelength ranges (Chen et al. 2012). The mid-infrared data was taken from the *Spitzer Space Telescope* GLIMPSE (Galactic Legacy Infrared Mid-Plane Survey Extraordinaire) programme, which provides photometric measurements in four wavelength bands (3.6, 4.5, 5.8, and 8.0 μm; Benjamin et al. 2003; Churchwell

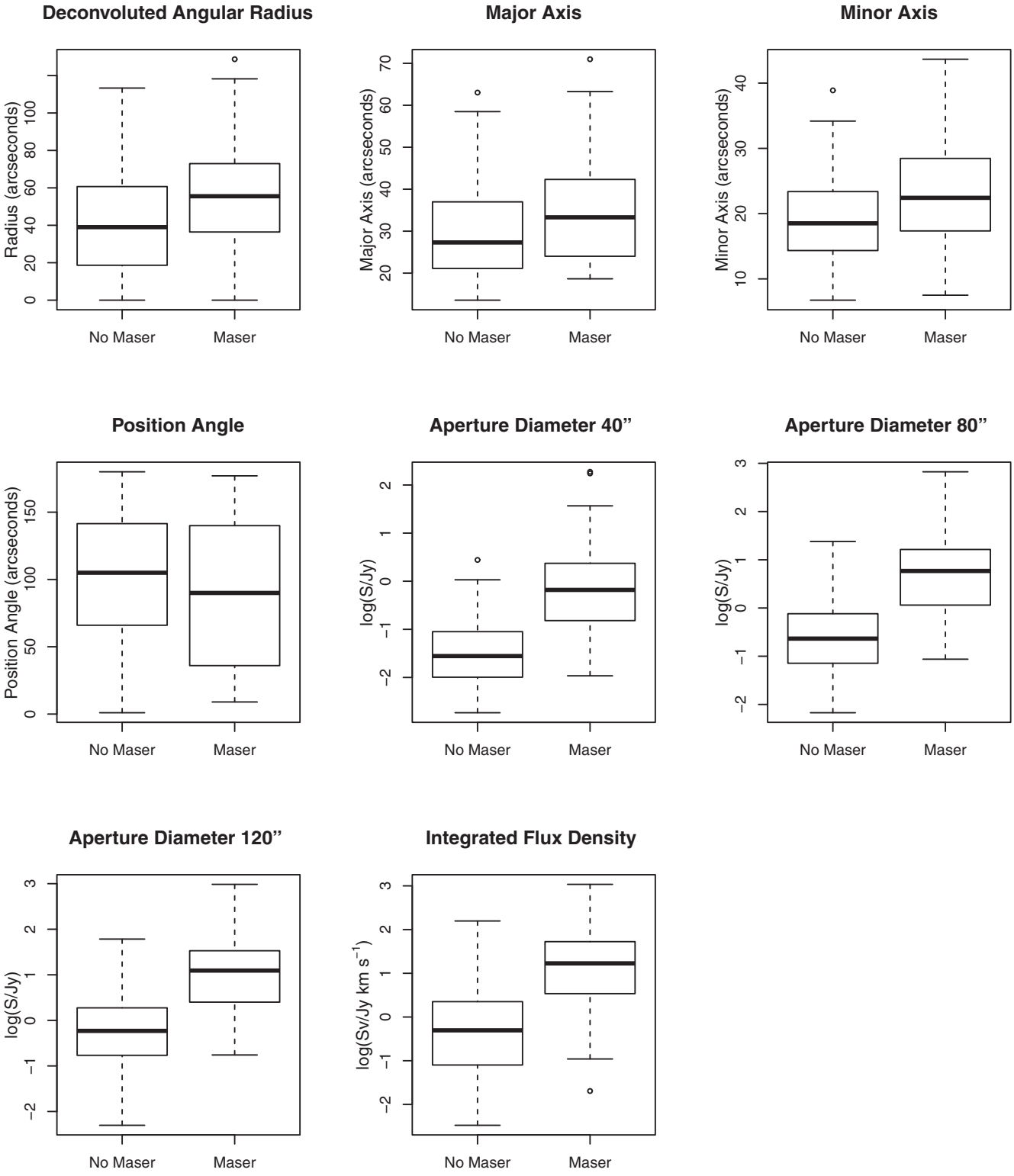

Figure 7. The variables used in the classification and prediction of Data Set 3. Some of the variables are on logarithmic scales to better illustrate the differences. For an explanation of boxplots, see Figure 1.

PASA, 33, e015 (2016)
doi:10.1017/pasa.2016.13

Table 6. The predictor variables that increased the classification accuracy of the various methods for Data Set 3. The value given for random forests is the mean decrease in accuracy, while logistic regression provides *p*-values. The most important variables in logistic regression and random forest models are shown in bold. For further explanation, see Table 2.

	Random forests	Logistic reg.	LDA	Norm. LDA
Major axis	4.124	0.2781	Y	Y
Minor axis	10.28	0.3867	Y	Y
Position angle	1.758	**0.0952**		Y
Angular radius	9.218	0.3252		
40 arcsec	**27.94**	**0.0312**	Y	Y
80 arcsec	**21.81**	0.4267	Y	
120 arcsec	14.24	0.6531	Y	
Int. flux den.	15.31	0.1158	Y	Y

Table 7. The results of cross-validating random forests, logistic regression, and LDA classification and prediction for Data Set 3 (class I methanol masers associated with *GLIMPSE* sources). Figure 8 shows the ROC curve for each of the models. For definitions of classification results, see Section 2.4.

	Random forests	Logistic reg.	LDA	Norm. LDA
True neg.	141	142	148	145
False pos.	11	10	4	7
False neg.	21	24	31	23
True pos.	41	38	31	39
Specificity (%)	92.8	93.4	97.4	95.4
Sensitivity (%)	66.1	61.3	50.0	62.9

et al. 2009), while the millimetre continuum data was from the Bolocam Galactic Place Survey (BGPS) (Aguirre et al. 2011). The motivation for this survey was a previous search for 95 GHz class I methanol masers by Chen et al. (2011). The authors targeted infrared sources for which GLIMPSE images show extended emission with an excess in the 4.5-μm band (thought to indicate an outflow from a high-mass YSO). It was found that those GLIMPSE sources with an associated BGPS source (54 of the 62 sources which lay within the BGPS region) were much more likely to exhibit class I methanol maser emission. Chen et al. (2011) also found that the GLIMPSE sources with redder mid-infrared colours were more likely to be associated with methanol masers and the higher the mass and density of the BGPS dust clump, the stronger the class I maser emission.

Chen et al. (2012) used the results of Chen et al. (2011) to identify 420 sources detected in both the *Spitzer* GLIMPSE and BGPS catalogues as likely to have an associated class I methanol maser. They then observed a random selection of 214 of these sources and detected 95 GHz class I methanol masers towards 62 (hence 152 non-detections). For the classification process, we used only the data from the BGPS catalogue (version 1.0) which contains a total of 8 358 sources (Aguirre et al. 2011). The predictor variables used in the classification models for Data Set 3 were the angular size of the major and minor axis of the dust clump, as well as its position angle, deconvolved angular radius, and 1.1-mm flux density within apertures of diameter 40, 80, and 120 arcsec and the integrated flux density. Boxplots of these variables are shown in Figure 7. As with the classification of the other two data sets, here each of the models were optimised by omitting superfluous variables, as well as those that decreased the models' accuracy. Both logistic regression and random forests utilised all eight variables, while LDA performed better without including all of them (see Table 6).

This data set was the primary focus of our analysis, as it has a training set with several hundred sources, including a large number of detections and there are also a large number of BGPS sources which have not been searched for class I methanol maser emission (8 144) which provide the opportunity to make testable predictions.

The variables with high importance in the random forests calculations were the flux densities (each of the 40, 80, and 120 arcsec aperture values and the integrated) and also the angular size of the minor axis. The most important variable was the flux density within 40 arcsec (the smallest angular scale measured by Bolocam). Logistic regression also found the flux density within 40 arcsec to be the most important variable with a *p*-value of 0.0312, with the next most significant variable being the position angle with a *p*-value of 0.0952, while the 80 arcsec flux had the next highest contribution. This is consistent with the results of Chen et al. (2012) which showed that class I masers were preferentially associated with sources with the highest beam averaged column density (which is directly proportional to the 40 arcsec flux density). There is no physical reason why the position angle of the dust clump would effect the likelihood of a dust clump having an associated class I methanol maser, but when this variable was omitted from the classification, the accuracy of the models decreased. However, while the *p*-value suggests that the position angle is a significant predictor variable in logistic regression, the change in accuracy was not significant compared to that of the other variables. This suggests that we can dismiss it as an artefact of the classification method, but it does serve as a reminder to view results such as this with a degree of scepticism. It is also worth noting that random forests presented it with the lowest variable importance.

Table 7 shows the results of the cross-validation of the different classification techniques used on Data Set 3 (see Section 2.4.2). Here, random forests offered the highest sensitivity, while surprisingly performing LDA on the untransformed data produced the highest specificity. This is the first instance in our studies where transforming the data set to be closer to a normal distribution decreased the performance of LDA, although the decrease was minor (2%) and likely not significant. The ROC curve in Figure 8 gives a more complete representation of the models' capabilities, showing that random forests, logistic regression, and LDA using the

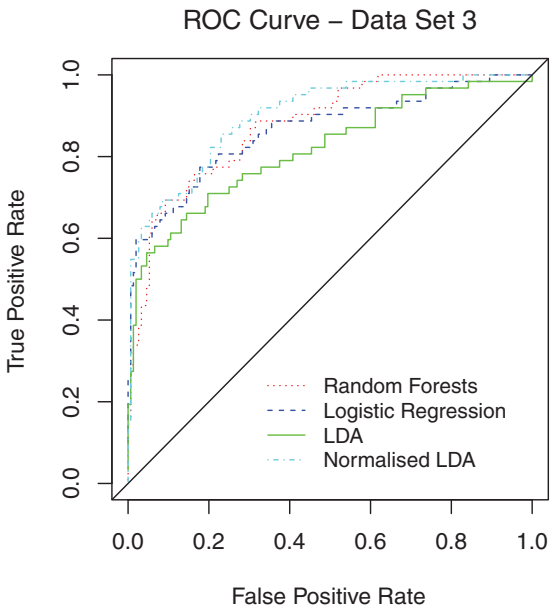

Figure 8. Receiver operating characteristic curves showing the results of the cross validation for Data Set 3. The diagonal line $y = x$ represents randomly classifying the samples, with half predicted as positive and half as negative.

transformed data performed to similar standards, while generally LDA on the untransformed data performed the worst.

Figure 9 shows the MDS plot for the random forest model generated using all the training data for Data Set 3. It is clear that the maser-associated dust clumps are generally located in the bottom-right region of the plot. Comparing this plot to the MDS plots for the other two Data Sets (Figures 3 and 6), we can see that the maser-associated sources are more clearly separated from those without a maser-association. The green squares in Figure 9 represent sources which the random forests model predicts to have an associated class I methanol maser, but for which the observations of Chen et al. (2012) did not detect a maser. Many of these sources lie very close on the MDS plot to others where a maser was detected and it may be that some of these non-detections have a weak class I methanol maser which was not detected by Chen et al. due to the limited sensitivity of those observations. The 95 GHz class I methanol masers are in the same transition family as the best studied class I methanol maser transition at 44 GHz. In general, the 44 GHz class I methanol masers have a peak flux density approximately a factor of 3 greater than the 95 GHz maser emission in the same source (Val'tts et al. 2000). These sources would be good candidates for sensitive observations in the 44 GHz transition to more robustly determine if they are associated with class I methanol masers.

The classification models we have developed can also be used to predict which of the BGPS sources that were not observed by Chen et al. (2012) are the best candidates for having an associated class I methanol maser. Since we have four different classification models, we can compare the results of each, as those sources identified by all, or most of the models would be expected to be the promising targets for further searches.

Table 8. The classification results on the training data subset (where the maser presence is known), and the number of predicted masers from the 8 144 sources for which maser presence is unknown, using Data Set 3 (class I methanol masers associated with *GLIMPSE* sources). For definitions of classification results, see Section 2.4.

	Random forests	Logistic reg.	LDA	Norm. LDA
True neg.	140	145	149	147
False pos.	12	7	3	8
False neg.	21	22	30	22
True pos.	41	40	32	40
Specificity (%)	92.1	95.4	98.0	96.7
Sensitivity (%)	66.1	64.5	51.6	64.5
Predictions	632	405	334	460

Table 9. Number of maser predictions on sources from Data Set 3 shared by two classification methods, with 242 sources predicted to be masers using all four methods.

	Random forests	Logistic reg.	LDA	Norm. LDA
Random forests	632	364	317	377
Logistic reg.		405	254	371
LDA			334	256
Norm. LDA				460

For the prediction model, as with Data Set 2, we grew a random forest using 3 000 trees (instead of the default 500, see Section 3.2). Table A1 in the Appendix lists the 739 BGPS sources which were predicted to have an associated class I methanol maser by one or more of the four classification models for the 8 144 BGPS sources which have not yet been searched. Table 8 shows that of the 8 144 potential BGPS target sources random forests predicts 632 to have an associated class I methanol maser and this is significantly more than any of the other classification models. There are 242 of the 8 144 BGPS sources which all models predict will have an associated class I methanol maser and these will be the prime targets for future searches. Table 9 shows the number of sources predicted to be masers by one or two classification methods, which should be considered if there is sufficient time to search additional targets.

4 DISCUSSION

We applied three different classification techniques to three different searches for interstellar masers to investigate each technique's performance. We show the classification and prediction results of LDA performed on both the normally

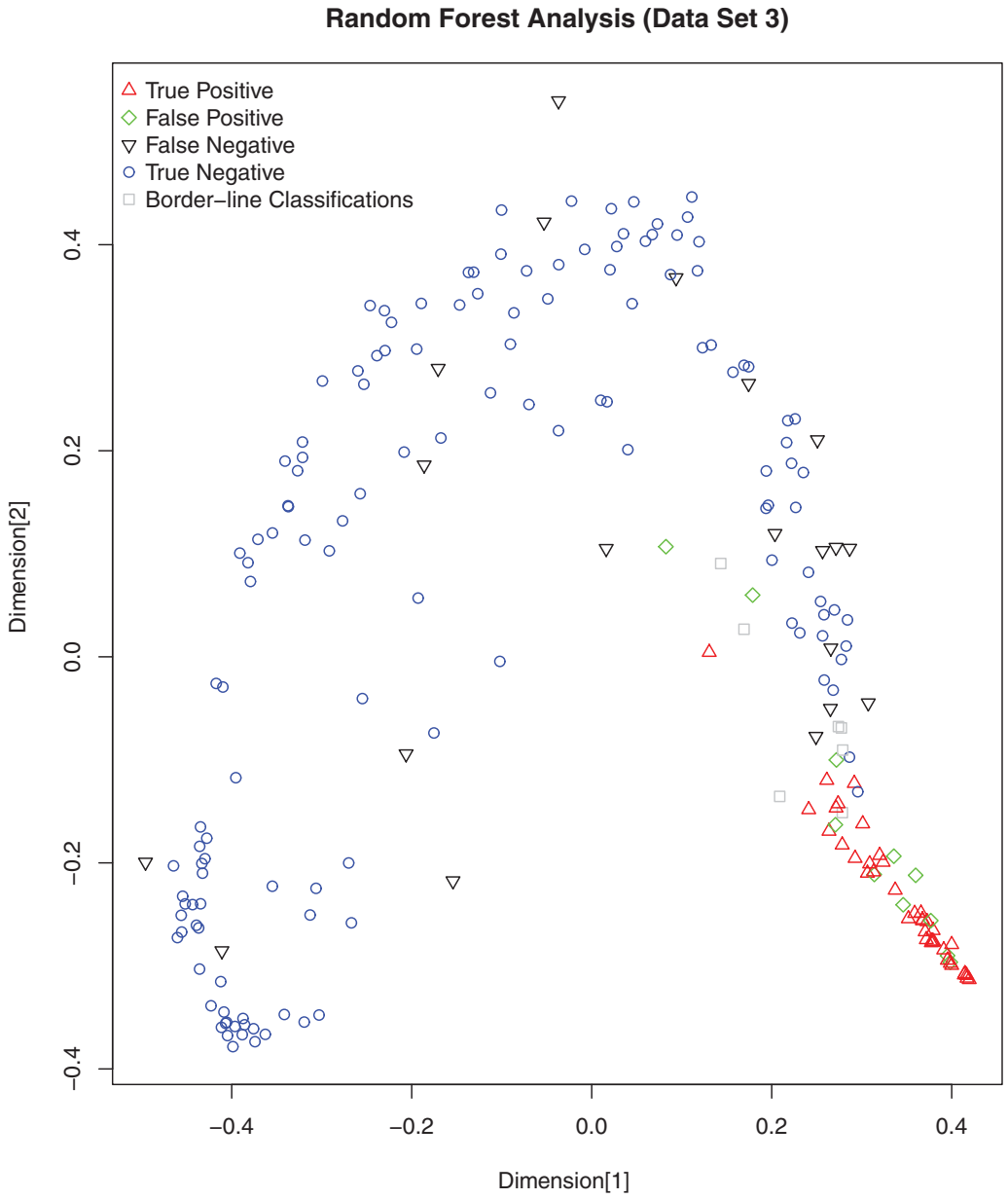

Figure 9. The MDS plot for the random forest model used to predict potential millimetre dust-clumps with associated class I methanol masers (Data Set 3). 'Border-line' classifications were samples with a predicted maser association between 45 and 55%. For details on multidimensional plots in random forest analysis, see Figure 3.

distributed data and the untransformed data to demonstrate the difference (see Section 2.1). In most cases, LDA performs significantly better when applied to transformed data.

All three methods of classification (both parametric and non-parametric) used on Data Set 1 returned high values for both sensitivity (correctly classifying sources associated with masers) and specificity (correctly classifying non-maser sources). The highest accuracy was achieved through the non-parametric method of random forests, which in this case classified every source correctly.

Data Set 2 had a relatively small training sample (32 sources) compared with the number of predictor variables (15), and we found here that LDA appeared to give the best results, while random forests and logistic regression performed quite poorly in correctly identifying sources associated with a maser.

For Data Set 3, which has more than 50 detections and more than 150 non-detections, the non-parametric random forests had the highest sensitivity, while the parametric method of LDA performed on the untransformed data had the lowest, but also had the highest specificity. Considering both sensitivity and specificity, logistic regression, and random forests were the most accurate methods.

Based on the predictions of Breiman (2001b), our initial expectation was that given sufficient training data the complex relationship between the predictor variables and the presence or absence of a related astrophysical phenomenon would be more accurately represented by a non-parametric approach than a simple linear model. However, the training data sets were relatively small, and so made it difficult for the models to capture and convey all the information contained within the predictor variables. We find that random forests does perform relatively better for the largest data set, but in this case, it is comparable with the accuracy of the non-parametric techniques, not superior to them. It may be that in order to outperform parametric methods, the non-parametric techniques require still larger amounts of training data. However, it is more likely that for Data Set 3 all techniques approximately reach the limit of the information available within the measured parameters of the data.

There are a number of factors related specifically to the data which will lead to limitations in the accuracy of any classification model developed using it. One factor is the intrinsic measurement uncertainty for parameters such as the flux density, angular size, etc., which can influence the results directly in the sense that it is always possible that given observations with greater sensitivity additional sources would be detected. However, the absence of these weaker sources does more than simply qualifying the question that is being answered by the classification model. For example, the intensity of astrophysical masers depends in a complex and non-linear manner on the physical parameters of the environment and some of these parameters may not be represented either directly or indirectly in any of the predictor variables being used as inputs to the classification methods. A second, less obvious factor which may limit the accuracy of classification techniques is that for derived parameters there are often implicit assumptions. For example, the calculation of the mass of the dust clumps used for Data Set 1 assumes that the emission at 1.2-mm wavelength is optically thin (likely a reasonable assumption), and that the dust is at a constant temperature of 40 K for all the dust. This second assumption is necessary because we do not have any information on the specific temperature distribution of the dust, but it inevitably leads to systematic errors in the relative mass calculated for regions where the true dust temperature is on average higher (or lower) than the assumed value. Similarly, the distance to individual sources has been estimated using kinematic distance models, which on average provide a reasonable estimate, but which can lead to significant errors for individual sources. It is also highly probable that our sensitivity values obtained after cross validation of the classification techniques were poor due to the unbalanced nature of the data, in that for all three data sets, the vast majority of the samples are not associated with masers.

Breen & Ellingsen (2011) tested the binomial generalised linear model of (Breen et al. 2007, Data Set 1) by searching for 22 GHz water masers towards 267 dust clumps. They found a high-detection rate towards dust clumps for which the binomial GLM predicted a probability of greater than 10% for the presence of a water maser (20 of 27 sources). They also found that while the detection rate dropped for sources for which the model predicted a lower probability of having an associated water maser, a substantial fraction of water masers (approximately 70%) were detected towards sources for which the model predicted a probability of less than 1%. Breen & Ellingsen (2011) show that unreliable distance estimates for many of the dust clumps is in part responsible for the misclassification. This is consistent with the assertion we make above that the combination of measurement and systematic uncertainties in the underlying data ultimately limit the accuracy which can be obtained with any classification technique.

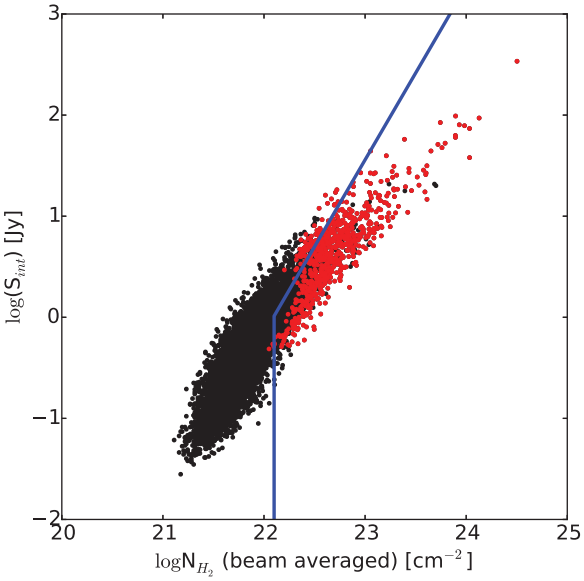

Figure 10. The integrated flux density versus the beam averaged H_2 column density for the 8 144 BGPS sources not searched for class I methanol masers by Chen et al. (2012). Sources for which one or more of the classification models predicts the presence of a class I methanol masers are represented with red dots, other sources are represented with black dots. The blue line shows the criteria developed by Chen et al. (2012) to identify BGPS sources likely to have an associated class I methanol maser.

When the different classification models we developed using Data Set 3 are applied to the 8 144 BGPS sources which have not been searched for class I methanol maser emission a total of 739 sources are predicted to have an associated maser by one or more of the models, with 242 sources predicted by all four models (see Section 3.3 and Appendix A). Figure 12 of Chen et al. (2012) plots the integrated flux density against the beam averaged H_2 column density for Data Set 3 and shows that the maser associated sources are restricted to a limited range for these two predictor variables. Figure 10 shows the integrated flux density versus the beam averaged H_2 column density for the 8 144 BGPS sources not observed by Chen et al. Those sources for which one or more of the classification models predict an associated class I methanol maser are indicated with a red dot, with sources which no

model predicts to have an associated class I maser are indicated with a black dot. Figure 10 shows that there is a high level of agreement between the predictions of the classification models and the empirical criteria developed by Chen et al. (2012). In total 1 200 BGPS sources meet the criteria identified by Chen et al. (2012), approximately a factor of two more than identified by any of the classification models. In their calculation of the beam averaged column density, Chen et al. (2012) assumed a constant temperature of 20 K for the dust clumps and used the 40 arcsec flux density measurement as the intensity of the dust continuum emission. This means that the calculated beam average column density is directly proportional to the BGPS 40 arcsec flux density measurement. The relationship derived by Chen et al. (2012) suggested this is the most important predictor variable for the presence (or otherwise) of class I methanol maser emission towards these sources. The results presented here also support this.

Ultimately, determining the relative accuracy of these classification models, and whether they are superior to directly derived criteria [such as those of Chen et al. (2012)] is to test them through future observations. There are currently approximately 400 different class I methanol maser sources which have been identified throughout the Galaxy (see Chen et al. 2013, and references therein), so a search targeted towards the candidate BGPS sources we have identified is likely to significantly increase the number of known sources.

5 CONCLUSIONS

In this paper, we present three major findings regarding the utilisation of different classification techniques on different size astronomical data sets. (1) For small data sets parametric methods (such as LDA and logistic regression perform better than random forests (a non-parametric method). (2) For larger data sets, random forests has the capability to out-perform the parametric methods trialled here. (3) In almost all cases, transforming the data to be closer to a normal distribution significantly increases the accuracy of LDA. In the case where using transformed data slightly decreased the accuracy of the model, the classification results were very similar. Since the process of transforming data is relatively easy, this is a step that should be definitely employed if LDA is utilised. This step has typically not been included when LDA has been applied to astronomical data used in past studies.

Our results suggest that where there is very limited training information parametric models which can only predict based on simple combinations of the input variables are more accurate than non-parametric methods. However, where there is more training data (such as Data Sets 1 and 3), non-parametric models can perform as well (likely better in some circumstances) than parametric techniques. Our results for Data Set 3 show that random forests is comparable in accuracy to the parametric methods, rather than exceeding them as expected (see Breiman 2001b).

Frequently in astrophysics relationships are sought between two or three variables in the form of correlations between them, such as the radio:far-infrared correlation for galaxies, or colour–colour selection criteria for HII regions. In the past, this has often been because of limited numbers of predictor variables being available for large samples of data, however, this is now less of an issue. Mathematical classification techniques such as those utilised here potentially offer significant improvements over simple correlation relationships, but the most appropriate technique to apply depends heavily on the nature of the data available and the goal of the investigation (e.g. detection prediction, physical understanding of relationship between variables). Our models determined which predictor variables were important in the classification process, and for all three Data Sets our results agreed with the previous studies of Breen et al. (2007), Ellingsen et al. (2010), and Chen et al. (2012), respectively.

For the specific goal of identifying millimetre dust clumps which are more likely to have an associated class I methanol maser, we find that on the basis of cross-validation tests and the predictions the models produce on the training data, both the non-parametric method of random forests, and the parametric methods of logistic regression and LDA are well suited for the task of identifying likely targets for future searches. 242 sources out of the 8 144 in Data Set 3, were predicted by all four of our techniques to have associated masers. The results of future searches for class I methanol masers towards BGPS sources will allow a direct test of each of the classification models and allow us to determine the validity of these conclusions.

ACKNOWLEDGEMENTS

EMM undertook much of this work funded through a University of Tasmania Dean's Summer Research Scholarship. Shari Breen is the recipient of an Australian Research Council DECRA Fellowship (project number DE130101270). This research has made use of NASA's Astrophysics Data System Abstract Service.

References

Aguirre, J. E., et al. 2011, ApJS, 192, 4
Bailey, S., Aragon, C., Romano, R., Thomas, R. C., Weaver, B. A., & Wong, D. 2007, ApJ, 665, 1246
Benjamin, R. A., et al. 2003, PASP, 115, 953
Borra, S., & Di Ciaccio, A. 2010, Computational Statistics and Data Analysis, 54, 2976
Breen, S. L., & Ellingsen, S. P. 2011, MNRAS, 416, 178
Breen, S. L., et al. 2007, MNRAS, 377, 491
Breen, S. L., Ellingsen, S. P., Caswell, J. L., & Lewis, B. E. 2010, MNRAS, 401, 2219
Breiman, L. 2001a, Technical report, Random Forests, University of California Berkeley, http://oz.berkeley.edu/~breiman/randomforest2001.pdf
Breiman, L. 2001b, StaSc, 16, 199
Breiman, L. 2001c, Machine Learning, 45, 5

Breiman, L., & Cutler, A. 2013, Technical report, Random Forests, University of California Berkeley, http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm
Carliles, S., Budavári, T., Heinis, S., Priebe, C., & Szalay, A. S. 2010, ApJ, 712, 511
Chen, X., Ellingsen, S. P., Shen, Z.-Q., Titmarsh, A., & Gan, C.-G. 2011, ApJS, 196, 9
Chen, X., Gan, C.-G., Ellingsen, S. P., He, J.-H., Shen, Z.-Q., & Titmarsh, A. 2013, ApJS, 206, 9
Chen, X., et al. 2012, ApJS, 200, 5
Churchwell, E., et al. 2009, PASP, 121, 213
Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., J. G., & Lawler, J. J. 2007, Ecology, 88, 2783
Feigelson, E. D., & Babu, J. 2012, Modern Statistical Methods for Astronomy: With R Applications (Cambridge: Cambridge Univ. Press)
Efron, B., & Tibshirani, R. 1994, An Introduction to the Bootstrap Sample (UK: Chapman and Hall/CRC)
Einasto, M., Liivamägi, L. J., Saar, E., Einasto, J., Tempel, E., Tago, E., & Martínez, V. J. 2011, A&A, 535, A36
Ellingsen, S. P. 2005, MNRAS, 359, 1498
Ellingsen, S. P. 2006, ApJ, 638, 241
Ellingsen, S. P., Breen, S. L., Caswell, J. L., Quinn, L. J., & Fuller G. A. 2010, MNRAS, 404, 779
Ellingsen, S. P., Voronkov, M. A., Cragg, D. M., Sobolev, A. M., Breen, S. L., & Godfrey, P. D. 2007, in IAU Symp., Vol. 242, IAU Symposium, ed. J. M. Chapman, & W. A. Baan (Cambridge: Cambridge University Press), 213 (arXiv 0705.2906), doi:10.1017/S1743921307012999
Fisher, R. A. 1922, Phil. Trans. Royal Soc., 222, 309
Goddi, C., Moscadelli, L., & Sanna, A. 2011, A&A, 535, L8
Gruendl, R. A., & Chu, Y.-H. 2009, ApJS, 184, 172
Hastie, T., Tibshirani, R., & Friedman, J. 2001, The Elements of Statistical Learning: Data Mining, Inference and Prediction (Berlin: Springer)
Hosmer, D. W., & Lemeshow, S. 2000, Applied Logistic Regression (2nd edn.; NJ: John Wiley & Sons)
Johnston, S., et al. 2007, PASA, 24, 174
Kobel, P., Hirzberger, J., Solanki, S. K., Gandorfer, A., & Zakharov, V. 2009, A&A, 502, 303
Liaw, A., & Wiener, M. 2002, R News, 2, 18
Lo, N., et al. 2009, MNRAS, 395, 1021
Meixner, M., et al. 2006, AJ, 132, 2268
Mirabal, N., Frías-Martinez, V., Hassan, T., & Frías-Martinez, E. 2012, MNRAS, 424, L64
Mookerjea, B., Kramer, C., Nielbock, M., & Nyman, L.-Å. 2004, A&A, 426, 119
Morgan, A. N., Long, J., Richards, J. W., Broderick, T., Butler, N. R., & Bloom, J. S. 2012, ApJ, 746, 170
R Core Team 2013, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org
Robitaille, T. P., Whitney, B. A., Indebetouw, R., & Wood, K. 2007, ApJS, 169, 328
Robitaille, T. P., Whitney, B. A., Indebetouw, R., Wood, K., & Denzmore, P. 2006, ApJS, 167, 256
Song, H., Tan, C., Jing, J., Wang, H., Yurchyshyn, V., & Abramenko, V. 2009, Sol. Phys., 254, 101
Surcis, G., Vlemmings, W. H. T., van Langevelde, H. J., & Hutawarakorn Kramer, B. 2012, A&A, 541, A47
Titmarsh, A. M., Ellingsen, S. P., Breen, S. L., Caswell, J. L., & Voronkov, M. A. 2013, ApJ, 775, L12
Val'tts, I. E., Ellingsen, S. P., Slysh, V. I., Kalenskii, S. V., Otrupcek, R., & Larionov, G. M. 2000, MNRAS, 317, 315
Yuan, Y., Shih, F. Y., Jing, J., & Wang, H.-M. 2010, RAA, 10, 785

APPENDIX

A CLASSIFICATION MODEL PREDICTIONS

Table A1 summarises the predictions for each of the classification models for class I methanol masers associated with Bolocam sources.

Table A1. Bolocam Galactic Plane sources for which one or more of the mathematical classification models predicted the presence of an associated class I methanol maser (probability of a maser >0.5). The maser probability for each model is listed, those which exceed 0.5 are in bold type. This list contains a total of 739 sources that were predicted to be masers by at least one of the four methods (242 of which were predicted by all methods), from a total of 8 144 sources in version 1.0.1 of the Bolocam catalogue.

Bolocam catalogue #	BGPS name	Random forests	Logistic regression	LDA	Normalised LDA
4	G000.010+00.157	0.43	0.44	0.40	**0.60**
5	G000.016−00.017	**0.84**	0.49	**0.64**	0.42
7	G000.020+00.033	**0.81**	0.29	0.47	0.17
8	G000.020−00.051	**0.92**	**0.71**	**0.75**	0.45
18	G000.052+00.027	**0.78**	**0.72**	**0.91**	**0.64**
20	G000.054−00.209	**0.82**	0.47	**0.77**	0.31
22	G000.066−00.079	**0.86**	**0.87**	**1.00**	**0.76**
24	G000.070+00.175	**0.55**	0.06	0.27	0.02
27	G000.072+00.047	**0.60**	0.27	0.37	0.15
32	G000.094−00.109	**0.55**	0.12	0.22	0.09
35	G000.098+00.073	**0.55**	0.12	0.06	0.11
38	G000.104−00.005	**0.90**	**0.73**	**0.80**	**0.80**
39	G000.106−00.085	**0.94**	**0.96**	**0.99**	**0.76**
41	G000.110+00.001	**0.84**	**0.77**	**0.77**	**0.76**
43	G000.118+00.085	**0.51**	0.13	0.31	0.08
47	G000.120−00.513	0.21	0.11	0.05	**0.54**
48	G000.122−00.113	**0.58**	0.37	0.19	0.23
56	G000.140+00.021	**0.87**	0.20	**0.52**	0.27
57	G000.140−00.085	**0.83**	0.43	**0.54**	0.27
61	G000.156−00.091	**0.85**	0.29	**0.71**	0.44
62	G000.162−00.039	**0.56**	0.23	0.38	0.16
72	G000.184−00.003	**0.62**	0.23	0.14	0.09
79	G000.208−00.003	**0.71**	**0.71**	0.05	0.48
81	G000.212−00.517	**0.93**	**0.93**	**0.97**	**0.87**
83	G000.216−00.019	**0.79**	0.13	0.34	0.10
84	G000.216−00.045	**0.75**	0.14	0.31	0.12
87	G000.228−00.475	**0.61**	0.15	**0.69**	0.23
89	G000.234−00.089	**0.54**	0.11	0.31	0.03
91	G000.246−00.043	**0.72**	0.30	0.18	0.27
96	G000.254+00.013	**0.96**	**0.99**	**1.00**	**0.88**
99	G000.262+00.027	**0.98**	**1.00**	**1.00**	**0.98**
102	G000.274−00.085	**0.60**	0.09	0.13	0.16
103	G000.278−00.063	**0.58**	0.10	**0.51**	0.17
106	G000.282−00.481	**0.64**	**0.62**	**0.80**	**0.71**
109	G000.292−00.025	**0.58**	0.12	0.25	0.03
112	G000.296+00.043	**0.86**	**0.78**	0.50	**0.75**
115	G000.318−00.101	**0.72**	0.18	0.42	0.13
116	G000.320−00.201	**0.86**	**1.00**	**0.99**	**0.99**
123	G000.332−00.011	**0.61**	0.08	**0.60**	0.09
124	G000.332−00.075	**0.76**	0.22	**0.56**	0.22
126	G000.338+00.097	**0.52**	0.16	0.24	0.05
127	G000.340+00.053	**0.71**	**0.74**	**0.85**	**0.73**
130	G000.368−00.083	**0.59**	0.28	0.26	0.09
135	G000.378+00.041	**0.96**	**1.00**	**0.56**	**0.97**
141	G000.394−00.083	**0.59**	0.06	0.29	0.06
148	G000.412−00.503	0.44	**0.77**	0.34	**0.79**
149	G000.414+00.051	**0.86**	0.39	**0.97**	**0.71**
168	G000.472+00.019	**0.94**	**0.77**	0.03	**0.57**
170	G000.482−00.005	**0.95**	**1.00**	**1.00**	**0.97**
171	G000.492−00.111	**0.58**	0.18	0.37	0.27
173	G000.498+00.017	**0.78**	**0.89**	**0.82**	**0.87**

Table A1. Continued.

Bolocam catalogue #	BGPS name	Random forests	Logistic regression	LDA	Normalised LDA
175	G000.500+00.187	0.29	**0.58**	0.09	**0.60**
186	G000.530+00.181	**0.89**	**0.98**	**0.75**	**0.90**
190	G000.546−00.003	**0.92**	**0.70**	0.46	0.41
193	G000.558−00.067	**0.53**	0.16	0.38	0.14
196	G000.572+00.023	**0.68**	0.07	0.38	0.10
199	G000.586−00.125	**0.56**	0.07	0.14	0.13
203	G000.590+00.007	**0.66**	0.40	0.13	0.21
206	G000.598−00.113	**0.59**	0.17	0.07	0.49
207	G000.606−00.033	**0.86**	0.01	0.38	0.16
209	G000.608+00.001	**0.83**	0.40	0.08	0.13
211	G000.610−00.057	**0.95**	**0.99**	**0.67**	**0.96**
218	G000.630−00.095	**0.85**	0.24	0.18	**0.73**
222	G000.648+00.027	**0.90**	0.00	0.01	0.42
223	G000.656−00.045	**0.77**	**1.00**	**1.00**	**1.00**
225	G000.670−00.141	**0.88**	0.01	0.21	0.30
226	G000.674−00.097	**0.91**	0.09	0.36	**0.66**
227	G000.680−00.029	**0.94**	**1.00**	**1.00**	**1.00**
228	G000.684−00.169	**0.65**	0.03	0.08	0.14
229	G000.686−00.111	**0.84**	**0.90**	0.09	**0.60**
237	G000.738−00.051	**0.90**	**0.86**	0.17	0.48
238	G000.738−00.093	**0.84**	0.10	**0.66**	0.43
239	G000.738−00.157	**0.59**	0.18	**0.60**	0.12
241	G000.748+00.017	**0.76**	0.17	0.15	0.08
244	G000.760−00.069	**0.78**	0.00	0.07	0.16
245	G000.762+00.013	**0.68**	0.04	0.10	0.01
249	G000.772−00.109	**0.67**	0.22	0.08	0.18
250	G000.772−00.251	**0.75**	**0.54**	**0.80**	**0.69**
251	G000.776−00.187	**0.70**	0.31	0.44	0.24
258	G000.798−00.156	**0.51**	0.38	0.40	0.22
260	G000.802−00.098	**0.82**	0.25	**0.60**	0.13
261	G000.812+00.024	**0.59**	0.05	0.35	0.03
263	G000.826−00.212	**0.89**	**0.58**	**0.55**	**0.51**
266	G000.834−00.152	**0.84**	0.06	**0.72**	0.19
268	G000.836−00.200	**0.74**	0.05	0.16	0.09
269	G000.840+00.184	**0.68**	**0.86**	0.26	**0.82**
277	G000.862−00.054	**0.70**	0.19	0.46	**0.62**
280	G000.868−00.040	**0.72**	0.00	0.02	0.02
285	G000.886−00.036	**0.77**	0.44	0.09	**0.78**
296	G000.906−00.022	**0.52**	0.26	0.10	0.33
317	G000.950−00.080	**0.59**	0.00	0.28	0.01
346	G001.010−00.240	**0.82**	**0.94**	**0.77**	**0.92**
349	G001.020−00.122	**0.50**	0.05	0.35	0.02
350	G001.024+00.068	**0.53**	0.05	0.23	0.01
367	G001.092−00.030	**0.65**	0.01	0.34	0.06
377	G001.128−00.108	**0.66**	**1.00**	**0.98**	**0.98**
390	G001.150−00.126	0.40	0.30	0.05	**0.60**
408	G001.194−00.074	**0.55**	0.17	0.33	0.12
427	G001.234+00.056	**0.56**	0.23	0.49	0.20
471	G001.320−00.142	**0.55**	0.12	0.28	0.08
481	G001.338+00.096	**0.52**	0.17	0.41	0.14
489	G001.354+00.260	**0.53**	0.16	**0.64**	0.08
513	G001.406+00.328	0.14	0.12	0.05	**0.77**
536	G001.476+00.040	**0.58**	0.12	0.34	0.04
548	G001.518−00.194	0.27	0.15	**0.53**	0.15
572	G001.600+00.022	**0.72**	0.09	0.43	0.37
578	G001.610−00.172	**0.63**	0.14	**0.58**	0.09

Table A1. Continued.

Bolocam catalogue #	BGPS name	Random forests	Logistic regression	LDA	Normalised LDA
585	G001.652−00.066	**0.67**	0.07	**0.79**	0.17
596	G001.676−00.130	**0.56**	0.12	0.27	0.07
603	G001.696−00.386	**0.57**	0.29	0.30	0.25
604	G001.698−00.366	**0.68**	0.27	**0.52**	0.24
612	G001.734−00.412	**0.73**	0.21	**0.66**	0.39
663	G002.144+00.006	**0.54**	0.50	0.15	0.42
700	G002.444+00.126	0.29	0.48	0.37	**0.55**
723	G002.534+00.198	**0.56**	**0.60**	**0.55**	**0.60**
735	G002.616+00.132	**0.83**	**0.74**	0.34	**0.68**
834	G003.094+00.164	**0.67**	0.28	**0.52**	0.24
920	G003.310−00.402	**0.89**	**0.61**	**0.73**	**0.66**
929	G003.350−00.080	**0.65**	**0.99**	**0.56**	**0.99**
937	G003.410+00.880	**0.51**	0.35	0.13	0.17
946	G003.438−00.352	**0.95**	**1.00**	**0.98**	**0.99**
986	G003.910−00.002	**0.61**	0.33	0.10	0.24
987	G003.932−00.008	**0.64**	0.11	0.09	0.10
1018	G004.418+00.124	0.43	**0.51**	0.13	**0.89**
1020	G004.434+00.126	**0.91**	**0.95**	**0.59**	**0.89**
1039	G004.681+00.277	**0.67**	**0.55**	0.15	0.49
1060	G004.885−00.171	0.06	0.30	0.04	**0.58**
1114	G005.621−00.081	0.36	**0.66**	0.29	**0.86**
1116	G005.641+00.239	**0.96**	**0.99**	**0.96**	**0.95**
1129	G005.833−00.511	**0.75**	0.39	**0.72**	**0.62**
1130	G005.837−00.397	0.28	0.25	**0.53**	0.22
1135	G005.883−00.357	**0.54**	**0.51**	0.23	0.17
1136	G005.887−00.391	**0.98**	**1.00**	**1.00**	**1.00**
1138	G005.897−00.319	**0.79**	**0.70**	0.37	**0.50**
1140	G005.901−00.443	**0.99**	**1.00**	**0.99**	**0.99**
1141	G005.903−00.429	**0.99**	**1.00**	**1.00**	**0.99**
1142	G005.911−00.543	**0.70**	0.31	0.18	0.24
1175	G006.191−00.359	**0.97**	**1.00**	**0.93**	**0.96**
1188	G006.249−00.123	**0.69**	0.45	0.13	0.28
1216	G006.553−00.097	**0.93**	**0.98**	**0.87**	**0.95**
1240	G006.799−00.255	**0.99**	**1.00**	**0.93**	**0.98**
1250	G006.919−00.225	**0.87**	**0.64**	**0.61**	**0.64**
1269	G007.167+00.133	0.22	0.33	0.14	**0.64**
1281	G007.269−00.529	**0.59**	**0.64**	0.33	**0.51**
1286	G007.289−00.529	0.14	0.29	0.13	**0.58**
1305	G007.475+00.061	**0.85**	**0.96**	**0.81**	**0.83**
1314	G007.632−00.110	0.48	**0.68**	0.11	**0.77**
1316	G007.636−00.194	**0.61**	0.13	0.09	0.19
1326	G007.992−00.268	**0.96**	**0.96**	**0.76**	**0.91**
1337	G008.141+00.224	**0.75**	**1.00**	**1.00**	**0.99**
1347	G008.282+00.164	**0.55**	0.10	0.11	0.05
1354	G008.352−00.318	0.10	0.49	0.35	**0.62**
1358	G008.400−00.290	**0.69**	**0.70**	0.44	**0.64**
1359	G008.407−00.350	**0.71**	**0.54**	**0.60**	**0.64**
1366	G008.506−00.280	0.19	0.14	0.07	**0.66**
1377	G008.670−00.356	**0.95**	**1.00**	**1.00**	**1.00**
1383	G008.734−00.364	**0.60**	0.34	0.34	0.24
1399	G008.874−00.494	**0.56**	0.14	0.24	0.11
1421	G009.620+00.194	**0.95**	**1.00**	**0.99**	**1.00**
1435	G009.986−00.030	**0.50**	**0.65**	0.13	0.42
1452	G010.134−00.376	**0.57**	0.41	0.19	**0.84**
1454	G010.150−00.408	**0.60**	0.32	0.16	0.24
1455	G010.152−00.344	**0.73**	**1.00**	**0.62**	**0.98**
1456	G010.166−00.360	**0.83**	**0.99**	**0.99**	**0.93**
1459	G010.192−00.390	**0.59**	**0.60**	0.08	0.41
1462	G010.204−00.348	**0.71**	**0.75**	0.16	**0.62**
1465	G010.212−00.310	**0.69**	0.20	0.04	0.44
1474	G010.286−00.120	**0.83**	**0.91**	**1.00**	**0.80**

Table A1. Continued.

Bolocam catalogue #	BGPS name	Random forests	Logistic regression	LDA	Normalised LDA
1476	G010.300−00.148	**0.98**	**1.00**	**1.00**	**1.00**
1480	G010.324−00.162	**0.98**	**1.00**	**0.99**	**0.88**
1483	G010.343−00.144	**0.78**	**0.99**	**0.73**	**0.94**
1495	G010.446−00.018	**0.82**	**0.80**	0.22	**0.63**
1507	G010.625−00.338	**0.56**	**0.96**	**0.70**	**0.96**
1518	G010.681−00.028	0.23	**0.55**	0.27	**0.70**
1562	G010.973−00.094	0.26	0.27	0.29	**0.59**
1566	G010.989−00.084	**0.51**	0.25	0.26	0.19
1574	G011.035+00.062	0.22	0.30	0.26	**0.52**
1584	G011.083−00.536	0.34	0.45	**0.55**	0.46
1590	G011.111−00.398	**0.67**	**0.80**	**0.65**	**0.75**
1655	G011.904−00.140	**0.85**	**0.88**	**0.63**	**0.74**
1659	G011.947−00.036	**0.60**	**0.90**	0.34	**0.96**
1676	G012.113−00.128	**0.65**	0.22	0.08	0.17
1683	G012.209−00.104	**0.96**	**1.00**	**0.76**	**0.99**
1684	G012.215−00.118	**0.85**	**0.89**	0.01	**0.53**
1708	G012.403−00.466	0.38	**0.52**	0.41	0.48
1710	G012.419+00.506	**0.96**	**1.00**	**0.99**	**0.98**
1747	G012.681−00.182	**0.87**	**0.99**	**1.00**	**0.96**
1758	G012.721−00.216	**0.91**	**0.91**	**0.61**	**0.85**
1762	G012.739−00.102	0.15	0.25	**0.51**	0.30
1771	G012.773+00.334	**0.91**	**0.81**	**0.79**	**0.67**
1780	G012.809−00.200	**0.93**	**1.00**	**1.00**	**1.00**
1792	G012.853−00.226	**0.79**	**0.97**	0.18	**0.91**
1796	G012.861−00.272	**0.57**	**0.88**	0.21	**0.61**
1801	G012.879−00.288	**0.53**	0.37	0.10	0.40
1804	G012.891−00.224	**0.65**	0.28	0.16	**0.61**
1805	G012.895−00.282	**0.62**	0.39	0.09	0.37
1810	G012.909−00.260	**0.77**	**1.00**	**1.00**	**0.99**
1813	G012.917−00.334	**0.66**	0.39	0.29	0.44
1833	G012.999−00.358	**0.67**	**0.84**	**0.54**	**0.77**
1869	G013.211−00.142	**0.98**	**0.98**	**0.95**	**0.82**
1871	G013.217+00.036	**0.76**	0.23	**0.84**	0.26
1876	G013.245−00.084	**0.93**	**0.94**	**0.59**	**0.88**
1883	G013.275−00.336	**0.56**	0.12	0.36	0.04
1894	G013.333−00.038	**0.51**	0.28	0.25	0.43
1905	G013.387+00.066	0.28	**0.54**	0.36	0.48
1954	G013.874+00.281	**0.97**	**1.00**	**1.00**	**0.98**
1974	G013.971−00.411	0.28	0.30	0.05	**0.62**
1984	G014.012−00.175	0.40	**0.55**	0.17	0.36
1985	G014.016−00.133	**0.57**	**0.52**	**0.64**	**0.60**
1995	G014.089−00.557	**0.52**	0.41	0.15	**0.85**
1997	G014.102+00.087	**0.79**	**0.93**	0.45	**0.85**
2007	G014.181−00.529	**0.66**	**0.59**	0.14	**0.57**
2009	G014.183−00.503	**0.74**	0.21	0.24	0.40
2011	G014.194−00.193	**0.62**	**0.96**	**0.70**	**0.91**
2016	G014.227−00.513	**0.93**	**1.00**	**0.88**	**0.96**
2019	G014.244−00.071	**0.78**	**0.88**	**0.67**	**0.81**
2027	G014.327−00.533	**0.55**	0.31	0.11	0.11
2050	G014.466−00.089	**0.76**	0.48	0.27	0.43
2051	G014.474−00.007	**0.60**	0.27	0.44	0.14
2054	G014.492−00.139	**0.80**	**0.77**	**0.51**	**0.67**
2072	G014.606+00.012	**0.51**	**0.89**	**0.75**	**0.95**
2081	G014.633−00.574	**0.90**	**1.00**	**0.98**	**0.95**
2082	G014.634+00.308	**0.67**	**0.72**	0.46	**0.68**
2101	G014.736−00.102	0.02	0.05	0.10	**0.59**
2106	G014.760−00.180	0.13	0.27	0.06	**0.56**
2136	G014.918+00.068	0.36	**0.51**	0.23	0.49
2146	G014.973−00.746	**0.56**	0.18	0.09	**0.89**

Table A1. Continued.

Bolocam catalogue #	BGPS name	Random forests	Logistic regression	LDA	Normalised LDA
2147	G014.983−00.692	**0.85**	**1.00**	**0.96**	**0.98**
2150	G014.991−00.738	**0.89**	**0.91**	**0.73**	**0.70**
2151	G015.004+00.010	**0.71**	0.24	0.07	0.15
2152	G015.013−00.674	**0.79**	**1.00**	**1.00**	**1.00**
2153	G015.021−00.620	**0.85**	**0.99**	0.24	**0.77**
2155	G015.031−00.670	**0.82**	**1.00**	**1.00**	**1.00**
2156	G015.031−00.746	**0.59**	**0.68**	0.25	**0.93**
2157	G015.045−00.650	**0.89**	**0.94**	0.00	**0.97**
2159	G015.057−00.624	**0.83**	**0.91**	0.44	**0.56**
2162	G015.079−00.604	**0.69**	**0.52**	0.43	**0.74**
2165	G015.093−00.676	**0.83**	0.36	0.36	**0.84**
2167	G015.095−00.710	**0.99**	**0.99**	0.22	**0.79**
2168	G015.097−00.734	**0.84**	**0.65**	**0.73**	**0.70**
2169	G015.099−00.558	**0.55**	0.43	0.30	0.33
2170	G015.099−00.600	**0.82**	0.43	0.42	0.30
2171	G015.101−00.656	**0.90**	**0.94**	**0.64**	**0.80**
2181	G015.137−00.674	**0.85**	**0.66**	**0.73**	0.35
2184	G015.153−00.660	**0.64**	**0.67**	0.17	0.30
2189	G015.182−00.158	0.26	0.47	0.33	**0.67**
2190	G015.195−00.628	**0.94**	**1.00**	**0.71**	**0.96**
2191	G015.201−00.442	**0.56**	**0.54**	**0.50**	**0.55**
2193	G015.205−00.626	**0.82**	0.14	**0.89**	**0.57**
2195	G015.234−00.612	**0.74**	0.39	0.37	**0.88**
2198	G015.250−00.602	**0.68**	0.48	0.05	**0.67**
2224	G015.557−00.463	0.23	**0.55**	0.23	**0.67**
2234	G015.665−00.499	**0.64**	**0.62**	0.34	**0.59**
2248	G016.144+00.009	**0.63**	0.17	0.11	0.15
2274	G016.362−00.355	0.20	0.03	0.05	**0.51**
2275	G016.364−00.209	**0.86**	**0.92**	**0.97**	**0.84**
2311	G016.821−00.344	**0.66**	**0.75**	0.46	**0.62**
2312	G016.832+00.080	**0.67**	0.40	0.13	0.31
2320	G016.926+00.298	0.22	0.05	0.06	**0.53**
2325	G016.946−00.074	0.36	**0.78**	0.11	**0.86**
2343	G017.366−00.034	0.26	0.07	0.05	**0.59**
2351	G017.638+00.154	**0.95**	**1.00**	**0.98**	**0.98**
2365	G018.091−00.302	**0.71**	0.23	**0.60**	0.17
2375	G018.150−00.286	**0.84**	**0.94**	0.43	**0.70**
2377	G018.173−00.298	**0.54**	**0.62**	0.45	**0.57**
2386	G018.260−00.246	**0.55**	0.36	**0.82**	0.48
2387	G018.277−00.262	**0.52**	**0.76**	0.14	**0.82**
2388	G018.302−00.390	**0.97**	**1.00**	**0.99**	**0.98**
2396	G018.462−00.002	**0.85**	**0.99**	0.35	**0.91**
2424	G018.608−00.074	0.19	**0.55**	0.21	**0.70**
2430	G018.655−00.060	**0.78**	**0.76**	0.44	**0.72**
2431	G018.666+00.032	0.26	0.30	**0.56**	**0.57**
2442	G018.738−00.225	**0.90**	**0.96**	**0.83**	**0.92**
2455	G018.830−00.483	**0.84**	0.38	**0.96**	**0.56**
2456	G018.834−00.299	0.33	0.38	0.14	**0.58**
2510	G019.077−00.287	**0.86**	**1.00**	**0.93**	**0.97**
2561	G019.364−00.031	**0.91**	**0.93**	**0.91**	**0.89**
2573	G019.474+00.171	**0.94**	**1.00**	**0.67**	**0.99**
2601	G019.609−00.233	**0.98**	**1.00**	**0.97**	**1.00**
2602	G019.612−00.137	0.38	**0.70**	0.34	**0.76**
2603	G019.614−00.257	**0.63**	**0.62**	0.45	**0.56**
2612	G019.702−00.263	**0.72**	**0.65**	0.28	**0.61**
2619	G019.756−00.129	0.16	**0.54**	0.28	**0.57**
2673	G020.366−00.011	0.07	0.39	0.16	**0.59**
2718	G020.734−00.059	**0.82**	**0.93**	0.29	**0.90**
2720	G020.750−00.091	**0.80**	**0.81**	**0.54**	**0.86**
2722	G020.763−00.059	**0.53**	0.46	0.10	0.31

Table A1. Continued.

Bolocam catalogue #	BGPS name	Random forests	Logistic regression	LDA	Normalised LDA
2741	G020.984+00.097	0.30	0.23	**0.61**	0.16
2784	G021.385−00.253	**0.60**	**0.88**	0.28	**0.89**
2788	G021.423−00.541	**0.76**	**0.71**	0.45	**0.71**
2819	G021.878+00.007	**0.67**	**0.72**	0.41	**0.62**
2854	G022.353+00.067	**0.65**	**0.80**	0.39	**0.87**
2860	G022.379+00.447	0.32	**0.50**	0.19	0.45
2864	G022.417+00.315	**0.57**	0.13	**0.61**	0.14
2907	G022.725−00.274	**0.53**	0.27	0.10	0.15
2971	G023.012−00.410	**0.93**	**0.99**	**0.98**	**0.94**
3016	G023.202−00.000	**0.86**	**0.77**	**0.60**	**0.80**
3018	G023.208−00.378	**0.94**	**1.00**	**0.70**	**1.00**
3026	G023.268+00.078	**0.60**	**0.88**	0.32	**0.75**
3027	G023.272−00.258	**0.87**	**0.81**	**0.81**	**0.77**
3029	G023.274−00.212	**0.52**	0.29	0.40	0.32
3039	G023.321−00.298	**0.50**	0.44	**0.58**	**0.55**
3053	G023.368−00.290	**0.77**	**0.83**	0.26	**0.68**
3065	G023.414−00.228	**0.89**	0.03	**0.91**	0.31
3077	G023.456+00.064	**0.69**	**0.89**	0.33	**0.85**
3078	G023.456−00.018	0.24	0.05	0.06	**0.73**
3086	G023.484+00.096	**0.75**	**0.57**	0.36	0.46
3116	G023.571+00.014	**0.81**	**0.86**	0.47	**0.73**
3141	G023.658−00.142	**0.53**	0.11	0.13	0.12
3155	G023.711+00.170	**0.73**	**0.97**	0.49	**0.93**
3183	G023.870−00.124	**0.74**	**0.50**	**0.84**	0.45
3186	G023.888+00.060	0.15	0.14	0.14	**0.65**
3189	G023.902+00.064	0.27	**0.66**	0.11	**0.72**
3200	G023.955+00.150	**0.90**	**0.96**	**0.64**	**0.83**
3205	G023.992−00.092	0.35	0.15	**0.70**	0.31
3212	G024.018+00.048	**0.56**	0.32	0.26	0.30
3307	G024.402−00.190	**0.53**	0.29	0.15	0.25
3313	G024.414+00.102	**0.57**	0.31	**0.63**	0.47
3320	G024.439+00.228	**0.61**	0.13	0.36	0.05
3322	G024.443−00.228	**0.95**	**0.99**	**0.87**	**0.95**
3326	G024.461+00.196	**0.73**	**0.60**	0.19	0.38
3329	G024.472+00.490	**0.68**	0.38	**0.97**	**0.66**
3337	G024.494−00.040	**0.83**	**1.00**	**0.98**	**0.99**
3343	G024.510−00.220	**0.84**	**0.66**	**0.60**	**0.60**
3357	G024.545−00.248	**0.53**	**0.57**	**0.56**	0.42
3409	G024.757+00.091	**0.69**	**0.86**	0.25	**0.74**
3440	G024.943+00.075	**0.52**	**0.75**	0.05	**0.92**
3461	G025.155−00.275	**0.61**	0.34	0.49	0.39
3474	G025.227+00.289	**0.53**	**0.83**	0.30	**0.71**
3497	G025.353−00.193	**0.63**	0.35	**0.61**	0.47
3502	G025.384−00.181	**0.98**	**1.00**	**0.96**	**0.94**
3507	G025.400−00.141	**0.92**	**1.00**	**0.98**	**0.95**
3511	G025.411+00.103	**0.61**	0.39	0.17	0.24
3519	G025.456−00.211	**0.54**	**0.60**	**0.73**	**0.52**
3576	G025.713+00.045	**0.70**	**0.75**	0.37	**0.72**
3582	G025.737+00.213	**0.54**	0.24	0.32	0.11
3588	G025.797+00.245	**0.57**	**0.52**	0.35	**0.51**
3591	G025.805−00.041	**0.56**	**0.88**	0.32	**0.87**
3594	G025.827−00.179	**0.93**	**1.00**	**0.92**	**1.00**
3645	G026.209+00.025	0.24	0.09	0.06	**0.58**
3679	G026.510+00.281	**0.99**	**1.00**	**0.91**	**1.00**
3685	G026.545−00.293	**0.58**	0.41	0.13	0.47
3690	G026.562−00.303	**0.80**	**0.73**	0.47	**0.85**
3766	G027.187−00.083	0.40	**0.90**	0.25	**0.93**
3774	G027.283+00.149	**0.51**	0.41	**0.52**	**0.52**
3782	G027.367−00.167	**0.94**	**1.00**	**0.99**	**1.00**
3807	G027.562+00.080	**0.95**	**0.69**	**0.84**	**0.64**

Table A1. Continued.

Bolocam catalogue #	BGPS name	Random forests	Logistic regression	LDA	Normalised LDA
3852	G027.903−00.016	0.15	0.44	0.36	**0.61**
3864	G027.977+00.076	**0.69**	0.35	**0.52**	0.46
3899	G028.149+00.148	0.12	0.08	0.03	**0.62**
3913	G028.201−00.052	**0.84**	**1.00**	**0.99**	**1.00**
3921	G028.241+00.058	**0.64**	0.09	0.30	0.03
3925	G028.285−00.364	**0.92**	**0.97**	**0.91**	**0.90**
3929	G028.305−00.388	**0.58**	0.29	**0.57**	0.42
3936	G028.337+00.116	0.43	**0.67**	0.29	**0.72**
3939	G028.344+00.058	**0.52**	0.34	0.18	0.31
3955	G028.397+00.078	**0.99**	**1.00**	**0.97**	**0.97**
3998	G028.565−00.236	**0.53**	**0.62**	**0.66**	**0.61**
4006	G028.609+00.016	**0.92**	**0.92**	0.46	**0.85**
4014	G028.651+00.026	**0.90**	**0.96**	0.39	**0.83**
4048	G028.811+00.169	**0.63**	0.26	**0.61**	0.40
4049	G028.817+00.363	0.34	**0.81**	0.18	**0.87**
4055	G028.831−00.255	**0.86**	**1.00**	**0.66**	**0.99**
4061	G028.863+00.065	**0.73**	**0.98**	**0.60**	**0.94**
4063	G028.881−00.025	**0.64**	0.28	0.12	0.19
4121	G029.225+00.023	**0.54**	0.36	0.10	0.15
4152	G029.397−00.095	**0.58**	**0.56**	0.20	**0.53**
4154	G029.435−00.177	**0.60**	0.12	0.12	0.05
4236	G029.855−00.056	**0.60**	0.36	0.41	0.50
4239	G029.863−00.048	**0.58**	**0.60**	0.32	**0.70**
4243	G029.888−00.000	0.24	0.06	**0.52**	0.04
4252	G029.913−00.046	**0.87**	**0.70**	**0.90**	**0.55**
4254	G029.920−00.016	**0.54**	0.34	0.09	0.19
4258	G029.933−00.064	**0.90**	**0.79**	**0.55**	0.46
4259	G029.937−00.790	0.29	0.08	0.07	**0.53**
4261	G029.943+00.072	0.14	0.06	0.04	**0.62**
4266	G029.955−00.018	**0.97**	**1.00**	**0.97**	**1.00**
4272	G029.975−00.050	**0.93**	**0.93**	**0.72**	**0.75**
4281	G030.004−00.270	**0.85**	**0.82**	**0.89**	**0.80**
4384	G030.387−00.106	**0.84**	**0.71**	**0.50**	**0.73**
4449	G030.536+00.021	0.39	**0.86**	0.32	**0.89**
4468	G030.590−00.043	**0.81**	**1.00**	**0.91**	**0.99**
4488	G030.652−00.203	**0.66**	**0.81**	**0.74**	**0.78**
4499	G030.688−00.261	0.39	**0.60**	0.46	**0.83**
4500	G030.688−00.039	**0.87**	**0.63**	**0.93**	**0.78**
4509	G030.704−00.067	**0.91**	**1.00**	**1.00**	**1.00**
4518	G030.719−00.081	**0.94**	**1.00**	**0.91**	**0.99**
4526	G030.746−00.059	**0.89**	**0.99**	**0.96**	**0.90**
4527	G030.746+00.001	**0.72**	0.09	**0.53**	0.11
4530	G030.756−00.051	**0.90**	**1.00**	**0.97**	**0.96**
4533	G030.760+00.207	**0.78**	**0.58**	0.48	**0.56**
4537	G030.768−00.039	**0.96**	**0.91**	0.09	**0.85**
4541	G030.776−00.215	**0.60**	0.50	**0.52**	**0.60**
4546	G030.788−00.025	**0.81**	**0.94**	**0.99**	**0.83**
4547	G030.788+00.205	**0.96**	**0.99**	0.33	**0.93**
4553	G030.802+00.115	0.32	0.10	0.04	**0.79**
4555	G030.808−00.027	**0.98**	**0.97**	**0.75**	**0.87**
4560	G030.820−00.055	**0.94**	**1.00**	**1.00**	**1.00**
4566	G030.830+00.135	**0.51**	0.05	0.21	0.02
4573	G030.850−00.081	**0.74**	**0.76**	0.34	**0.56**
4582	G030.868+00.115	**0.94**	**1.00**	0.50	**0.97**
4583	G030.870−00.155	0.34	0.29	0.09	**0.88**
4586	G030.878+00.059	**0.58**	0.20	0.40	0.10
4594	G030.896+00.139	**0.56**	0.43	0.30	**0.50**
4598	G030.900+00.163	0.49	**0.55**	0.21	**0.60**
4633	G030.974−00.139	**0.67**	0.32	0.43	0.28
4636	G030.980+00.215	0.24	0.44	0.23	**0.68**

Table A1. Continued.

Bolocam catalogue #	BGPS name	Random forests	Logistic regression	LDA	Normalised LDA
4654	G031.028+00.265	**0.53**	0.06	0.11	0.04
4662	G031.050+00.357	**0.58**	0.35	0.38	0.28
4695	G031.160+00.049	**0.55**	0.42	0.32	0.46
4722	G031.246−00.111	**0.68**	**0.98**	**0.61**	**0.99**
4736	G031.282+00.063	**0.91**	**1.00**	**0.94**	**1.00**
4760	G031.398−00.257	**0.87**	**1.00**	**0.67**	**0.99**
4764	G031.414+00.307	**0.88**	**1.00**	**1.00**	**1.00**
4911	G032.021+00.063	**0.81**	0.40	**0.58**	0.48
4916	G032.044+00.059	**0.96**	**1.00**	**0.94**	**0.91**
4926	G032.119+00.091	0.31	**0.53**	0.39	**0.71**
4933	G032.152+00.135	**0.90**	**0.99**	**0.91**	**0.97**
4975	G032.474+00.205	**0.61**	0.40	**0.53**	0.38
5041	G032.744−00.075	**0.90**	**0.99**	**0.74**	**0.91**
5053	G032.798+00.193	**0.93**	**1.00**	**0.99**	**1.00**
5057	G032.820−00.329	0.34	**0.62**	0.30	**0.80**
5120	G033.133−00.091	**0.60**	**0.99**	**0.85**	**0.99**
5171	G033.414−00.002	**0.58**	0.24	0.32	0.19
5229	G033.652−00.025	**0.55**	0.18	0.46	0.22
5263	G033.810−00.187	0.20	0.49	0.33	**0.51**
5278	G033.914+00.107	**0.82**	**1.00**	**1.00**	**0.99**
5306	G034.091+00.015	**0.83**	0.47	**0.72**	0.45
5321	G034.191−00.594	**0.63**	0.41	0.11	0.37
5340	G034.258+00.154	**0.92**	**1.00**	**1.00**	**1.00**
5346	G034.283+00.184	0.33	**0.55**	0.05	0.20
5384	G034.454+00.006	0.49	**0.55**	0.39	**0.90**
5385	G034.457+00.248	**0.57**	0.49	**0.52**	0.45
5433	G034.712−00.596	**0.76**	**0.56**	**0.73**	0.40
5467	G034.820+00.350	**0.92**	**0.94**	**0.78**	**0.87**
5530	G035.026+00.350	**0.92**	**1.00**	**0.54**	**0.97**
5538	G035.045−00.478	**0.58**	0.05	0.25	0.02
5627	G035.466+00.138	**0.86**	**0.98**	**0.91**	**0.92**
5653	G035.576+00.066	**0.84**	0.35	**0.64**	0.35
5654	G035.576−00.032	**0.92**	**1.00**	**0.67**	**0.99**
5657	G035.579+00.006	**0.84**	**0.50**	**0.69**	0.36
5695	G035.750+00.152	**0.81**	**0.58**	**0.67**	**0.62**
5700	G035.794−00.176	**0.64**	**0.85**	0.45	**0.79**
5756	G036.405+00.020	0.36	0.46	0.27	**0.81**
5849	G037.547−00.112	**0.66**	**0.83**	**0.54**	**0.83**
5850	G037.555+00.200	**0.85**	**0.93**	**0.50**	**0.90**
5853	G037.599+00.426	**0.64**	0.31	0.08	0.21
5864	G037.737−00.112	**0.51**	**0.81**	0.49	**0.88**
5874	G037.820+00.412	**0.83**	**0.90**	0.25	**0.75**
5879	G037.875−00.400	**0.99**	**1.00**	**0.97**	**0.98**
5931	G038.694−00.454	0.21	0.22	0.19	**0.53**
5956	G038.920−00.352	**0.81**	**0.93**	**0.81**	**0.92**
5972	G039.256−00.059	**0.60**	0.12	0.16	0.06
5980	G039.389−00.143	0.44	0.41	0.19	**0.59**
6006	G039.883−00.347	0.18	**0.52**	0.33	**0.62**
6024	G040.283−00.221	**0.97**	**1.00**	**0.74**	**0.99**
6029	G040.622−00.139	**0.70**	**0.83**	**0.56**	**0.82**
6082	G041.741+00.095	**0.62**	0.23	0.11	0.17
6117	G043.164−00.031	**0.96**	**1.00**	**1.00**	**1.00**
6118	G043.169+00.009	**0.91**	**1.00**	**1.00**	**1.00**
6119	G043.177−00.521	**0.98**	**0.99**	**0.92**	**0.92**
6120	G043.237−00.047	**0.89**	**0.99**	**0.90**	**0.88**
6122	G043.307−00.213	**0.73**	**0.97**	0.21	**0.97**
6126	G043.795−00.125	**0.98**	**1.00**	**0.64**	**0.97**
6142	G044.307+00.041	**0.81**	**0.64**	0.42	**0.53**
6162	G045.069+00.133	**0.99**	**1.00**	**0.79**	**0.99**
6165	G045.121+00.133	**0.98**	**1.00**	**1.00**	**0.99**

Table A1. Continued.

Bolocam catalogue #	BGPS name	Random forests	Logistic regression	LDA	Normalised LDA
6172	G045.453+00.061	**0.95**	**1.00**	**0.92**	**0.98**
6176	G045.465+00.047	**0.97**	**1.00**	**0.58**	**0.94**
6177	G045.477+00.135	**0.89**	**0.93**	**0.90**	**0.84**
6202	G045.805−00.355	**0.57**	0.25	0.19	0.26
6254	G048.579+00.056	**0.66**	0.37	**0.65**	0.36
6256	G048.603+00.024	**0.74**	**0.92**	**0.54**	**0.87**
6286	G048.895−00.410	**0.53**	0.20	0.09	0.09
6287	G048.914−00.280	**0.71**	0.04	**0.90**	**0.51**
6291	G048.989−00.300	**0.86**	**1.00**	**0.99**	**0.97**
6292	G048.997−00.312	**0.73**	0.32	0.07	**0.92**
6299	G049.070−00.350	**0.51**	0.21	0.17	0.11
6300	G049.075−00.276	0.27	0.09	0.15	**0.65**
6310	G049.170−00.208	**0.60**	0.35	**0.68**	0.43
6312	G049.192−00.336	**0.93**	**0.88**	**0.74**	**0.94**
6313	G049.210−00.342	**0.89**	**0.97**	**1.00**	**0.92**
6321	G049.264+00.312	**0.63**	0.09	0.10	0.07
6334	G049.367−00.302	**0.87**	**1.00**	**1.00**	**0.98**
6336	G049.371−00.350	**0.70**	0.25	**0.52**	0.20
6337	G049.375−00.262	**0.87**	**0.84**	**0.66**	**0.55**
6339	G049.389−00.320	**0.92**	**0.87**	**0.86**	**0.97**
6340	G049.390−00.310	**0.88**	**0.97**	**0.53**	**0.87**
6344	G049.402−00.214	**0.67**	0.14	0.34	0.10
6362	G049.489−00.370	**0.77**	**1.00**	**0.96**	**1.00**
6363	G049.489−00.386	**0.95**	**1.00**	**1.00**	**1.00**
6365	G049.529−00.346	**0.63**	0.33	0.37	0.18
6371	G049.561−00.276	**0.78**	**0.59**	**0.61**	0.43
6389	G050.283−00.390	0.18	**0.50**	0.19	**0.61**
6402	G051.375−00.011	**0.52**	0.25	**0.63**	0.35
6406	G052.752+00.336	**0.62**	**0.59**	0.09	0.35
6410	G053.036+00.112	**0.63**	**0.71**	0.18	**0.79**
6425	G053.259+00.040	0.43	0.32	0.11	**0.57**
6446	G053.957+00.032	**0.55**	0.38	0.13	0.33
6448	G054.108−00.049	**0.55**	0.10	0.26	0.05
6452	G054.120−00.075	**0.51**	0.14	0.08	**0.54**
6467	G056.250−00.160	0.25	0.03	0.11	**0.56**
6486	G059.786+00.067	**0.82**	**1.00**	**0.96**	**0.98**
6495	G060.887−00.129	**0.90**	**1.00**	**1.00**	**0.93**
6497	G061.475+00.090	**0.90**	**1.00**	**1.00**	**1.00**
6502	G063.115+00.340	**0.59**	**0.61**	0.29	0.49
6506	G071.149+00.402	**0.75**	0.36	0.13	0.29
6508	G072.954−00.028	**0.57**	0.14	0.14	0.10
6521	G075.757+00.339	**0.90**	**1.00**	**1.00**	**0.99**
6523	G075.784+00.341	**0.95**	**1.00**	**1.00**	**0.99**
6528	G075.835+00.399	**0.84**	**1.00**	**1.00**	**0.98**
6529	G075.841+00.367	**0.85**	**0.74**	**0.69**	**0.77**
6530	G075.843+00.359	**0.80**	**0.66**	**0.61**	**0.96**
6547	G076.156−00.287	**0.84**	**0.59**	**0.72**	**0.67**
6550	G076.186+00.095	**0.88**	**0.53**	**0.78**	0.46
6555	G076.358−00.601	**0.84**	**0.69**	**0.68**	**0.60**
6556	G076.382−00.623	**0.92**	**1.00**	**1.00**	**1.00**
6562	G077.475−01.083	**0.57**	0.45	0.05	0.22
6569	G077.820−01.313	**0.55**	0.43	0.14	0.39
6588	G077.978+00.577	0.26	0.03	0.04	**0.61**
6599	G078.034+00.617	**0.58**	0.23	0.06	0.18
6602	G078.106−00.317	**0.68**	**0.69**	**0.66**	**0.73**
6604	G078.114−00.637	**0.89**	**0.57**	**0.71**	0.42
6631	G078.379+01.017	0.41	**0.50**	**0.54**	0.40
6652	G078.888+00.709	**0.95**	**1.00**	**0.99**	**0.99**
6657	G078.978+00.351	**0.83**	**0.95**	**0.99**	**0.84**
6669	G079.132−00.369	**0.87**	**0.95**	**0.87**	**0.87**

Table A1. Continued.

Bolocam catalogue #	BGPS name	Random forests	Logistic regression	LDA	Normalised LDA
6683	G079.289+01.301	**0.50**	0.36	0.45	0.35
6685	G079.296+00.283	**0.68**	0.48	0.39	**0.63**
6686	G079.308+01.307	**0.82**	**0.55**	0.28	0.38
6687	G079.313+00.279	**0.84**	**0.72**	**0.52**	**0.67**
6690	G079.335+00.341	**0.51**	0.35	**0.60**	**0.57**
6698	G079.483−00.719	**0.55**	0.49	0.18	0.42
6703	G079.563−00.767	**0.94**	**1.00**	**0.54**	**0.95**
6704	G079.643+00.473	**0.62**	**0.95**	0.12	**0.71**
6712	G079.879+01.179	**0.93**	**0.92**	**0.93**	**0.84**
6718	G079.981+00.811	0.45	**0.52**	0.32	**0.50**
6721	G079.986+00.839	**0.60**	0.13	0.48	0.07
6730	G080.364+00.445	**0.87**	0.37	**0.65**	0.27
6741	G080.635+00.686	**0.73**	**0.55**	**0.64**	**0.52**
6747	G080.829+00.568	**0.74**	**0.69**	**0.55**	**0.61**
6753	G080.863+00.384	**0.79**	**0.96**	**0.69**	**0.94**
6754	G080.864+00.422	**0.79**	**0.98**	**0.90**	**0.97**
6762	G080.941−00.126	**0.69**	**0.61**	**0.75**	**0.72**
6765	G080.954−00.154	**0.61**	0.45	0.19	0.34
6788	G081.117−00.140	**0.72**	0.28	**0.61**	0.18
6796	G081.174−00.100	**0.93**	**0.99**	**0.74**	**0.99**
6808	G081.260+00.984	0.33	0.06	0.04	**0.75**
6815	G081.302+01.052	**0.96**	**1.00**	**0.98**	**0.99**
6820	G081.344+00.760	**0.84**	**1.00**	**0.91**	**0.95**
6839	G081.451+00.470	**0.57**	0.10	0.33	0.06
6840	G081.457+00.018	**0.61**	0.29	0.32	**0.57**
6844	G081.477+00.022	**0.71**	0.09	**0.77**	0.33
6859	G081.542+00.986	0.33	**0.70**	0.17	**0.71**
6863	G081.549+00.096	**0.64**	**0.68**	0.48	**0.60**
6872	G081.582+00.104	**0.68**	**0.80**	0.35	**0.71**
6901	G081.680+00.540	**0.93**	**1.00**	**1.00**	**1.00**
6909	G081.721+00.572	**0.97**	**1.00**	**1.00**	**1.00**
6920	G081.753+00.593	**0.86**	**1.00**	**1.00**	**0.98**
6926	G081.765+00.641	0.23	**0.55**	0.08	0.20
6934	G081.783+00.621	**0.68**	0.37	0.12	**0.85**
6941	G081.831+00.853	0.32	**0.53**	0.15	**0.66**
6947	G081.844+00.881	**0.83**	0.30	**0.63**	0.19
6955	G081.875+00.783	**0.84**	**1.00**	**1.00**	**1.00**
7069	G084.548+00.104	0.36	**0.67**	0.20	**0.68**
7097	G084.774−01.184	**0.55**	0.28	0.09	0.17
7098	G084.784−01.104	0.37	0.14	0.03	**0.80**
7099	G084.805−01.112	**0.84**	**0.73**	0.19	**0.51**
7101	G084.829−01.092	**0.73**	0.06	0.45	0.29
7104	G084.844−01.084	**0.66**	0.22	0.26	0.42
7111	G084.951−00.692	0.37	**0.55**	0.17	**0.80**
7121	G085.042−00.144	**0.72**	0.21	**0.78**	0.28
7126	G085.073−00.140	0.18	0.14	**0.59**	0.22
7140	G085.412+00.002	**0.53**	**0.78**	0.41	**0.70**
7146	G089.635+00.171	**0.50**	0.08	0.14	0.06
7149	G098.978+03.960	**0.61**	0.35	0.17	0.20
7150	G099.115+03.926	**0.62**	**0.53**	0.22	**0.52**
7151	G099.981+04.168	**0.98**	**1.00**	**0.97**	**0.93**
7170	G110.113+00.050	0.48	0.20	**0.68**	0.43
7213	G111.284−00.664	**0.82**	0.14	**0.81**	0.25
7232	G111.447+00.798	0.49	0.05	**0.54**	0.03
7235	G111.484+00.746	**0.62**	0.18	**0.80**	0.33
7243	G111.522+00.800	**0.91**	**0.81**	0.40	**0.57**
7244	G111.528+00.818	**0.79**	**0.57**	**0.51**	**0.56**
7247	G111.537+00.756	**0.95**	**1.00**	**0.99**	**1.00**
7248	G111.545+00.776	**0.95**	**1.00**	**1.00**	**1.00**
7252	G111.573+00.750	**0.68**	**0.96**	**0.84**	**0.98**

Table A1. Continued.

Bolocam catalogue #	BGPS name	Random forests	Logistic regression	LDA	Normalised LDA
7257	G111.597+00.806	**0.57**	0.08	**0.66**	0.08
7260	G111.615+00.374	**0.84**	**0.98**	**0.94**	**0.96**
7305	G111.787+00.586	**0.52**	0.12	0.31	0.15
7322	G111.882+00.992	**0.84**	**0.56**	**0.73**	**0.56**
7331	G111.945+00.808	**0.53**	0.18	0.38	0.06
7351	G133.694+01.215	**0.98**	**1.00**	**1.00**	**1.00**
7352	G133.715+01.217	**0.93**	**1.00**	**1.00**	**1.00**
7361	G133.736+01.271	**0.64**	0.25	0.35	0.13
7364	G133.748+01.197	**0.98**	**1.00**	**0.72**	**0.92**
7367	G133.784+01.421	**0.70**	0.05	0.49	0.03
7374	G133.890+01.137	0.26	0.46	0.10	**0.82**
7380	G133.949+01.063	**0.87**	**1.00**	**1.00**	**1.00**
7392	G134.203+00.753	**0.79**	0.17	**0.53**	0.36
7394	G134.211+00.729	**0.67**	0.04	0.16	0.26
7396	G134.218+00.787	**0.73**	0.26	0.30	0.33
7456	G138.295+01.556	**0.91**	**0.96**	**0.82**	**0.85**
7459	G138.503+01.646	**0.53**	**0.62**	0.38	**0.60**
7460	G188.792+01.027	**0.75**	**0.56**	**0.79**	**0.52**
7461	G188.948+00.883	**0.81**	**1.00**	**1.00**	**0.99**
7465	G189.030+00.781	**0.96**	**1.00**	**0.99**	**0.94**
7466	G189.032+00.793	**0.91**	0.47	**0.72**	**0.66**
7474	G189.776+00.343	**0.91**	**1.00**	**0.85**	**0.98**
7481	G189.804+00.355	**0.82**	**0.71**	0.37	**0.58**
7482	G189.810+00.369	**0.55**	0.17	0.19	0.26
7483	G189.831+00.343	**0.60**	0.18	0.36	0.16
7486	G189.864+00.499	**0.60**	0.24	0.37	0.20
7492	G189.951+00.331	**0.61**	0.14	0.25	0.09
7501	G192.581−00.043	**0.98**	**1.00**	**0.99**	**0.98**
7502	G192.596−00.051	**0.78**	**1.00**	**0.93**	**0.98**
7531	G349.836−00.528	**0.79**	**0.63**	**0.68**	**0.51**
7536	G349.978−00.560	0.38	0.05	0.09	**0.52**
7538	G349.988−00.558	0.22	**0.64**	0.08	0.31
7540	G350.016+00.432	**0.78**	**1.00**	0.24	**0.96**
7545	G350.110+00.090	**0.98**	**1.00**	**1.00**	**0.97**
7546	G350.120+00.060	**0.90**	**0.99**	**0.69**	**0.84**
7549	G350.177+00.014	**0.59**	**0.58**	**0.66**	**0.56**
7558	G350.298+00.122	**0.72**	0.35	0.15	0.26
7559	G350.329+00.100	**0.88**	**0.99**	0.47	**0.85**
7560	G350.341+00.138	**0.59**	0.42	0.14	0.27
7571	G350.521−00.350	0.39	**0.55**	0.37	**0.51**
7577	G350.689−00.492	**0.94**	**0.96**	**0.77**	**0.81**
7591	G350.783−00.028	**0.68**	**0.96**	0.45	**0.87**
7603	G350.975+00.546	0.32	0.35	0.12	**0.71**
7604	G350.978−00.540	**0.68**	0.31	0.10	0.30
7605	G351.040−00.338	**0.96**	**1.00**	**0.95**	**0.97**
7621	G351.465−00.458	**0.72**	0.30	**0.83**	0.19
7628	G351.555+00.206	**0.92**	**1.00**	**0.99**	**0.99**
7632	G351.581−00.352	**1.00**	**1.00**	**1.00**	**1.00**
7635	G351.614+00.164	**0.95**	**1.00**	**0.98**	**0.89**
7645	G351.775−00.538	**0.98**	**1.00**	**1.00**	**1.00**
7646	G351.785−00.514	**0.98**	**1.00**	**0.85**	**0.88**
7650	G351.799−00.488	**0.72**	**0.77**	0.24	**0.65**
7651	G351.802−00.448	**0.96**	**0.97**	**0.85**	**0.87**
7674	G352.098+00.162	**0.70**	**0.85**	0.26	**0.78**
7677	G352.112+00.178	0.34	0.49	0.16	**0.52**
7697	G352.317−00.444	**0.97**	**1.00**	**0.95**	**0.92**
7711	G352.519−00.154	**0.73**	**0.95**	0.34	**0.83**
7714	G352.584−00.184	**0.64**	**0.69**	0.32	**0.60**
7716	G352.608−00.192	**0.59**	0.19	0.17	0.13
7721	G352.684−00.118	**0.56**	**0.66**	0.14	0.39

Table A1. Continued.

Bolocam catalogue #	BGPS name	Random forests	Logistic regression	LDA	Normalised LDA
7725	G352.858−00.202	**0.98**	**1.00**	**0.97**	**0.95**
7728	G352.876−00.516	**0.57**	0.34	0.05	0.18
7733	G353.019+00.504	**0.56**	**0.50**	0.25	**0.53**
7741	G353.067+00.508	**0.74**	**0.76**	0.19	**0.52**
7745	G353.069+00.452	**0.84**	**0.76**	0.17	**0.73**
7747	G353.079+00.422	**0.66**	0.45	0.07	**0.53**
7748	G353.091+00.446	**0.52**	0.32	0.15	0.09
7749	G353.117+00.366	**0.67**	0.40	**0.71**	0.42
7751	G353.216−00.246	**0.53**	**0.85**	0.18	**0.81**
7758	G353.316−00.256	**0.64**	0.13	**0.61**	0.25
7759	G353.334−00.294	**0.53**	0.16	0.10	0.31
7760	G353.343−00.288	**0.58**	0.49	0.26	**0.65**
7761	G353.343−00.290	**0.70**	0.31	0.32	0.43
7763	G353.362−00.088	**0.50**	0.22	0.19	0.38
7764	G353.365−00.166	**0.81**	**0.99**	0.46	**0.87**
7765	G353.367−00.336	**0.72**	0.26	0.34	**0.56**
7767	G353.384−00.336	**0.66**	**0.63**	0.47	**0.75**
7770	G353.400−00.070	**0.89**	**1.00**	0.45	**0.89**
7771	G353.412−00.360	**0.91**	**1.00**	**1.00**	**1.00**
7772	G353.432−00.088	**0.74**	**0.86**	0.14	**0.68**
7779	G353.548−00.016	**0.74**	0.44	0.46	0.40
7791	G353.834+00.268	**0.63**	**0.60**	0.18	**0.56**
7794	G353.978+00.260	**0.54**	0.20	0.21	0.06
7806	G354.208−00.036	0.35	**0.75**	0.22	**0.74**
7811	G354.343+00.474	**0.53**	0.12	0.24	0.05
7820	G354.422+00.032	**0.71**	0.23	0.22	0.22
7832	G354.600+00.474	**0.92**	**0.95**	0.09	**0.80**
7834	G354.617+00.472	**1.00**	**1.00**	**0.63**	**0.99**
7836	G354.662+00.484	**0.72**	**0.97**	**0.66**	**0.91**
7837	G354.672+00.242	**0.61**	0.47	0.06	0.19
7839	G354.711+00.292	**0.85**	**0.99**	0.20	**0.84**
7840	G354.725+00.302	**0.94**	**1.00**	**0.56**	**0.96**
7843	G354.769+00.326	**0.57**	0.44	0.19	**0.74**
7849	G354.826+00.352	0.47	0.23	0.07	**0.78**
7855	G354.946−00.540	**0.68**	**0.84**	0.47	**0.65**
7859	G355.129−00.300	**0.61**	0.31	0.08	0.33
7860	G355.186−00.418	**0.99**	**1.00**	**0.85**	**0.99**
7874	G355.268−00.270	**0.89**	**0.95**	**0.51**	**0.88**
7877	G355.346+00.148	**0.90**	**1.00**	0.31	**0.97**
7881	G355.413+00.102	**0.56**	**0.87**	0.10	**0.79**
7897	G355.742+00.132	**0.61**	0.49	0.23	**0.51**
7901	G355.828−00.500	0.22	0.29	**0.60**	0.35
7906	G356.007−00.424	**0.73**	0.26	0.15	0.18
7909	G356.304−00.206	0.21	0.48	0.16	**0.58**
7917	G356.430+00.104	**0.52**	0.31	0.14	0.21
7920	G356.482+00.190	**0.55**	**0.58**	0.12	**0.63**
7931	G356.662−00.264	**0.84**	**0.97**	**0.69**	**0.84**
7950	G357.557−00.322	**0.98**	**1.00**	**0.75**	**0.98**
7961	G357.968−00.164	**0.98**	**1.00**	**0.86**	**0.99**
7963	G357.997−00.154	0.46	**0.53**	**0.55**	0.49
7976	G358.389−00.484	**0.98**	**1.00**	**0.91**	**0.91**
7980	G358.461−00.392	**0.97**	**1.00**	**0.82**	**0.98**
7983	G358.513−00.374	**0.57**	0.47	**0.73**	**0.51**
8019	G358.725−00.130	**0.50**	0.29	0.36	0.22
8131	G359.141+00.028	0.43	**0.70**	0.32	**0.79**
8193	G359.372+00.275	**0.56**	0.19	0.18	0.18
8198	G359.384−00.021	0.19	0.15	0.06	**0.87**
8203	G359.418+00.089	0.40	0.18	**0.53**	0.10
8207	G359.424−00.171	0.48	0.11	**0.59**	0.10
8212	G359.444−00.105	**0.70**	**1.00**	**1.00**	**0.99**

Table A1. Continued.

Bolocam catalogue #	BGPS name	Random forests	Logistic regression	LDA	Normalised LDA
8217	G359.470−00.037	**0.65**	0.28	0.32	**0.52**
8220	G359.475+00.009	0.32	0.20	0.32	**0.82**
8222	G359.480−00.151	**0.81**	0.35	0.16	0.20
8226	G359.490−00.035	**0.71**	0.14	0.36	0.09
8230	G359.500−00.141	**0.60**	0.00	0.05	0.01
8243	G359.557−00.095	**0.71**	0.01	0.23	0.09
8245	G359.566−00.161	**0.57**	0.11	0.23	0.05
8247	G359.576+00.001	**0.55**	0.07	0.32	0.03
8248	G359.576−00.209	**0.55**	0.09	0.17	0.02
8252	G359.602−00.221	**0.95**	**0.80**	0.37	**0.70**
8258	G359.617−00.243	**0.89**	**1.00**	**1.00**	**0.99**
8261	G359.636−00.131	**0.65**	0.30	**0.66**	0.30
8263	G359.639+00.017	0.40	0.17	0.06	**0.57**
8288	G359.713+00.045	**0.53**	0.20	0.42	0.11
8289	G359.716−00.375	**0.82**	**0.74**	**0.54**	**0.74**
8299	G359.752+00.037	**0.51**	0.07	0.36	0.03
8319	G359.822+00.029	0.49	0.05	**0.51**	0.12
8329	G359.864+00.019	**0.53**	0.14	0.35	0.22
8332	G359.867−00.085	**0.89**	0.16	0.25	**0.95**
8335	G359.891−00.071	**0.80**	**0.98**	**1.00**	**0.94**
8337	G359.900+00.015	**0.70**	0.15	0.23	0.11
8338	G359.906+00.041	**0.73**	0.43	0.15	0.45
8342	G359.912−00.305	0.10	0.32	0.49	**0.56**
8345	G359.920+00.025	**0.75**	0.32	0.13	0.16
8353	G359.944+00.171	**0.87**	**0.82**	0.49	**0.77**
8354	G359.946+00.153	**0.94**	**1.00**	0.39	**0.95**
8355	G359.946−00.047	**0.91**	**1.00**	**1.00**	**1.00**
8356	G359.947+00.023	**0.78**	0.14	0.45	0.08
8361	G359.971−00.459	**0.59**	**0.99**	**0.83**	**0.97**
8366	G359.978−00.071	**0.85**	0.00	**1.00**	**0.84**
8367	G359.985+00.023	**0.65**	0.11	0.43	0.10
8370	G359.994+00.107	**0.60**	0.11	0.26	0.04