The impact of training data characteristics on ensemble classification of land cover

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

Andrew Mellor

BSc (Hons), Aberystwyth University,
MSc Applied Science, RMIT University

School of Science

College of Science, Engineering and Health

RMIT University

July 2017

This dissertation is affectionately dedicated to Violet
(age 5 weeks)

Abstract

Supervised classification of remote sensing imagery has long been recognised as an essential technology for large area land cover mapping. Remote sensing derived land cover and forest classification maps are important sources of information for understanding environmental processes and informing natural resource management decision making. In recent years, the supervised transformation of remote sensing data into thematic products has been advanced through the introduction and development of machine learning classification techniques. Applied to a variety of science and engineering problems over the past twenty years (Lary et al., 2016), machine learning provides greater accuracy and efficiency than traditional parametric classifiers, capable of dealing with large data volumes across complex measurement spaces. The Random forest (RF) classifier in particular, has become popular in the remote sensing community, with a range of commonly cited advantages, including its low parameterisation requirements, excellent classification results and ability to handle noisy observation data and outliers, in a complex measurement space and small training data relative to the study area size.

In the context of large area land cover classification for forest cover, using multisource remote sensing and geospatial data, this research sets out to examine proposed advantages of the RF classifier - insensitivity to training data noise (mislabelling) and handling training data class imbalance. Through margin theory, the research also investigates the utility of ensemble learning – in which multiple base classifiers are combined to reduce generalisation error in classification – as a means of designing more efficient classifiers, improving classification performance, and reducing reference (training and test) data redundancy. The first part of the thesis (chapters 2 and 3) introduces the experimental setting and data used in the research, including a description (in chapter 2) of the sampling framework for the reference data used in classification experiments that follow. Chapter 3 evaluates the performance of the RF classifier applied across 7.2 million hectares of public land study area in Victoria, Australia. This chapter describes an open-source framework for deploying the RF classifier over large areas and processing significant volumes of multi-source remote sensing and ancillary spatial data.

The second part of this thesis (research chapters 4 through 6) examines the effect of training data characteristics (class imbalance and mislabelling) on the performance of RF, and explores the application of the ensemble margin, as a means of both examining RF classification performance, and informing training data sampling to improve classification accuracy. Results of binary and multiclass experiments described in chapter 4, provide insights into the behaviour of RF, in which training data are not evenly distributed among classes and contain systematically mislabelled instances. Results show that while the error rate of the RF classifier is relatively insensitive to mislabelled training data (in the multiclass experiment, overall 78.3% Kappa with no mislabelled instances to 70.1% with 25% mislabelling in each class), the level of associated confidence falls at a faster rate than overall accuracy with increasing rates of mislabelled training data. This study section also demonstrates that imbalanced training data can be introduced to reduce error in classes that are most difficult to classify.

The relationship between per-class and overall classification performance and the diversity of members in a RF ensemble classifier, is explored through experiments presented in chapter 5. This research examines ways of targeting particular training data samples to induce RF ensemble diversity and improve per-class and overall classification performance and efficiency. Through use of the ensemble margin, this study offers insights into the trade-off between ensemble classification accuracy and diversity. The research shows that boosting diversity among RF ensemble members, by emphasising the contribution of lower margin training instances used in the learning process, is an effective means of improving classification performance, particularly for more difficult or rarer classes, and is a way of reducing information redundancy and improving the efficiency of classification problems.

Research chapter 6 looks at the application of the RF classifier for calculating Landscape Pattern Indices (LPIs) from classification prediction maps, and examines the sensitivity of these indices to training data characteristics and sampling based on the ensemble margin. This research reveals a range of commonly used LPIs to have significant sensitivity to training data mislabelling in RF classification, as well as margin-based training data sampling.

In conclusion, this thesis examines proposed advantages of the popular machine learning classifier, Random forests - the relative insensitivity to training data noise (mislabelling) and its ability to handle class imbalance. This research also explores the utility of the ensemble margin for designing more efficient classifiers, measuring and improving classification performance, and designing ensemble classification systems which use reference data more efficiently and effectively, with less data redundancy. These findings have practical applications and implications for large area land cover classification, for which the generation of high quality reference data is often a time consuming, subjective and expensive exercise.

Declaration

I certify that except where due acknowledgement has been made, the work is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; the content of the thesis/project is the result of work which has been carried out since the official commencement date of the approved research program; any editorial work, paid or unpaid, carried out by a third party is acknowledged; and, ethics procedures and guidelines have been followed. I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Andrew Mellor

7 July 2017

Acknowledgements

I would first like to thank and acknowledge my panel of supervisors Prof. Simon Jones, Dr Andrew Haywood and Assoc. Prof. Chris Bellman for their sage advice and guidance over the course of my PhD. I would also like to thank my examiners and those who peer-reviewed published chapters of this dissertation, for their valuable comments and feedback.

I would particularly like to recognise Dr Andrew Haywood, whose ideas, advice and technical guidance and support have helped shape my research and whose wise counsel has kept me going on this journey. I also owe a great deal to my collaborator and co-author Prof. Samia Boukir for her expertise, passion and interest in my research and for helping bridge our respective research disciplines of computer science and remote sensing.

I would like to extend my gratitude to my RMIT friends and colleagues (past and present) including Assoc. Prof Alex Lechner, Dr Phil Wilkes, Dr Lola Suarez, Dr Mariela Soto-Berelov, Dr Will Woodgate, Laurie Buxton, Tapasya Arya and Dr Elizabeth Clarke. Thanks also to Neil Flood and his colleagues at the University of Queensland and Queensland Government, for their technical support in my research. I would also like to acknowledge the Victorian Department of Environment, Land, Water and Planning for providing me access to spatial data used in this research.

I have greatly appreciated the constant support of my friends and family during my long PhD journey and the keen interest they have shown in my research. Thank you Ange for your support. And finally, Sim and Hazel - I could not have done it without you both - thank you for your unwavering love, support and patience.

Thank you, everyone.

Contents

Abstract iii

Declaration vi

Acknowledgements vii

Contents viii

List of Figures xi

List of Tables xiii

Chapter 1. Introduction 1

1.1. Background: Large area land cover mapping 2

1.2. Remote sensing for large area land cover classification 3

1.3. Machine Learning for remote sensing land cover classification 4

1.4. Random forests 5

1.5. RF classification reference data 8

1.6. Research aims and experimental setting 12

1.7. Research Questions 13

1.8. Thesis structure 13

Chapter 2. Experimental Setting and Sampling Design 15

2.1. Experimental setting 16

2.2. Victorian Forest Monitoring Program 16

2.3. Design-based sampling 16

2.4. VFMP Sampling Design 17

2.5. Target population 18

2.6. Sampling Stratification 18

2.7. Sampling 21

2.8. Summary 22

Chapter 3. The Performance of Random Forests in an Operational Setting for Large Area Sclerophyll Forest Classification 24

3.1. Introduction 25

3.2. Random Forests 29

3.3. Open-Source Software 30

3.4. Methods 31

3.5. Random Forest Model 37

3.6. Results and Discussion 39

3.7. Conclusions 44

Chapter 4. Exploring issues of training data imbalance and mislabelling on random forest performance for large area land cover classification using the ensemble margin 46

4.1. Introduction 47

4.2. Random Forests 50

4.3. Ensemble Margin 51

4.4. Study Site and Data 52

4.5. Methods 54

4.6. Results and Discussion 61

4.7. Conclusion 82

Chapter 5. Exploring Diversity in Ensemble Classification: Applications in Large Area Land Cover Mapping 83

5.1. Introduction 84

5.2. Random Forests 85

5.3. Ensemble Margin 86

5.4. Ensemble diversity 87

5.5. Study Area and Data 88

5.6. Experiments 92

5.7. Results and Discussion 94

5.8. Conclusion 105

Chapter 6. Sensitivity of forest Landscape Pattern Indices to training data characteristics in the Random forest classifier 106

6.1. Introduction 107

6.2. Study Areas 109

6.3. Data 110

6.4. Random forest 112

6.5. Ensemble margin 112

6.6. Landscape Pattern Indices 113

6.7. Experiment 1: Margin-based training data sampling 114

6.8. Experiment 2: Training data mislabeling 114

6.9. Analysis of sensitivity of experiments 115

6.10. Results and discussion 115

6.11. Conclusion 125

Chapter 7. Thesis Synthesis and Conclusions 126

7.1. Research Questions 127

7.2. Summary 129

7.3. Future research 131

References 132

List of Figures

Figure 1-1 Random forest classifier training phase, adapted from Parnell et al. (2011) 7
Figure 1-2 Random forest classifier classification phase, adapted from (Nguyen et al., 2013) 7
Figure 1-3 The number of near-polar orbiting, land imaging civilian satellites operational as of 1 August 1972 to 2013 (Belward and Skøien, 2015). 11
Figure 2-1 Location of sampling units (plots) across Victoria's public land Forest Monitoring Program 20
Figure 2-2 VFMP sampling units by IBRA Bioregion 20
Figure 2-3 Primary components (field plot and aerial photoplot) of the VFMP sampling unit 22
Figure 3-1 Australian forest structural definitions (Australian Surveying and Land Information Group, 1990). 26
Figure 3-2 Victorian Interim Biogeographic Regionalisation for Australia (IBRA Bioregions) and aerial photographic interpretation (API) land cover maps (1:25,000) 34
Figure 3-3 Implemented Random Forests model forest probability map (a) inset forest probability map (0–100); (b) final forest classification, based on a binary threshold. 39
Figure 3-4 Random Forests predictor variable importance measures. 42
Figure 4-1 Effect of binary class imbalance on overall classification accuracy 64
Figure 4-2 Effect of binary class imbalance on mean margin 65
Figure 4-3 Binary classification unsupervised margin cumulative frequency distribution curve, comparing correctly and misclassified instance confidence, for optimal and critical training sizes. 66
Figure 4-4 Effect of multiclass imbalance on overall multiclass classification accuracy 67
Figure 4-5 Effect of multiclass imbalance on mean margin 68
Figure 4-6 Unsupervised margin cumulative frequency distribution curves associated with correctly and misclassified instances, comparing balanced versus 50% increase/decrease open/closed 71
Figure 4-7 Unsupervised margin cumulative frequency distribution curves associated with correctly and misclassified instances, comparing balanced versus 90% increase/decrease open/closed 72
Figure 4-8 Unsupervised margin cumulative frequency distribution curves, comparing balanced and ratio-imbalanced (10 open: 90 closed) for optimal cases 73
Figure 4-9 Unsupervised margin cumulative frequency distribution curves, comparing balanced and ratio-imbalanced (10 open: 90 closed) for critical cases 74
Figure 4-10 Effect of class mislabelling on binary classification overall accuracy 76
Figure 4-11 Effect of class mislabelling on binary classification mean margin 76
Figure 4-12 Effect of class mislabelling on multiclass classification overall accuracy 78
Figure 4-13 Effect of class mislabelling on multiclass classification mean margin 78
Figure 5-1 Study area map: Victorian Interim Biogeographic Regionalisation for Australia (IBRA Bioregions) and Aerial Photographic Interpretation (API) land cover maps. 90
Figure 5-2 Aerial photography examples of forest canopy cover used in the multiclass classification (a) Woodland, 20-50% canopy cover; b) Open, 51-80% canopy cover; c) Closed, >80% canopy cover; d) Shrub (land cover dominated by woody vegetation shrub species, up to 2 m in height). Scale various around 1:25,000 91
Figure 5-3 Flow chart illustrating training margins experiment (2) 93
Figure 5-4 Flow chart illustrating minimum node size experiment (3) 94
Figure 5-5 Ensemble and mean base classifier accuracies, mean margin and KW diversity plotted against mtry 96
Figure 5-6 Mean tree accuracy as a function of training set size by lowest and highest unsupervised margins, and random sampling 97
Figure 5-7 Ensemble accuracy as a function of training set size by lowest and highest unsupervised margins, and random sampling 98

Figure 5-8 Ensemble KW diversity as a function of training set size by lowest and highest unsupervised margins, and random sampling 99
Figure 5-9 Ensemble accuracy for the open canopy class as a function of training set size by lowest and highest unsupervised margins, and random sampling 102
Figure 5-10 Ensemble accuracy for the closed canopy class as a function of training set size by lowest and highest unsupervised margins, and random sampling 102
Figure 5-11 Proportion of training samples by class and lowest unsupervised margins by percentile 103
Figure 5-12 Ensemble and mean base classifier accuracies and KW diversity as a function of minimum node size 104
Figure 6-1 Study areas map 110
Figure 6-2 Scatter plot showing curve linear trend between Number of Forest Patches and training data sampling margin percentile (Naringal) 118
Figure 6-3 Scatter plot showing curve linear trend between Edge Density and training data sampling margin percentile (Naringal) 118
Figure 6-4 Scatter plot showing curve linear trend between overall model accuracy and training data sampling margin percentile 119
Figure 6-5 Scatter plot showing no significant relationship between Total area of forest and training data sampling margin percentile (Naringal) 120
Figure 6-6 Naringal forest extent map from classification training data sampled from the 40th percentile margin values 121
Figure 6-7 Naringal forest extent map from classification training data sampled from the 90th percentile margin values 121
Figure 6-8 Scatter plot showing curve linear relationship between the number the forest patches and training data sampling margin percentile (Newstead). 123
Figure 6-9 Newstead forest extent map from classification training data sampled from the 90th percentile margin values 124

List of Tables

Table 2-1 Victorian Forest Monitoring Program sample points by stratum, adapted from (Haywood et al., 2016, 2017) 19
Table 3-1 Random Forests (RF) predictor variables 36
Table 3-2 Random Forests accuracy assessment. CI, confidence interval; OOB, out-of-bag. 40
Table 4-1 Optimal and critical training and test set sizes used for binary and multiclass experiments 55
Table 4-2 Training set sizes for each class for multiclass imbalance (experiments 2 and 3) 56
Table 4-3 RF model performance results for binary classification imbalance (experiment 1) 63
Table 4-4 Binary imbalance confusion matrices and margin-weighted confusion matrices for evenly balanced and imbalanced training data in optimal and critical cases 63
Table 4-5 RF model performance results for optimal and critical multiclass classification imbalance experiments 68
Table 4-6 Multiclass imbalance confusion matrices and margin-weighted confusion matrices for optimal case (balanced and 50% imbalanced) 70
Table 4-7 RF model performance results for optimal and critical binary class mislabelling experiments ... 75
Table 4-8 RF model performance results for optimal and critical size multiclass class mislabelling experiments 79
Table 4-9 Multiclass mislabelling confusion matrices and margin weighted confusion matrices for optimal case 79
Table 4-10 Multiclass mislabelling confusion matrices and margin weighted confusion matrices for critical case 80
Table 5-1 Mean tree, ensemble accuracies (%) and Kappa statistic results for the number of predictor variables experiment 95
Table 5-2 Mean tree and ensemble accuracies (%), and Kappa statistic results for the training margin experiments 99
Table 5-3 Mean tree and ensemble accuracies (%) and Kappa statistic results for the minimum node size experiment 104
Table 6-1 Description of Landscape Pattern Indices (LPIs) 113
Table 6-2 Nature of the trend between margin-based training selection (40th to 90th percentile and random sampling) and LPIs for the Naringal Study area. Curve Linear (CL), Linear (L) or Not-significant (NS). 116
Table 6-3 Nature of the trend between margin-based training selection (40th to 90th percentiles and random sampling) and LPIs for the Newstead Study area. Curve Linear (CL), Linear (L) or Not-significant (NS). 116
Table 6-4 Nature of the trend between mislabeled training data (from 0% up to 30%) and LPIs for the Naringal Study area. Curve Linear (CL), Linear (L) or Not-significant (NS). 124
Table 6-5 Nature of the trend between mislabeled training data (from 0% up to 30%) and LPIs for the Newstead Study area. Curve Linear (CL), Linear (L) or Not-significant (NS). 125

Chapter 1. **Introduction**

1.1. Background: Large area land cover mapping

Timely and accurate large area land cover maps provide critical information to meet a range of environmental, social and economic needs. Such maps are essential inputs to a range of scientific applications, a source of input parameters for models and provide a basis of policy analysis (Wulder et al., 2008). Maps at a range of global, regional, national and sub-national scales, which characterise land cover and support land cover change assessment, support the needs of natural resource managers, scientists, policy makers and researchers (Vogelmann et al., 2004; Ståhl et al., 2016). The applications of such maps include assessment of global carbon budgets and climate modelling, assessing food security (Liu et al., 2008), predicting fire behaviour and hydrological modelling. Large area mapping products provide critical inventory data and information for understanding environmental processes and for effective natural resource management, land use planning and decision making (Lowry et al., 2007).

Satellite-based (remote sensing) earth observation has been recognised as an essential technology for large area, contiguous land cover mapping, which allows for frequent re-measurement for monitoring (DeFries and Townshend, 1994; Boyd and Danson, 2005; Hansen and Loveland, 2012; Chen et al., 2015). Remote sensing derived vegetation and forest maps (and forest cover change products) in particular, are important for understanding the spatial configuration and fragmentation of forest cover (Riitters et al., 2012), modelling forest productivity (Tramontana et al., 2015), invasive species and forest health (Coops et al., 2010) and locating priority areas for biodiversity conservation.

Remote sensing derived forest cover maps and monitoring systems are an important part of many national and regional forest inventory programs - used as a surrogate for field-based observations, to improve the precision of statistical estimates derived from field (plot) measurements and for creating spatially explicit forest cover maps (Deppe, 1998; McRoberts et al., 2005; McRoberts and Tomppo, 2007; Tomppo et al., 2010; Haywood et al., 2016). Forest extent is an indicator under the Montreal Process' seven criteria used to characterise sustainable forest management (Howell et

al., 2008) to which twelve countries are signatories, together representing about 60 per cent of the world's forests and 90 per cent of the world's temperate and boreal forests (Montréal Process Working Group, 2015).

1.2. Remote sensing for large area land cover classification

Remote sensing classification - the transformation of image data into thematic map products has been a fundamental aspect of remote sensing since multi-spectral imagery first became available in the early 1970s (Wilkinson, 2005). Supervised classification in particular, is one of the most common forms of analysis undertaken with remote sensing data (Foody and Mathur, 2004). Supervised remote sensing image classification is broadly defined as the guided categorisation of pixels in an image (or remotely sensed data), to generate a particular set of labels of land cover themes (Lillesand and Kiefer, 1994). A review of image classification methods by Lu & Weng (2007), describes the complexity of this classification process, which requires many factors to be considered. These range from the determination of a suitable classification system, the selection of suitable training samples and image processing feature extraction, to post classification processing and accuracy assessment.

A review of remote sensing classification experiments by Wilkinson (2005), identified advances in three main areas of satellite image classification:

1. The development of particular components of classification algorithms - including training strategies;
2. Augmentation of classification algorithms through novel systems-level approaches;
3. The use of multiple types of ancillary data (including numerical and categorical data).

This fifteen year review (published in 2005) however, found as a whole, no significant upward trend in classification results across the hundreds of experiments reviewed (Wilkinson, 2005).

1.3. Machine Learning for remote sensing land cover classification

Machine learning (ML) techniques – the advanced application of statistics to learning for identifying patterns in data and then making predictions from those patterns – have been used in a variety of science and engineering problems for nearly twenty years (Lary et al., 2016) and over the past decade have become increasingly popular techniques for remote sensing classification (e.g. Foody and Cutler, 2006; Foody et al., 2016; Ghimire et al., 2012; Graves et al., 2016; Rodriguez-Galiano et al., 2012; Rogan et al., 2008). Despite criticism directed at many ML techniques, considered 'black-boxes' which are unable to generate practical prediction equations (Lary et al., 2016), ML algorithms have proved to be more accurate and efficient techniques over traditional parametric approaches, particularly when dealing with large volumes of data across complex measurement spaces (Foody et al., 1995; Rogan et al., 2008).

Unlike more traditional parametric classifiers, non-parametric ML algorithms make no assumptions as to the frequency distribution of input data. ML techniques do require prior knowledge about the nature of the relationships between the data (Lary et al., 2016). Traditional parametric techniques (such as Maximum Likelihood Classification) assume a normal distribution of data and as such, are limited in their application to multi-modal input data (Belgiu and Drăguţ, 2016). With respect to remote sensing data, which rarely have normal distributions, simple classifiers are also constrained in their application to dealing with the complex interactions between scene complexity, scale and aggregation (Marceau et al., 1994). Indeed, the application of traditional remote sensing classifiers are limited in heterogeneous landscapes which are characterised by land cover classes which are difficult to discriminate because of both low inter-class separability, as well as high intra-class variability (Ghimire et al., 2012). Other challenges include the complexity of measurement space and error and variability in calibration (reference) data (DeFries and Cheung-Wai Chan, 2000). Moisture, elevation and temperature (environmental) gradients and topographic heterogeneity also present challenges for image classification (Ghimire et al., 2012).

ML algorithms applied in remote sensing classification include Artificial Neural Network (ANN) (Foody and Arora, 1997; Yuan et al., 2009), deep learning neural

networks (Yu et al., 2017), Adaboost (Chan and Paelinckx, 2008; Haywood and Stone, 2011), Classification and Regression Tree (CART) (Lawrence and Wright, 2001). In recent years however, Support Vector Machines (SVM) and Random forests (RF) have stood out as the most popular ML classification algorithms used in the field of remote sensing. A Scopus database search across title, abstract and keywords "SVM" AND "Remote sensing" returned the highest number of publications, with an yearly average of 142 between 2010 and 2015. Over the same period, a search of "Random forests" AND "remote sensing" showed the highest annual increase in publications in remote sensing, with an annual average increase of 33% (compared to 27% for SVM). Moreover, across all fields (i.e. constraining the search terms to the algorithm name only), since 2010, the number of publications based on the search "Random forests" have increased on average 22% each year.

1.4. Random forests

Random forests (RF) (Breiman, 2001) is an ensemble machine learning technique that combines a collection of decision trees (created using random bootstrap samples of training data), and determines an output class through modal vote (classification) or mean prediction (regression) of the individual trees. Building on research by Amit & Geman (1997) and Ho (1998), Breiman (2001) developed Random forests, defining the classifier as consisting of a collection (or ensemble) of tree structured classifiers

$$\{h(\boldsymbol{x}, \Theta_{k),} k = 1, \ldots\}$$

where Θ_k are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input.

Individual decision trees in a random forest ensemble are constructed by partitioning a subset training data (bagging sample) at each decision tree node, into increasingly homogeneous subsets, using randomly drawn predictor variables. The node-splitting predictor variable selected from the variable subset is one which results in the greatest increase in training data purity (variance or Gini) before and after the tree node split (Cutler et al., 2007). Purity here is defined as the relative homogeneity of training data in each sub-node after node splitting. This decision tree construction continues until there are no further gains in training data purity. Two key model

parameters need to be defined in training the random forest classifier (following notation in the *randomForest* library (Liaw and Wiener, 2002) available in statistical software package R (R Core Team, 2013).

1. The number of trees generated in the random forest ensemble (ntree)
2. The number of randomly selected predictor (or input) variables used at each decision tree split (mtry) - of this predictor variable subset, that which forms the best split is selected.

Figure 1-1 and Figure 1-2 illustrate the training and classification phases of the random forest classifier.

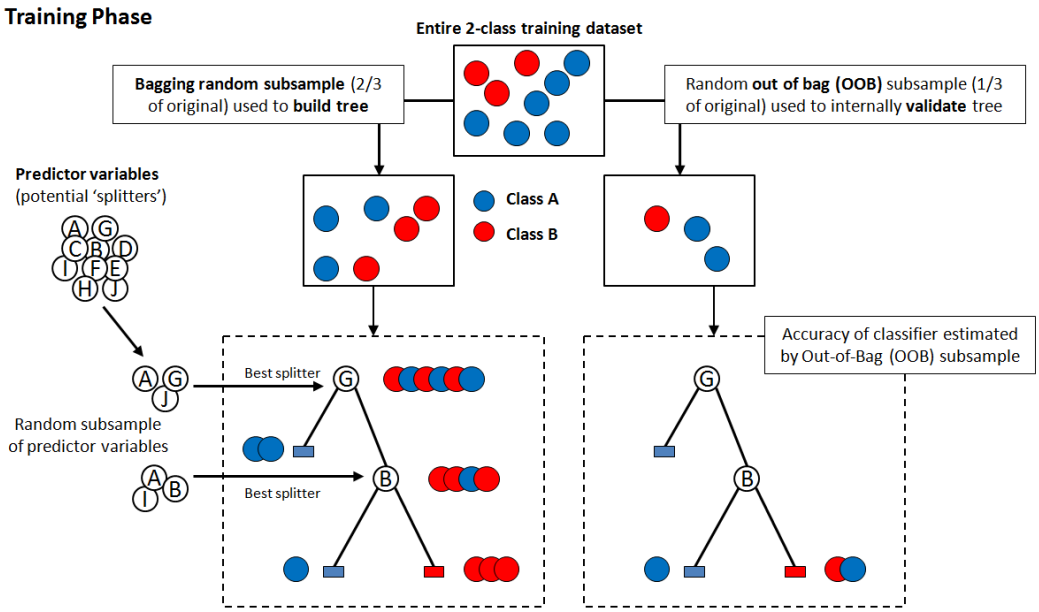


Figure 1-1 Random forest classifier training phase, adapted from Parnell et al. (2011)

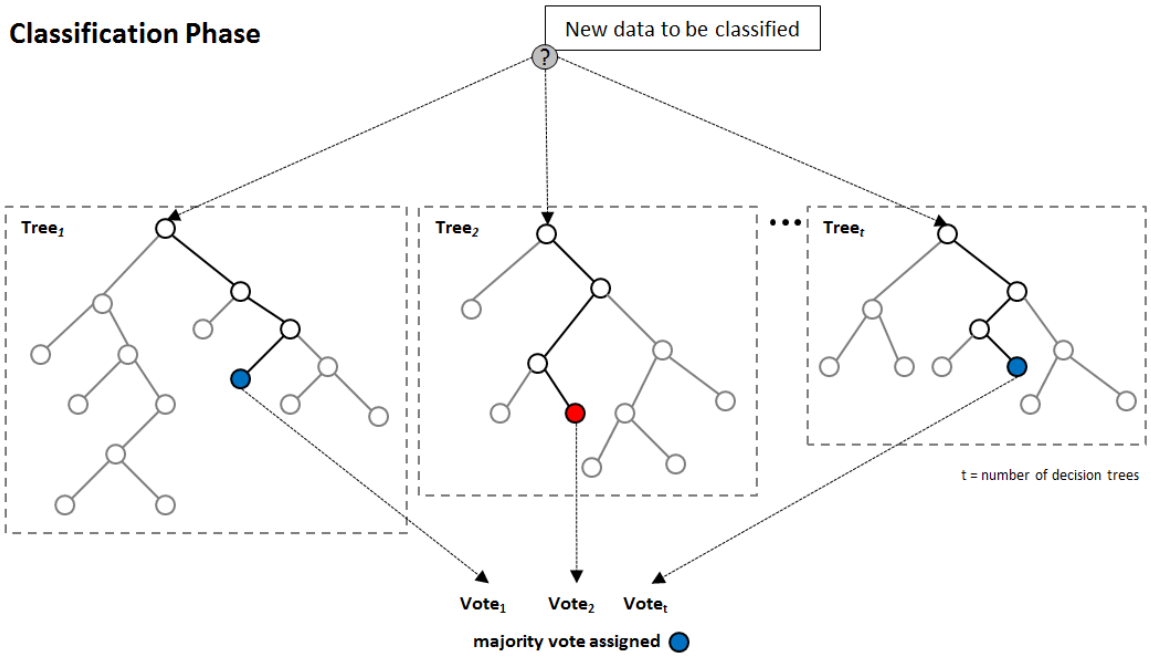


Figure 1-2 Random forest classifier classification phase, adapted from (Nguyen et al., 2013)

Advantages of RF over other machine learning and traditional classifiers have been widely cited in the literature. Chiefly among its attributes are the excellent classification results, efficiency and processing speed (Pal, 2005; Du et al., 2015a; Chutia et al., 2016). Compared to other ML algorithms (such as Boosting and

Support Vector Machine), RF does not require a great deal of parameter adjustment and fine-tuning, with default parameterization often leading to excellent performance (Breiman, 2001; Svetnik et al., 2003; Statnikov et al., 2008) - this makes RF accessible, with good ease of use. Other cited advantages that demonstrate its performance and versatility include its applicability to both binary and multiclass prediction problems (Huang and Boutros, 2016); its handling of thousands of input variables (including a mixture of both categorical and continuous data), and providing estimates of their relative importance in the classification process; its ability to handle noisy observation data and outliers, in a complex measurement space and small training data relative to the study area size (DeFries and Cheung-Wai Chan, 2000; Rogan et al., 2008; Rodriguez-Galiano et al., 2012; Pelletier et al., 2017) and its ability to characterize complex variable interactions (Cutler et al., 2007). RF also demonstrates good predictive performance in applications with more variables than sample data (Huang and Boutros, 2016) and has been argued to not overfit (Peters et al., 2009). The RF algorithm grows an ensemble (forest) of decision trees which have high variance and low bias (Belgiu and Drăguț, 2016).

The RF algorithm can handle diverse multisource remote sensing and geographic data (e.g. soil and terrain variables), making it well-suited to land cover classification (Corcoran et al., 2013; Inglada et al., 2017). Coupled with another of its advantages – the ability to produce variable importance measures, which aid interpretation of the classification model – RF can be used to evaluate the contribution and influence of data sources, for both optimising the classifier and interpreting results (which is typically more challenging in ensemble classification compared to an individual classification tree (Strobl et al., 2007).

1.5. RF classification reference data

The RF classifier has been shown to perform better with large numbers of training samples (Deng and Wu, 2013; Du et al., 2015b). Moreover, van der Ploeg et al. (2014) compared the performance of different machine learning techniques (including SVM and RF) for binary problem solving in relation to the effective sample size (or 'data hungriness'), and concluded that far more events per variable (10 times as many in this study) were needed to achieve stable model performance

(Area Under Curve) compared to classical techniques such as linear regression. Indeed, in the context of this medical study, the authors proposed that such "modern modelling techniques should only be considered....if very large data sets with many events are available" (van der Ploeg et al., 2014). These findings are consistent with earlier research (Selker et al., 1995), which found ML algorithms' ultimate limitations were associated with a "data barrier" (the availability of the information in data).

In an experimental study using data from various application domains, Dietterich (2000) established that boosting is more accurate than bagging. Boosting approaches have been shown to reduce classification variance and bias (Gislason et al., 2006; Ghimire et al., 2012). However, they require large computational resources, overfit if there are insufficient training samples, and are sensitive to any outliers present in the training samples. Other studies have also highlighted the sensitivity of the RF classifier to spatial auto-correlation of training data (Colditz, 2015; Millard and Richardson, 2015), as well as the proportion of different classes within training samples (Dalponte et al., 2013) – highlighting the importance of reference data given its cost and resource requirements.

In the context of large area supervised land cover classification using Earth observation data, the generation of reference data (hereafter used to describe the combination of training and validation or test data) whether through ground-based or sampled from high spatial resolution imagery, is an expensive and time consuming process (Ghimire et al., 2012; Gomez et al., 2016) and the quality of reference data can substantially affect the quality of derived land cover maps (Foody et al., 2016). Indeed, labelling reference data samples is prone to error and can result in poor classification performance and bias (Bradley and Friedl, 1996; Pal and Mather, 2006). Moreover, where ground truth data is assumed to be accurate, but does in fact contain errors, the classification algorithm can be wrongly supposed to be the source of inaccuracy rather than the training data (Carlotto, 2009).

Three developments are facilitating the take up and ease-of-use of modern machine learning algorithms, such as RF, for large area land cover classification problems.

1. Access to cloud computing

Cloud computing - the practice of using a network of internet hosted, remotely accessed servers to store, manage and process data, provides significant opportunities to address the challenge of large scale data-intensive remote sensing applications (Sugumaran et al., 2015). Increasing spatial, temporal, spectral and radiometric remote sensing data resolutions, across a range of platforms, coupled with access to data processing algorithms, and rapidly increasing internet data access and speed, is a technological nexus - one that can be referred to as big data (Sugumaran et al. 2015). Kumar et al. (2013) defines the questions as no longer "how do we capture imagery?", but rather, "how do we handle the immense volume of imagery we already have and to which we're adding every day?".

Amazon Web Services (a subsidiary of Amazon.com) provides a suite of cloud-computing, storage and analytics services in 13 regions across the world, from 2015 made publically available the entire archive of Landsat 8 scenes. Machine Learning AWS also provides tools to build machine learning models, including data analysis, training and evaluation. Google Earth Engine is a cloud-computing platform for processing satellite imagery and other earth observation data. GEE contains over 200 public datasets, over 5 million images and more than 5 petabytes of data. GEE's suite of tools include a suite of supervised classification algorithms (including Random forest, CART and SVM) and workflow for building, training, applying and assessing classification algorithms (Google Earth Engine Team, 2015).

2. Open Source software

Increasing ease of access to machine learning algorithms like RF via open-source software environments (including through cloud-computing services), allows users to access and readily automate classifiers through a set of adjustable parameters, which makes RF straightforward to apply for relatively inexperienced users (Qi et al., 2006). Several implementations of the RF classifier are now available, including the most popular randomForest (Liaw and Wiener, 2002) on the statistics package R (R Core Team, 2013), as well as implementations in Python, such as *scikit learn Ensemble forest* (scikit-learn developers, 2016) and through the Machine Learning Tool Kit (MILK) (Coelho, 2017) and *Fast random forest* in the WEKA Environment.

3. Remote Sensing Data

Launched in July 1972, Landsat 1 became the first global satellite earth observing mission (Belward and Skøien, 2015). The number of near polar orbiting operational earth observing satellite missions grew rapidly after 1972, to eight in August 1982, twenty a decade later, thirty-nine by August 2002 and eighty-three by 2012. Figure 1-3 (Belward and Skøien, 2015) shows the number of satellites operating by year and illustrates the rapid increase overtime.

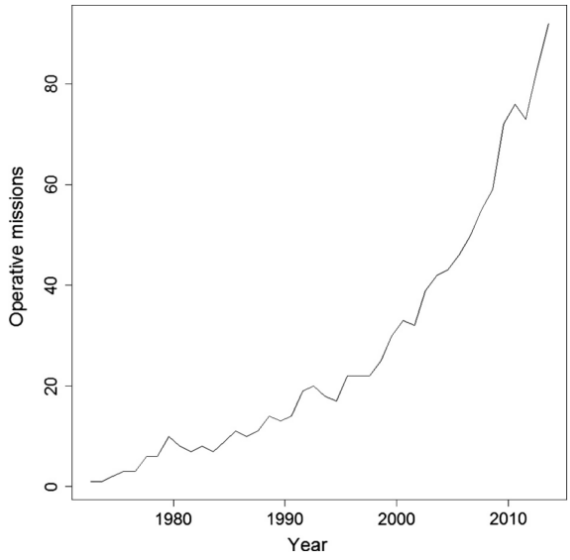


Figure 1-3 The number of near-polar orbiting, land imaging civilian satellites operational as of 1 August 1972 to 2013 (Belward and Skøien, 2015).

Commensurate with the increase in earth observing platforms has been the increase in available remote sensing data. A policy change in 2008 resulted in the all new and archived United States Geological Survey (USGS) held Landsat satellite image data becoming freely available to any user (Wulder et al., 2012). The significance of this policy change cannot be underestimated - as at June 30 2016, over 42 million Landsat scenes have been downloaded by users worldwide (U.S. Geological Survey, 2017). Open data policies, like the Landsat Data Policy (http://landsat.usgs.gov/documents/Landsat_Data_Policy.pdf), have increased the practicality of combining multiple data from multiple sensors and support data assimilation approaches for generating information, which, unlike in the meteorological community, are under-represented in terrestrial remote sensing (Wulder et al., 2012). Wulder et al. (2012) contend that the decision to make Landsat data freely available supports the efforts of international earth observing organisations in encouraging open data standards.

The range of open-access satellite imagery extends to the European Space Agency's Sentinel program (including 10 metre multispectral data) and Synthetic Aperture Radar (European Space Agency, 2016); MODIS (Moderate Resolution Imaging Spectroradiometer) aboard the Terra and Aqua satellites, acquiring data across 36 spectral bands over the entire Earth's surface every 1-2 days (NASA, 2016). Together with Landsat 8, the Sentinel satellite constellations will provide potential for landscape-scale observation data every three to four days (Turner et al., 2015). The combination of Landsat 8 and two Sentinel satellite sensors (2A and 2B) offer a global median average revisit interval of 2.9 days and maximum revisit interval of 7 days (Li and Roy, 2017).

Low-cost and accessible cloud-computing infrastructure, the free availability of open-access versions of a range of popular ML classification algorithms, and open-access policies for moderate resolution multi-spectral remote sensing data and a range of other spatial data, are all factors which promote the uptake of ML classifiers and provide great opportunities for improving the accuracy, currency and quality of large area land cover maps for a range of applications.

1.6. Research aims and experimental setting

In the context of large area classification using multisource remote sensing and geospatial data, the primary aim of this research is to examine two of the proposed advantages for RF described in this introduction - the relative insensitivity to training data noise (mislabelling) and its ability to handle class imbalance. This research will also investigate the utility of ensemble learning (and associated margin theory) – in which multiple base classifiers are combined to reduce generalisation error in classification – to design more efficient classifiers, improve classification performance, to reduce reference data redundancy and design ensemble classification systems which use reference data more efficiently and effectively.

The experimental setting and data used in this research (introduced and described in detail in chapters 2 and 3) provides a unique real-world testing environment through which to explore and apply ML concepts – typically constrained to simulation-based studies in the field of information science – to a large area remote sensing problem, using reference data (stratified, unbiased and proportional to the study area) and an

environment which is both realistic and a representative testing environment to provide insights for classification problems applied in alternative geographic settings where greater reference data typically constraints apply.

1.7. Research Questions

Three research questions are explored in this thesis:

Question 1: How do training data characteristics of class imbalance and class mislabelling affect RF performance?

This question is explored through the application of margin theory, employed as a measure of confidence in classification results, to supplement traditional classification performance measures used in remote sensing classification.

Question 2: What is the relationship between ensemble diversity and classification performance?

This question seeks to examine the degree of influence that ensemble diversity has on classification performance, and how ensemble classifier diversity can be controlled to improve the efficiency and effectiveness of classification training data.

Question 3: What is the relationship between training data characteristics (used to construct RF ensemble classification models) and Landscape Pattern Indices (LPIs) calculated from RF derived prediction maps?

This questions looks at the application of RF classification models to generate LPIs, and examines the sensitivity of these indices to training data characteristics and sampling based on the ensemble margin.

1.8. Thesis structure

This thesis is presented such that each chapter (with the exclusion of the introduction and synthesis) may be read independently. The research chapters match the published (or prepared for publication) versions, with changes only to formatting in order to maintain a consistent style through the thesis. Cited references are compiled into a single bibliography at the end of the thesis.

The thesis comprises seven chapters, of which four are research chapters (three of which have been published in peer-reviewed journals). There is no stand-alone literature review chapter, as these are included in the introduction sections of each research chapter.

Chapter 2 describes the experimental setting for this research - including the sampling framework for the reference (training and test) data used in classification experiments that follow. This chapter summarises the advantages and opportunities afforded by the experimental design to explore the key research questions introduced in Chapter 1. Chapter 3 evaluates the performance of the Random forest (RF) classifier applied across 7.2 million hectares of public land in Victoria, Australia. This chapter describes an open-source framework for deploying the RF classifier over large areas and processing significant volumes of multi-source remote sensing and ancillary spatial data.

Chapter 4 examines the effect of training data characteristics of class imbalance and mislabelling on the performance of Random forests. Through different experiments applied to binary and multiclass problems, this research chapter examines the sensitivity of RF classification performance to training class imbalance and training data mislabelling. Chapter 4 also introduces the ensemble margin, and derived metrics that can be used as ancillary measures of classification performance.

Chapter 5 explores the relationship between per-class and overall classification performance and the diversity of members in a RF ensemble classifier. This chapter brings together the understanding of the ensemble margin developed in Chapter 4, to look at ways to target particular training data samples to induce ensemble diversity and improve per-class and overall classification performance and efficiency.

Chapter 6 explores the application of the RF classifier for deriving landscape pattern indices from classification prediction maps and examines the sensitivity of these indices to training data characteristics and sampling based on the ensemble margin.

Chapter 7 provide a synthesis of the research and discussing the research findings and their implications in the context of recent technology and data developments, which have increased the accessibility of advanced classification algorithms such as RF.

Chapter 2. Experimental Setting and Sampling Design

2.1. Experimental setting

The following chapter describes the experimental context for this research - including the study area and the sampling framework for the reference (training and test) data used in large area land cover classification experiments that follow. This chapter summarises the advantages and opportunities afforded by the experimental design to explore the key research questions introduced in Chapter 1.

2.2. Victorian Forest Monitoring Program

The reference data used in the research experiments described in chapters 3 through 6, is drawn from the Victorian Forest Monitoring Program (VFMP). The VFMP (Haywood et al., 2016; Haywood and Stone, 2017) is a strategic forest inventory established in the State of Victoria in south east Australia. The VFMP combines field measurement plots with remote sensing data across the State's public land forests, the information from which is used to assess Victoria's progress towards achieving sustainable forest management objectives and targets (Haywood et al., 2016). The VFMP and other similar strategic forest inventories have been established in many jurisdictions around the world (e.g. in north America and Scandinavia) – historically with the primary objective of monitoring and assessing forest (i.e. timber) resources. More recently however, there has been a shift in public focus and awareness towards the essential role that forests also play in climate regulation, as a source of biological and genetic diversity, in the storage and maintenance of carbon cycles, and the provision of cultural, tourism and amenity values (Myers, 1996; Boyd and Danson, 2005). The increasing need for consistent data with which to make comparisons between land and forest management regimes or between different jurisdictions is also driving the need to establish and maintain forest data collection systems – which also support national and international forest policy and decision making.

2.3. Design-based sampling

The VFMP uses a design-based sampling framework (also known as a probability-based sampling design) - a classical approach to sampling (Cochran, 1977), for which the objective is to describe the characteristics of a real and explicitly defined population. Such sampling is necessary to address the impracticality of collecting

reference data for a census of an entire region (Stehman, 2000). In design-based frameworks, sampling locations are selected through probability sampling and statistical inference used to, for example, estimate a spatial mean, is based on sampling design (Brus, 2010). Design-based inference typically assumes a finite population of elements to which one or more fixed target quantities are linked (Ståhl et al., 2016). In contrast to design-based sampling, model-based sampling does not have requirements on a method for selecting sampling locations, and typically are selected by purposive (targeted) sampling, for instance on a centred grid (Brus, 2010). Model-based approaches, sometimes characterized as model dependent approaches (Hansen et al., 1983), use predictions based on models and ancillary variables to produce estimates (McRoberts, 2010).

Simple random sampling and systematic sampling are sampling approaches which provide a foundation for most probability or design-based sampling. The VFMP applies stratified random sampling for its design-based approach. In stratified random sampling, the total population is divided into mutually exclusive, non-overlapping strata, from which simple random samples are taken. Each potential sample unit can only be assigned to one stratum and all unit are included. Among the advantages of stratified random sampling include minimizing sample selection bias and reducing over and under-representation of certain population segments.

2.4. VFMP Sampling Design

The VFMP is a plot-based design made up of permanent observational units located on a state-wide grid (Haywood and Stone, 2017). The guiding principle of the VFMP design is the consistency of data collected through monitoring, whereby the same attributes are measured over space and time, with the same standards and in a statistically defensible manner and at an acceptable level of precision. For the VFMP, the desired stratum level target precision (standard error) is 12.5%. The VFMP's sampling framework has the following key elements (which are described in further detail below) (Haywood et al., 2017).

1. **Target population**: the public land estate of Victoria

2. **Stratification**: Two-way stratification of the target population with each stratum adequately sampled for statistical reliability through variable sampling intensity.

3. **Plot design**: comprising two components, a) ground-based - from which a range of direct measurements of forest structure and composition are taken, and b) a remotely sensed photo-plot

2.5. Target population

The target population (study area, or sampling frame) comprises 7.1 million hectares of public land, covering about one third of the state of Victoria, in south East Australia. This includes about 3.9 million hectares of mostly forested parks and conservation reserves – managed primarily for ecosystem and biodiversity conservation, as well as tourism, recreation and cultural and historic values. State forests cover about a further 3.1 million hectares – land management in State forests include the water catchments and water supply, flora and fauna conservation, as well as the provision of timber (The State of Victoria Department of Environment and Primary Industry, 2013). The target population is assumed to consist of an infinite number of points within the public land estate. Chapter 3 includes a more detailed description of the study area climatologically and environmental and topographic characteristics.

2.6. Sampling Stratification

The target population was stratified with respect to two factors, bioregion and tenure. Firstly, the target population was stratified into 11 IBRA (Interim Biogeographic Regionalisation for Australia) Bioregions – these are large and geographically distinct areas of land which share common geology, landform, climatic and ecological characteristics (Cummings and Hardy, 2000). The target population was further stratified into the two major public land tenure (Parks and Reserves, including national, state, and regional parks, and State forest (described above). Figure 2-1 shows the distribution of sampling plots (units) located across Victoria's major public land tenures. Figure 2-2 shows the sampling plots (units) and IBRA Bioregions (the primary stratification unit).

Within each stratum, sample units were placed at the intersections of a grid which utilised the VicGrid coordinate system (DSE, 2000) whose spacing varied between 2 km and 20 km and was selected to produce a per stratum sample size of approximately 30 samples. The target within-stratum sample size of 30 samples was based on the assumption of a coefficient of variation for a quantitative trait measured in the VFMP (such as biomass) of at least 70% and a stratum-level target precision (or standard error) of no more than 12.5% (Haywood et al., 2016). Within a geographically large stratum, sample points are more widely spaced to achieve the optimal and most resource efficient target number of sampling locations, compared to smaller strata. Table 2-1 shows the number and spacing of Victorian strategic forest inventory sample points by stratum. A more detailed description of the VFMP sampling design and its rationale can be found in Haywood et al. (2016). Unlike many other strategic forest inventories - which collect information about the state and dynamics of forests for management planning - the VFMP sampling (from field and remote sensing) deliberately extends to include all land covers types within the public land estate.

Table 2-1 Victorian Forest Monitoring Program sample points by stratum, adapted from (Haywood et al., 2016, 2017)

IBRA Bioregion	Parks and Reserves	Grid spacing (km)	State forest	Grid spacing (km)
		Sample Units		
Australian Alps	36	10	53	8
Flinders	26	4	*	-
Murray-Darling Depression	39	20	28	10
Naracoorte Coastal Plain	42	4	42	4
NSW South Western Slopes	43	4	31	4
Riverina	69	6	8	4
South East Coastal Plain	27	8	25	4
South East Corner	39	10	44	12
South East Highlands	49	12	42	18
Victorian Midlands	35	10	38	8
Victorian Volcanic Plains	30	6	40	2
Total	435		351	

* Flinders Bioregion does not contain any State forest

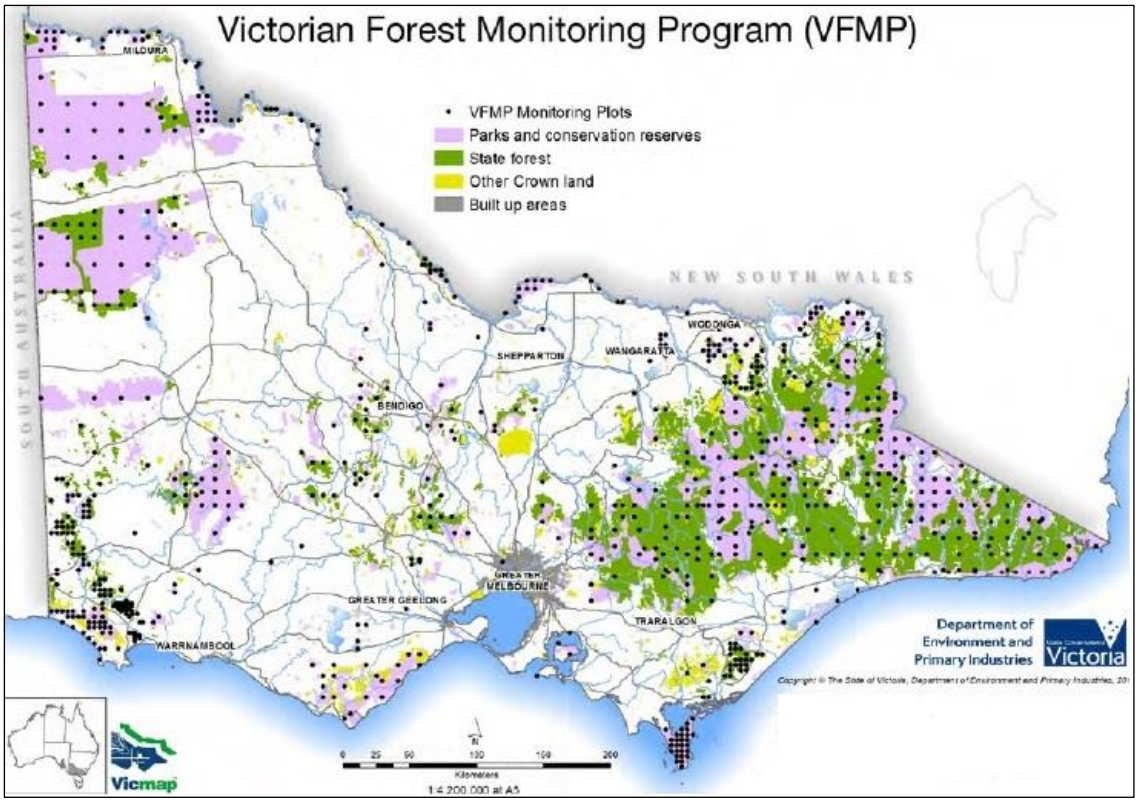


Figure 2-1 Location of sampling units (plots) across Victoria's public land Forest Monitoring Program

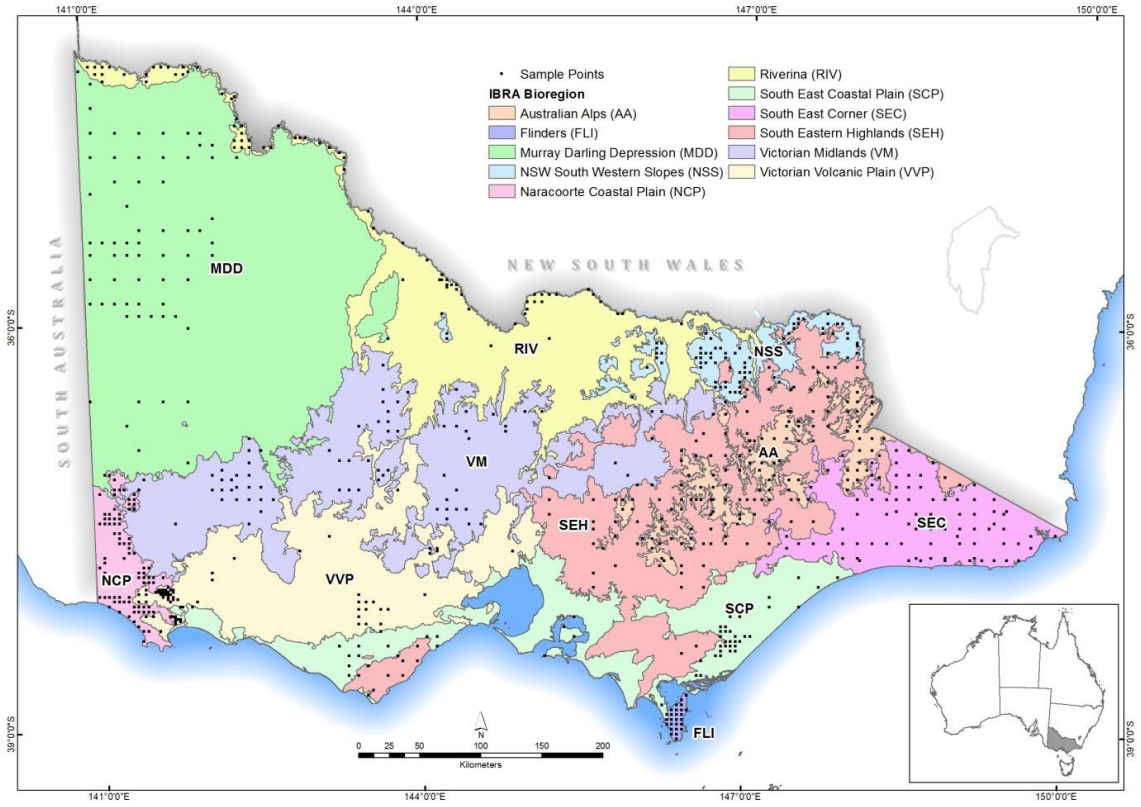


Figure 2-2 VFMP sampling units by IBRA Bioregion

2.7. Sampling

The plot design of each sample point comprises two main components – a multi-staged field-plot and an aerial photoplot. At the field-plot level, 215 variables are measured and assessed, within a 0.04 ha circular plot and soil and vegetation quadrats. These variables include physical and biotic characteristics (such as slope, aspect, topographic position and site disturbance), as well as tree measurements (e.g. species, diameter at breast height over bark, canopy health and cover), coarse woody debris, understory vegetation and groundcover attributes and soil. A detailed description of the field-plot inventory method and attributes measured is available in Haywood et al. (2016).

Above each field-plot point, 2 km x 2 km photoplot sampling units provide the primary source of land cover information for the VFMP inventory and the source of reference data used in the research experiments described in the following chapters. Digital high resolution (30 cm and 50 cm pixels) colour (RGB and Near Infrared) aerial photographs acquired over the period 2006 to 2010 were used to map landcover, following a classification system comprising broad forest type, height and canopy cover classes (Mellor and Haywood, 2010). A detailed description of the land cover mapping method applied to VFMP photoplots and used as the source of reference data in this study, is included in chapter 2 and documented in Farmer et al. (2013).

Figure 2-3 illustrates the primary sampling components of the VFMP ground plot, together with an example land cover photoplot map (above).

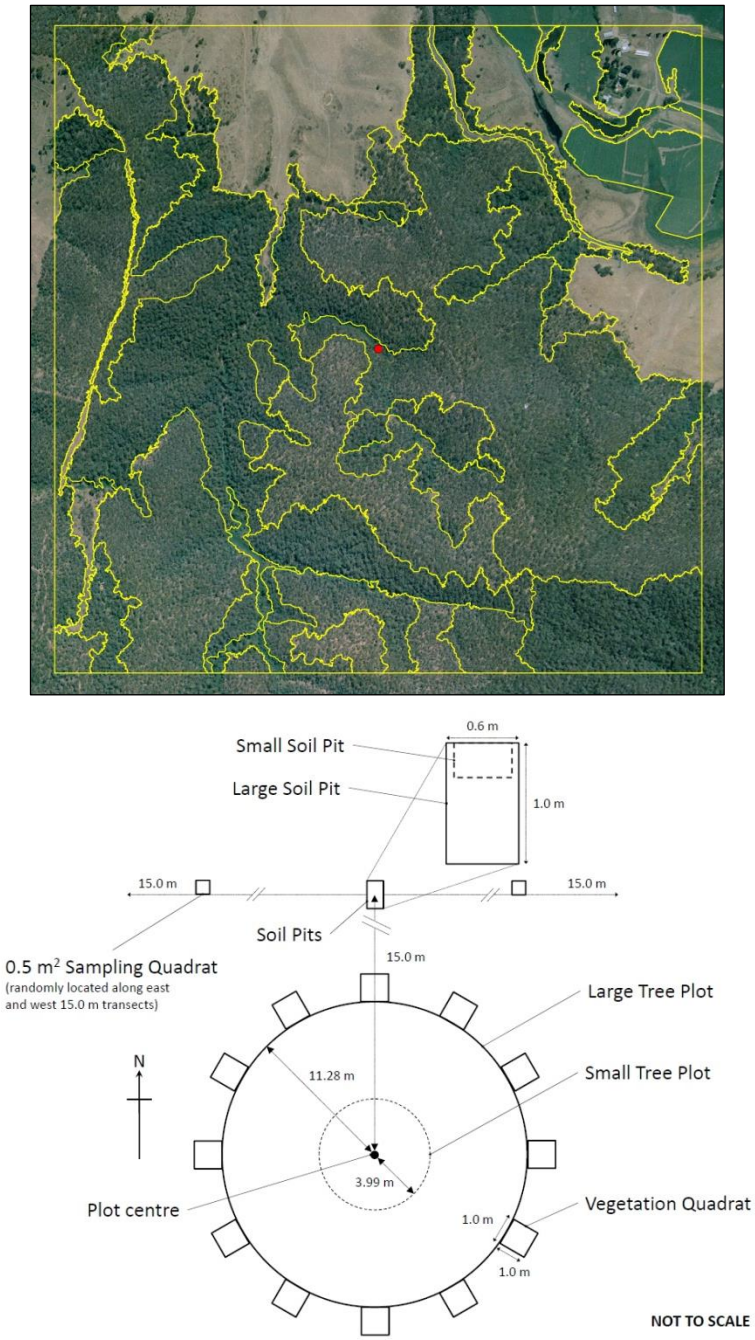


Figure 2-3 Primary components (field plot and aerial photoplot) of the VFMP sampling unit

2.8. Summary

The design-based statistical sampling framework of the VFMP and the nature of the photoplot sampling units from which reference data is collected, afford several advantages which provide a unique opportunity to explore how training data characteristics affect RF performance in this research. For example, the spread of sampling units is comprehensive and their geographic coverage extensive. The

systematic and stratified sampling framework helps ensure that sampling is balanced, unbiased and addresses heterogeneity characteristics of the large and diverse study area. Furthermore, the training data generated at sampling units is temporally consistent.

These training data sampling characteristics are not typical - particularly in large regions or jurisdictions in which areas are inaccessible or suitable high resolution data is scarce. Research findings from this exemplar reference dataset from a real-world experimental environment, might be used to design and parameterise more efficient ML classifiers in other jurisdictions, making more efficient and effective use of training data, which may of poorer quality (e.g. less geographic or class coverage, noisy and mislabelled and collected with temporal variability).

Chapter 3. The Performance of Random Forests in an Operational Setting for Large Area Sclerophyll Forest Classification

Based on the peer-reviewed published article:

Mellor, A., Haywood, A., Stone, C. and Jones, S., 2013. The performance of random forests in an operational setting for large area sclerophyll forest classification. *Remote Sensing*, *5*(6), pp.2838-2856.

3.1. Introduction

Forest extent is a measure commonly assessed in national forest inventories (NFI) (McRoberts, 2010) and, under the Montreal process (Howell et al., 2008), is a specific indicator used for monitoring and reporting sustainable forest management. For natural resource management agencies, current and accurate forest area estimates are critical for effective environmental monitoring. While ground-based (field plot) forest inventories can provide accurate and unbiased forest area estimates, spatially explicit remote sensing-derived forest extent maps can be used to assess the spatial configuration of forest at the landscape scale and used in combination with a high resolution sample (two-staged sampling) to improve forest area estimates (Deppe, 1998).

In Australia, under the Australian National Forest Inventory, forest is defined as "A land area, incorporating all living and non-living components, dominated by trees having usually a single stem and a mature or potentially mature stand height exceeding two metres and with existing or potential crown cover of overstory strata about equal to or greater than 20 percent. This definition includes native forests and plantations and areas of trees that are sometimes described as woodlands" (Department of Agriculture Fisheries and Forestry, 2012). The structural components in this definition encompass a wide range of forest types, from open low sparse canopy woodland to tall dense canopy forests (as illustrated by Figure 3-1, (Australian Surveying and Land Information Group, 1990)).

In Australia (and the state of Victoria, in particular), dry, damp and wet sclerophyll forests and woodlands comprise many of the forested ecosystems. The canopies in these ecosystems are dominated by eucalypt species and are characteristically open with irregular (asymmetrical) crown configurations and low foliage density (Jenkins and Coops, 2011). Canopy foliage is often clumped, leaves tend to concentrate around crown perimeters (Jacobs, 1955) and exhibit an erectophile (vertical) leaf angle distribution. In Victoria, as in much of Australia's forests, there is a high diversity of forest development phases, vertical and horizontal forest structures, topography and soil types (Behn et al., 2001), as well as dynamic phenological processes in understory vegetation (Bhandari, 2011).

These characteristics pose a number of challenges to the use of remote sensing in these environments for classifying and mapping forests. The mid- and under-story components, shadows and background soils all exhibit a strong influence on spectral reflectance characteristics. From a synoptic perspective, forest cover in Victoria can appear indistinguishable from shrub and other low and sparse woody vegetation species. Complexity and background noise in remote sensing signatures from open sclerophyll eucalypt forests is further intensified by the influence of dynamic understory elements and variation in forest structures (Jupp and Walker, 1997). The challenges and complexities associated with forest extent mapping across state and territories in Australia is evidenced by large differences and inconsistencies in forest extent maps and forest area estimates produced by state and federal government agencies and the variability in forest area estimates published in Australia's national five-yearly State of the Forests reports (Montreal Process Implementation Group for Australia, 2008). The processing of large area remote sensing datasets poses a further challenge for state land management agencies.

Source: Australian Land Information Group and JA Carnahan (1990). *Atlas of Australian Resources, Vegetation*. Australian Government Publishing Service, Canberra.

Figure 3-1 Australian forest structural definitions (Australian Surveying and Land Information Group, 1990).

Random Forests (RF) (Breiman, 2001) offers a possible solution to address these large area forest classification challenges, universal across many of Australia's forest ecosystems. Machine learning classifiers, such as RF, are increasingly being used for environmental mapping and modelling applications in fields, such as natural resource management and forestry (Main-Knorn et al., 2011; Clerici et al., 2012; Rodriguez-Galiano et al., 2012). RF is an ensemble decision tree classifier, which combines bootstrap sampling to construct many individual decision trees, from which a final class assignment is determined (Breiman, 2001).

RF can be used to learn complex non-linear relationships, such as those present in variable vertical forest structure and the association of overstory to understorey forest vegetation. RF has been demonstrated to be very effective for accurate land cover mapping across complex and heterogeneous landscapes and to be relatively insensitive to noise (Rodriguez-Galiano et al., 2012), making it suitable for application in complex and dynamic forest environments. As RF does not require normally distributed model training data, its application is appropriate for areas where species distributions of ecological communities follow non-linear patterns across the landscape (Austin and Meyers, 1996) and where complex terrain effects data normality (Khalyani et al., 2012). Other reported benefits of RF include its relative insensitivity to outliers (Breiman, 2001; Cutler et al., 2007), common characteristics of open canopies across large areas of dynamic and highly variable forest ecosystems. Furthermore, the RF classifier runs efficiently on large datasets (Rodriguez-Galiano et al., 2012), making it suitable for regional-scale mapping, comprising millions of hectares.

As only a random subset of variable data is used to construct each decision tree in a random forest classifier ensemble, correlation between decision trees is reduced, thereby improving predictive power and classification accuracy, whilst decreasing the computational complexity of the algorithm. As has been demonstrated in recent studies (Fahsi et al., 2000; Joy et al., 2003; Gislason et al., 2006; Sesnie et al., 2008), RF can incorporate multiple-sources of remote sensing data with ancillary continuous and categorical biophysical spatial data to improve classification performance and discriminate between forest and non-forest.

Moderate resolution multi-spectral imagery, such as Landsat Thematic Mapper (TM)/Enhanced Thematic Mapper (ETM+) has been commonly applied for estimating forest cover (Green and Sussman, 1990; Boyd and Danson, 2005), discrimination of some forest types (Lu et al., 2003), forest cover change detection (Tucker and Townshend, 2000; Rogan, 2002) and for model-based forest area estimation (McRoberts, 2010). Because of the challenges described above, limitations arise in classifying forest extent where different forest structures and composition and land cover types can appear spectrally alike using traditional remote sensing data analysis techniques. Improved forest classification accuracy and forest area estimates have been achieved for large areas using multi-temporal imagery, e.g., MODIS (Wulder et al., 2010; Maselli, 2011). The high temporal resolution of the MODIS sensor can provide valuable information about the phenological variability of different land covers and, as such, help address the challenge of forest canopy-to-understory discrimination in the type of open canopy forest environments described above.

In the context of open-canopy forest extent classification, textural information (spatial variation data derived from optical imagery) can provide additional information to a RF classifier, by differentiating vegetation that appears spectrally similar when integrated into a remote sensing image pixel, but whose spatial patterns differ (Culbert et al., 2009). Recent studies have used satellite image-derived texture indices to improve forest stand classification (Coburn and Roberts, 2004), biomass and carbon estimation (Lu, 2005; Proisy et al., 2007; Eckert, 2012) and forest structure derivation (Kayitakire et al., 2006). In a large heterogeneous landscape RF classification study, Rodríguez-Galiano et al. (2011), increased overall accuracy by 8% (and Kappa by 9%) by including textural information.

The conditional relationships between forest vegetation and biophysical factors can also be used to further improve forest/non-forest discrimination. Species-environment relationships are central to predictive geographical modelling (Guisan and Zimmermann, 2000). Topographic variables (e.g., elevation, slope and aspect) used in combination with spectral data have been demonstrated to enhance forest, habitat and vegetation classification (Fahsi et al., 2000; Joy et al., 2003; Gislason et al., 2006; Sesnie et al., 2008). Bioclimatic maps (e.g., temperature, precipitation) are an additional source of commonly used ancillary classification data.

These maps are typically developed through elevation-sensitive interpolation of climate station data and digital elevation models (Guisan and Zimmermann, 2000), which support the assumption that climate has a major influence on species distribution at broad geographic scales (Beaumont et al., 2005) and that similar compositions of vegetation can be expected to occur at sites with comparable soil, climate and topography (Franklin, 1995). In this paper, we evaluate the operational performance and utility of RF for classifying forest extent across Victoria, Australia, using remote sensing, topographic and climate predictor variables. The originality of this study lies firstly in the scale of the application of the RF algorithm, to construct, evaluate and implement an RF classifier to produce an accurate ~220,000 km^2 land management agency forest map. As far as we know, this scale of RF operation is unique. The second novel aspect to this study is in its application setting, which, to our knowledge, is the first time RF has been used in an operational environment at a regional scale comprising highly diverse and complex Australian forest ecosystems and topography, dominated by open canopy sclerophyll forests and woodland.

While studies on the production of forest and land cover maps derived from RF (or similar) classification techniques using multi-source remote sensing and ancillary data are published routinely in the academic literature, a secondary objective of this paper is to describe a framework for operational implementation of the RF algorithm using open-source software. The framework includes each phase of the RF classification process (from predictor variable pre-processing, through model development and implementation), to support transfer of this technology in an operational land management agency context and make use of the freely available and growing archives of remote sensing and geographic data.

3.2. Random Forests

Random Forests uses bootstrap (a form of sampling with replacement) aggregated sampling (bagging) to construct many individual decision trees from which a final class assignment is determined (Cutler et al., 2007). The RF algorithm constructs each decision tree using a bootstrap sample from available training data, with the remaining assigned as out-of-bag (OOB) samples. At each decision tree node, a random subset of predictor variables are tested to partition the observation data into

increasingly homogeneous subsets. The node-splitting variable selected from the variable subset is that which results in the greatest increase in data purity (variance or Gini) before and after the tree node split (Cutler et al., 2007). Tree building continues until there are no further gains in purity. A response variable can be predicted as an average (continuous variable classification) or model vote (categorical classification) among all decision trees built in the forest. The OOB sample data are used to compute accuracies and error rates averaged over all predictions (Cutler et al., 2007) and estimate variable importance in the classification. The computational complexity of the algorithm is reduced, as only a random subset of variables is used at each node split. This process also reduces correlation between trees, thereby improving both predictive power and classification accuracy. RF includes two methods to estimate the importance of each predictor variable in the model. The mean decrease in accuracy (MDA) importance measure is calculated as the normalised difference between OOB accuracy of the original observations to randomly permuted variables (Cutler et al., 2007). An alternative variable importance measure is calculated by summing all of the decreases in Gini impurity at each tree node split, normalised by the number of trees (Breiman and Cutler, 2001; Calle and Urrea, 2011).

3.3. Open-Source Software

By adopting an open-source framework for spatial data management, processing and analysis, users, such as land management agencies, can benefit from freely available software products and access to source code through which new algorithms can be integrated and manipulated. Stallman (1985) describes the four freedoms of the free and open-source software approach, as freedom to (i) run the program for any purpose, (ii) study how the program works, (iii) redistribute copies and (iv) improve the program and release such improvements to the public (Rocchini et al., 2013).

3.3.1. Geographic Resources Analysis Support System (GRASS)

GRASS (Geographic Resources Analysis Support System) (GRASS Development Team, 2012) is an open-source geographical information system capable of handling raster, topological vector, image processing and graphic data. Released under the GNU General Public License (GPL), GRASS is developed by a multi-national group

of developers and is one of the eight initial software projects of the Open Source Geospatial Foundation. GRASS has a modular structure into which may be plugged new routines programmed in a variety of languages (e.g., Python, C, shell), and there are over 300 modules and more than 100 add-on modules for the creation, manipulation and visualisation of both raster and vector data. The GRASS modules are designed under the UNIX philosophy (*i.e.*, that programs work together and handle text streams) and can be combined using shell scripting to create more complex or specialized modules by a user. GRASS supports an extensive range of raster and vector formats through GDAL/OGR libraries, including OGC-conformal (Open Geospatial Consortium) Simple Features for interoperability with other GIS.

3.3.2. R and Python

R (R Development Core Team, 2011) is an open-source language and software environment commonly used in research fields for statistical computing and graphics. One of the main advantages of R is its object-orientated approach, which allows results of statistical procedures to be stored as objects and used as input in further computations. R is a simple and effective formal complete programming language, and the R environment is, therefore, highly extensible. GRASS and R software can be integrated through the R package, *spgrass* (Bivand, 2007), an interface allowing GRASS GIS functions to be implemented within R code and data to be easily exchanged between the two software packages. Python (Python Software Foundation, 2011) is an object-orientated high-level programming language that is widely used as a scripting language in the spatial analysis environment. Python's popularity has led to the creation of many useful libraries, increasing its flexibility and interoperability, and it has well developed modules for linking with GRASS and R.

3.4. Methods

3.4.1. Study Area

The study area comprises approximately 7.2 million hectares of public land forests and parks tenure (hereafter, referred to as public land forests) in the state of Victoria, in southeast Australia. This area includes 4 million ha of national parks and

conservation reserves, managed primarily for ecosystem and biodiversity protection, tourism and recreation. The remaining 3.2 million ha are multiple-use state forest tenure, which include the provision of timber and non-timber forest products. Bounding extents of Victoria are north 141°47'36" E 33°58'54"S, east 149°58'36"E 37°30'20"S, south 146°17'13"E 39°9'33"S and west 140°57'29"E 34°28'23"S.

Public land forests extend to all parts of the state and range from low multi-stemmed Mallee woodland across flat and gently undulating topography in the Northwest and Box-Ironbark forests, characterised by sparse to dense canopies of box, ironbark and gum-barked eucalypts up to 25 m tall, on flat to undulating landscapes on rocky, auriferous soils across central Victoria. Highly variable medium and tall canopy damp sclerophyll forests are widespread across the study area, found on a range of loamy, clay-loam and sandy-loam soils. Tall (up to and above 75 m) wet sclerophyll forests are found mostly in the eastern part of the study area on deep loamy soils at higher elevations. Dry sclerophyll forests are prevalent throughout the east, central and southwest parts of the study area on clay-loam, sandy-loam and shallow rocky soils of exposed hillsides, with canopies typically less than 25m tall, with crooked, spreading trees (Viridans, 2000).

The study area is characterised by a range of different climate zones and diverse topography. The northwest region experiences semi-arid conditions, with low median annual rainfall (less than 250 mm in parts), with coastal areas experiencing a cooler temperate climate. Dry inland plains dominate much of the central and western parts of the state. The Victorian Alps—part of the Australian Great Dividing Range mountain system—extend east-west from the centre of the study area, with elevation up to 2,000 m. The Victorian Alps experience the lowest average temperatures and highest precipitation (greater than 1,400 mm/yr) in the study area. This variety of climate and topography is reflected in the variation in forest types and structure across the study area.

3.4.2. Training Data

Classification training data were derived from seven hundred and sixty-six 2 × 2 km land cover maps, systematically distributed across the Victorian Forest Monitoring Program (VFMP) (Haywood et al., 2016) random stratified grid (Figure 3-2). On-

screen digital aerial photographic interpretation (API) of high-resolution (30 cm and 50 cm pixels) colour aerial photographs (photoplots) across the study area (acquired over the period 2006 to 2010) were used to create the land cover maps, based on a land cover classification system (Mellor and Haywood, 2010) comprising broad forest type, canopy height and cover. The delineation of landscape objects into broad forest type/land cover classes, three canopy cover and three height classes, was undertaken by trained interpreters. Crown shape, size and arrangement, shadow and photographic image colour were all used for interpretation of the aerial photography. For the classification of forest, the Australian National Forest Inventory (NFI) forest definition (National Forest Inventory, 2003) was used, with an applied 0.5 ha minimum mapping unit, consistent with the UNFAO forest definition (Food and Agriculture Organization of the United Nations, 2001).

API data were aggregated into forest and non-forest training data classes. Mapping on pre- and post-2008 photography was adjusted to a baseline date of December 31, 2008, using ancillary GIS data to re-attribute and update API polygons, based on major known land cover changes associated with wildfire and clear fell logging. Training data API maps are further stratified by IBRA (Interim Biogeographic Regionalisation for Australia) Bioregions—relatively large, geographically distinct areas of land that share common characteristics, including geology, landform patterns, climate, ecological features and plant and animal communities. Eleven Bioregions are located within the study area. Figure 3-2 shows the distribution of VFMP sample land cover maps across the study area and Bioregions and example API land cover maps. For further information on the API method, refer to (Farmer et al., 2013). API vector data were converted to raster format to align with the 30 × 30 m pixels of Landsat satellite imagery (described in the following section).

3.4.3. Predictor Variables

Nineteen cloud-free Landsat TM scenes were used to build a study area mosaic; selected and downloaded from USGS Earth Explorer (U.S. Geological Survey, 2013). Satellite images were acquired between February and March 2009, corresponding to late summer conditions with relatively high scene sun angles (to minimise shadow and terrain effects) and designed to maximise spectral differences between overstory evergreen woody vegetation and seasonal understory vegetation.

Where cloud-free images were unavailable, the acquisition period was extended to December 2008 or the summer period in the preceding or following year. Images were downloaded in USGS L1T georectified and terrain-corrected format, at a spatial accuracy considered acceptable for the study (± one 30 m pixel). Landsat TM spectral bands 1–5 and 7 were pre-processed to minimise sources of between-scene spatial and temporal variation associated with different atmospheric conditions, topography, sensor location and sun elevation. A physical model was applied to convert image digital numbers (DNs) to surface reflectance standardised to a fixed viewing and illumination geometry, incorporating the Shuttle Radar Topography Mission (SRTM) Digital Elevation Model (CSIRO, 2011), using a methodology described in Flood et al., (2013). Pre-processed image tiles were mosaicked to create six study area surface reflectance Landsat TM bands.

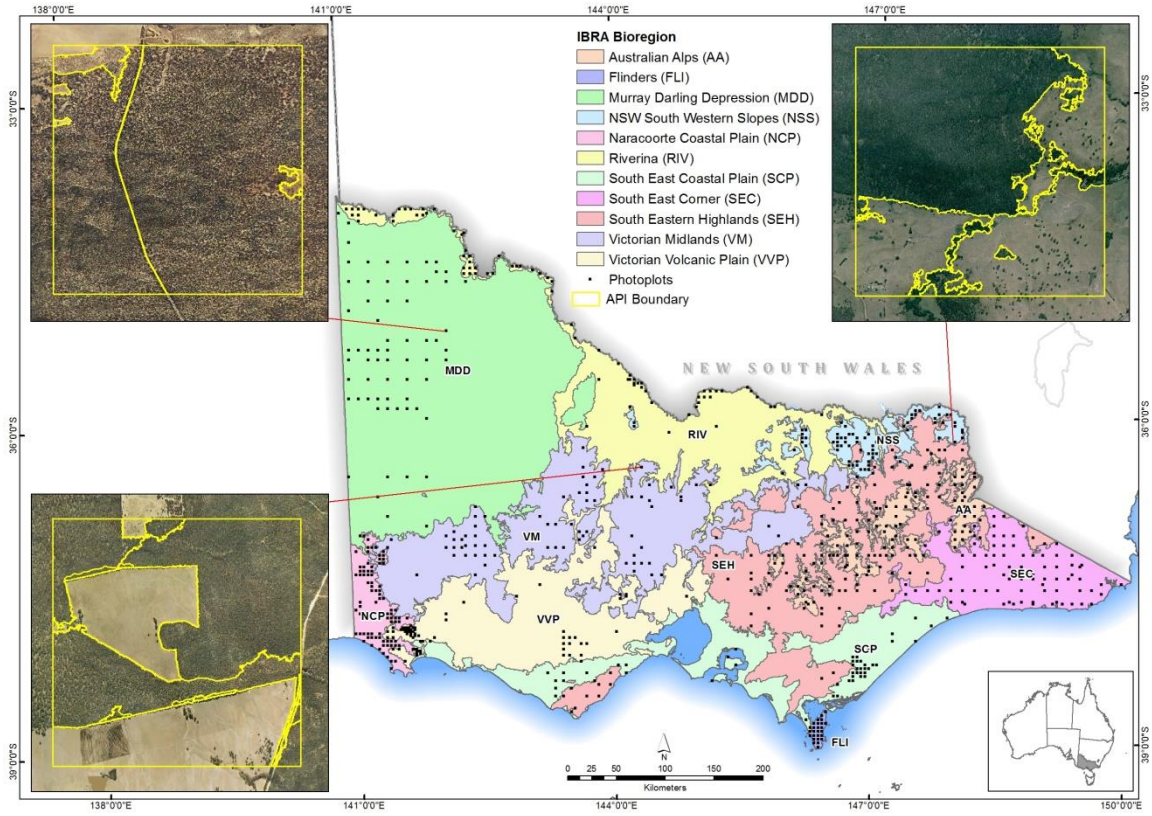


Figure 3-2 Victorian Interim Biogeographic Regionalisation for Australia (IBRA Bioregions) and aerial photographic interpretation (API) land cover maps (1:25,000)

Textural indices were derived from an NDVI layer produced using the Landsat TM surface reflectance bands 3 and 4, rescaled to a 6-bit raster (64 grey levels). Three first order (occurrence) texture measures were calculated using 3 × 3, 5 × 5 and 7 × 7 cell neighbourhood moving windows across the grey-scaled (Haralick, 1979) NDVI layer—these were variance, diversity (number of different values within the neighbourhood) and interspersion (proportion of cells in the neighbourhood, which differ from values assigned to the centre cell in the neighbourhood plus one). Three different sizes of neighbourhood windows were designed to capture the range in ecosystem textural variance across the study area.

Phenological temporal-variance in the study area was derived from state-wide multi-temporal MODIS NDVI data (MOD13Q1). A multi-temporal raster stack of twenty-three 250 m spatial resolution MODIS (16-day) NDVI images were extracted for Victoria, over the calendar year January 2008 to January 2009, from Australian mosaics (produced using the methodology described in (Paget and King, 2008)). To generate the temporal variance in NDVI, a one standard deviation raster was calculated from each annual multi-temporal image pixel-stack.

Elevation (metres), slope (degrees) and aspect (degrees) were derived from a one second (~30 m) smoothed digital elevation model (CSIRO, 2011). Climate surfaces were generated using the BIOCLIM component of the ANUCLIM (version 5.1) software package (Houlder, 2001), a correlative modelling tool that interpolates climate parameters using spatially explicit digital elevation data and point-based long-term monthly averages of climate variables. A full description of the process can be found in (Houlder, 2001; Beaumont et al., 2005). Elevation data raster cells were resampled to 250 m (an appropriate resolution for the distribution of climate stations across the study area) and used as an input to run the BIOCLIM climate model. A subset of the 35 climatic parameters generated by BIOCLIM was selected for inclusion in the model associated with precipitation, temperature, radiation and moisture. BIOCLIM and MODIS NDVI variance surfaces were resampled from 250 m spatial resolution, using the nearest neighbour method, to align with the 30 × 30 m Landsat TM data, elevation layers and textural indices.

3.4.4. Data Collation

Training and predictor variable data were collated in a GIS database—open-source GRASS Geographic Resources Analysis Support System (GRASS Development Team, 2012)—and exported into statistics package R (R Development Core Team, 2011) for model implementation and analysis, together with training sample raster pixel centroid coordinates. To reduce data redundancy and facilitate interpretation of the model, Pearson correlation coefficients were calculated between all paired combinations of predictor variables. Highly correlated variables ($r2 > 0.9$, $p < 0.001$) were further examined to calculate biserial correlation coefficients between these predictor variables and a dichotomous forest/non-forest training sample class. Of the highly correlated variable pairs, those with the weaker forest/non-forest relationship were excluded from the model. Table 3-1 shows the final predictor variables used in the RF model. Variables excluded from the model were the climate layers mean diurnal range, temperature seasonality and annual mean radiation; and textural indices variance (5 × 5 and 7 × 7 windows), diversity (3 × 3 and 7 × 7 windows) and interspersion (3 × 3, 5 × 5 and 7 × 7 windows).

Table 3-1 Random Forests (RF) predictor variables

Predictor Variable	**Units/Data Source**	**Spatial Resolution (m)**
Surface Reflectance		
Landsat TM band 1	0.45–0.52 µm	30
Landsat TM band 2	0.52–0.60 µm	30
Landsat TM band 3	0.63–0.69 µm	30
Landsat TM band 4	0.76–0.90 µm	30
Landsat TM band 5	1.55–1.75 µm	30
Landsat TM band 7	2.08–2.35 µm	30
Textural Indices		
Variance (3 × 3)	Landsat TM NDVI	30
Variance (5 × 5)		30
Diversity (3 × 3)		30
Phenological Variability		
NDVI Variance	MODIS NDVI	250
Topography and Climate		
Elevation	SRTM DEM	30
Slope	SRTM DEM	30
Aspect	SRTM DEM	30
Annual Precipitation	mm	250
Annual Temperature Range	°C	250

Annual Mean Temperature	°C	250
Annual Mean Moisture Index	0–1	250

3.5. Random Forest Model

3.5.1. Construction and evaluation

The randomForest package (Liaw and Wiener, 2002) in R (R Core Team, 2013) was used to build the RF model, for which there are several adjustable implementation parameters. The primary parameters being (i) number of predictor variables randomly sampled as candidates at each decision tree node split (parameter mtry); (ii) the number of decision trees (or base classifiers) constructed as part of the classifier ensemble (parameter ntree); and (iii) the type of model—classification, regression or unsupervised (parameter type). For model construction in this study, the default mtry value was used (equal to the square root of the total number of predictor variables). To optimize the number of trees (ntree) constructed in the final model, an initial decision tree ensemble was produced with 1,000 trees. Error estimates from the OOB sample showed stabilization of the overall error at 100 trees; therefore, 100 was used for the parameter ntree in the final model.

In addition to the RF model OOB test data, for performance evaluation, a 25% subset of training data was randomly sampled, left out of the training dataset (stratified evenly by forest and non-forest classes). The R package PresenceAbsence (Freeman and Moisen, 2008) was used to calculate the optimal threshold for converting forest probability (0–100) into a binary forest/non-forest classification, based on maximum Kappa. Kappa, percent correctly classified, user's and producer's accuracy and area under receiver operator curve were calculated to evaluate classification performance. The area under receiver operator curve (ROC) is a measure of a model's ability to discriminate presence (i.e., forest) and absence (i.e., non-forest) (Pearce and Ferrier, 2000), calculated from predicted forest probabilities. The ROC is a plot of sensitivity (true positive rate) against specificity (false positive rate). Poor model performance (i.e., where predictive ability is essentially random) returns a near-diagonal ROC plot (true positive rate equal to false positive rate). The area under ROC curve ranges from 0.5 (poor) up to 1. Producer's accuracy (or omission error, one minus

producer's accuracy) is the proportion of a land cover class on the ground (i.e., reference) that is correctly classified in the map (prediction). User's accuracy (or commission error, one minus user's accuracy), is the proportion of a mapped (predicted) class on a map, which matches the corresponding class on the ground (reference). Producer's accuracy measures classification scheme accuracy, while user's accuracy measures the output map generated from the classification (Shao and Wu, 2008).

3.5.2. Implementation

The RF model was implemented to predict and map forest probability across the study area. As R holds objects in virtual memory, there are limitations on the resources available for data processing. Therefore, the RPy Python package (Gautier, 2012) was used, allowing R functionality to be managed within the Python environment outside of R. The study area was divided into two hundred 40 km^2 tiles, and the RF model was implemented using parallel processing to calculate forest probability across multiple tiles simultaneously, after which the forest probability tiles were mosaicked together into a single forest probability layer.

Probability values (calculated from the proportion of decision tree votes among all base classifiers in the ensemble) were converted into binary forest and non-forest classes using the probability threshold calculated to maximise the Kappa statistic. To apply the forest definition 0.5 ha minimum mapping unit (MMU) and remove noise from the map, the forest/non-forest classification raster was first re-sampled from 30 m to 28.86 m, so that a 0.5 ha MMU area comprised six whole raster pixels. Horizontally, vertically and diagonally contiguous forest and non-forest cells were grouped together and attributed a count of the cells within each group. Raster cells within forest cell groups comprising less than six cells (i.e., less than 0.5 ha) were re-labelled as non-forest, and raster cells within non-forest cell groups comprising less than 6 cells were re-labelled as forest. Figure 3-3 shows the forest probability and final binary forest/non-forest maps.

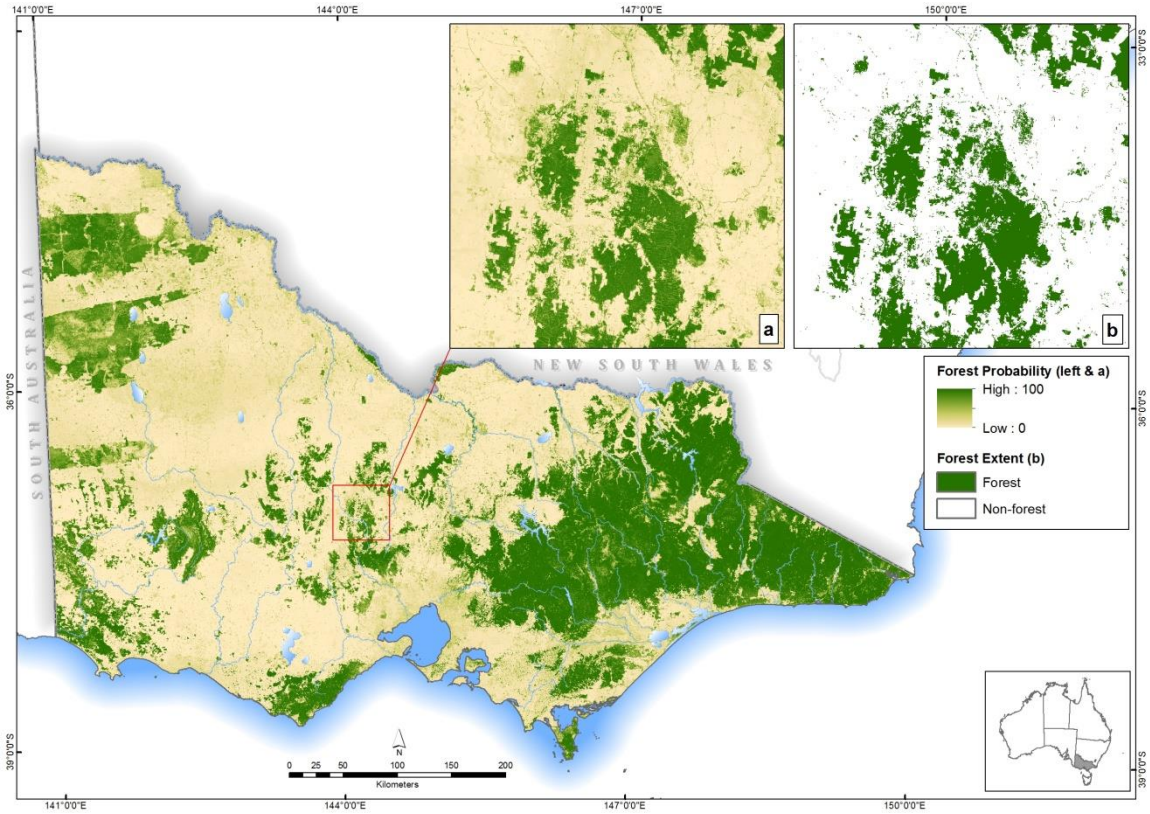


Figure 3-3 Implemented Random Forests model forest probability map (a) inset forest probability map (0–100); (b) final forest classification, based on a binary threshold.

3.6. Results and Discussion

3.6.1. Classification Accuracy

Overall accuracy (percent correctly classified) and Kappa results were high for forest and non-forest prediction using the RF model (Table 3-2). Overall accuracy of 96% was achieved, with a Kappa coefficient of 0.91. The threshold value for converting continuous forest probability scores into forest/non-forest classes, optimized to maximize overall Kappa, was 0.5. User's accuracy was marginally higher for the forest class than the non-forest class, indicating a greater tendency for the model to misclassify non-forest land cover as forest, leading to a slight overestimation of forest extent. A comparison of model performance (user's and producer's accuracy) between the test data and the RF OOB accuracy assessment shows marginally lower producer's and user's accuracy for non-forest classification, and user's accuracy in

the forest class was returned by the OOB; however, differences between the two accuracy assessment data sources are minor.

The high Kappa coefficient (0.91) for the forest/non-forest classification model is encouraging, and the model accuracy performance is consistent with studies that have successfully discriminated forest from non-forest land cover categories in other natural environments using RF (Gislason et al., 2006; Chan and Paelinckx, 2008). The area under curve (AUC) score (0.91) shows that the RF forest/non-forest classifier has excellent overall model accuracy.

Table 3-2 Random Forests accuracy assessment. CI, confidence interval; OOB, out-of-bag.

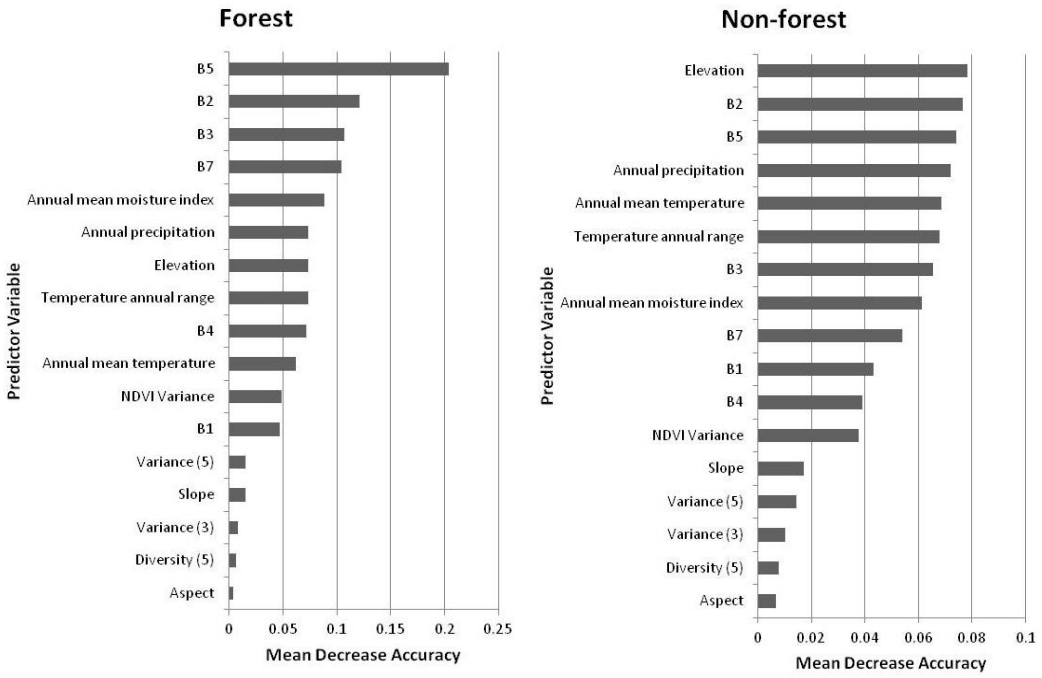
Kappa (CI 95%)	0.914 (0.909–0.919)	
AUC (CI 95%)	0.992 (0.991–0.992)	
Percent Correctly Classified (CI 95%)	95.7 (95.4–95.9)	
	Forest	**Non-forest**
Kappa maximised binary threshold value	0.5	
Sensitivity	94.42	**96.94**
Specificity	**96.94**	94.42
Test (Validation Data)		
Producer's accuracy (omission)	94.42	**96.94**
User's accuracy (commission)	**96.86**	94.56
Test OOB		
Producer's accuracy	94.60	**96.44**
User's accuracy	**96.51**	94.49

3.6.2. Variable Importance

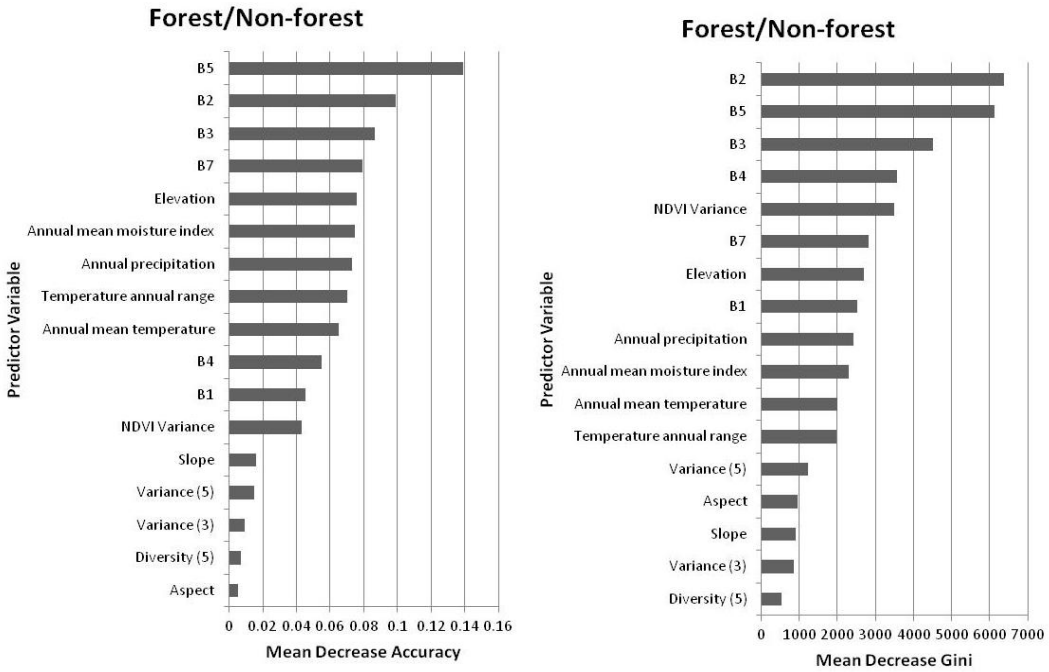
Landsat TM band 5 (shortwave infrared) was shown to be the most important variable in predicting forest (Figure 3-4(a)) based on the calculated mean decrease in accuracy (MDA) score. Band 5 was considerably more important than the next most important predictor variables—Landsat TM bands 2, 3 and 7, followed by elevation

and the four climate surfaces. The high importance of the middle-infrared band 5 (1.55–1.75 µm) in differentiating forest from non-forest at the pixel-level is likely to be associated with its vegetation and soil moisture sensitivity properties. For non-forest classification, based on MDA, elevation was the most important variable in the RF model, followed by bands 2 and 5. The influence of elevation may be associated with less rainfall at lower elevations, but is also very likely to reflect the land use history of the study area, whereby low flat land productive agricultural land has been extensively cleared (Woodgate and Black, 1988). Landsat bands 5, 2, 3 and 7 were the most important predictor variables for forest/non-forest differentiation (Figure 3-4(c)).

Landsat TM band 2 was the most important predictor variable, followed closely by band 5, based on the mean decrease Gini (MDG) measure (calculated for each predictor variable as the cumulative increase in data purity associated with each decision tree node split). Bands 3 and 4 were the next most important variables, followed by NDVI variance and band 7. In comparing the variable importance ranks between the two measures, MODIS NDVI variance was ranked 7 places higher in the MDG measure compared to MDA and band 4 (near-infrared), six places higher. These bands can be considered more important with respect to increasing the purity of training data samples after splitting at decision tree nodes, but less important based on the mean decrease accuracy.



a) b)



c) d)

Figure 3-4 Random Forests predictor variable importance measures.

(a) Mean decrease accuracy for forest prediction; (b) mean decrease accuracy for non-forest prediction; Random Forests predictor variable importance measures. (a) Mean decrease accuracy for forest prediction; (b) mean decrease accuracy for non-forest prediction; (c) mean decrease accuracy for forest and non-forest prediction; and (d) mean decrease Gini for forest and non-forest prediction.

The MODIS NDVI variance was included in the model as a means of discriminating seasonally dynamic grasses and understory vegetation from more phenologically 'stable' forest canopy reflectance. While results rank this variable as having a reasonably high degree of importance in decision tree node splitting (Gini purity), the low spatial resolution of this layer (250 m) and high spectral heterogeneity within MODIS pixels is likely to be a factor in its lower MDA importance ranking for forest prediction.

Results of this study on application of RF for large area forest classification are encouraging and demonstrate the classifier's utility in an operational land management agency context. Our results confirm the findings of other studies using RF, that this ensemble classifier can be used to learn complex non-linear

relationships. Variable importance measures demonstrate the successful integration of multiple sources of data in predicting forest—remote sensing spectral data and contextual topographic-climate variables.

This study demonstrates the feasibility of using an open-source framework for constructing and evaluating an RF model and its implementation to produce an accurate operational land management agency forest cover map. The framework established successfully integrates freely available spatial data—pre-processed and collated in GRASS—into the R statistical analysis environment. After construction and validation of an RF classifier, the resulting model was implemented in GRASS using an R-GRASS interface package, *spgrass* (Bivand, 2007), before finally using GRASS to filter the forest prediction map and apply the minimum mapping unit of the adopted forest definition to the final forest extent spatial product.

In this study, we evaluated the operational performance and utility of the ensemble decision tree classifier, Random Forests (RF), for producing an accurate large area (about 220,000 km^2) land management agency forest map. This study is unique in demonstrating the operational implementation of RF at the regional-scale within an open-source software framework, using GRASS GIS (GRASS Development Team, 2012) and R (R Development Core Team, 2011) statistics software. The framework described, comprising stages of data pre-processing, collation, modelling, evaluation and implementation, contributes to the deployment of affordable programs for collating and processing large volumes of multi-source remote sensing and ancillary GIS data to produce consistent and accurate forest cover maps across complex, noisy and heterogeneous landscapes.

We incorporated Landsat TM and MODIS satellite imagery, textural indices, modelled climate surfaces and topographic layers into an RF model, to accurately predict and map forest across an area comprising millions of hectares of complex and highly diverse forest ecosystems over varying topography, dominated by open canopy sclerophyll forests and woodland. Sample aerial photography land cover maps were used to derive training and test (validation) data. The overall accuracy and Kappa statistics for forest/non-forest classification were 96% and 0.91, respectively. Forest classification achieved a producer's accuracy of 96% and a user's accuracy of 94%. Estimated predictor variable importance measures derived

from the Gini Index and out-of-bag (OOB) training data, showed Landsat TM bands 5 and 2 to have the strongest influence in forest/non-forest class-separability.

3.7. Conclusions

Results show how the RF algorithm can be effectively used to learn the conditional, complex and non-linear relationships between forest vegetation and biophysical factors, to build an accurate forest classifier across highly diverse and dynamic ecosystems. In a land management agency context, the study demonstrates how the RF can be used to address the challenges and operational constraints of land cover classification, including the use of non-parametric and noisy data, its implementation using open-source software, and the integration of multi-source regional scale ancillary spatial data.

While these results are encouraging for the application of RF in an applied natural resource management context, there are several important areas of further research that warrant further investigation. Based on the "Strong Law of Large Numbers", Breiman (Breiman, 2001) showed that RF does not over-fit training data as more trees are grown. While results from OOB accuracy and test data support this, the performance of the RF model is based on the important assumption that training data is representative of forest and non-forest classes from across the study area. As proposed by Armston *et al.* (2009), in a study investigating the use of RF regression analysis to predict overstory foliage projective cover (FPC) from Landsat TM and ETM imagery, an important next step would be to undertake an independent assessment of the implemented classification model (forest extent map, Figure 4) from sites located away from training data. This would improve understanding of the extent to which spatial autocorrelation between training data samples (*i.e.*, contiguous or closely located pixels) lead to bias, as well as reduced variance and representativeness (Chen and Stow, 2002). In short, how do spatially auto-correlated model training and validation data over-estimate the accuracy and performance of the RF classifier across large heterogeneous landscapes? Other important directions for further research include: (1) the characteristics of RF training data, to better understand how the classifier manages noise and outliers; (2) understanding how different sampling techniques affect classifier performance; and (3) the implementation of the

classifier model on other acquired and calibrated remote sensing image dates and its utility for producing accurate multi-temporal forest extent maps in a monitoring context.

Chapter 4. Exploring issues of training data imbalance and mislabelling on random forest performance for large area land cover classification using the ensemble margin

Based on the peer-reviewed published article:

Mellor, A., Boukir, S., Haywood, A. and Jones, S., 2015. Exploring issues of training data imbalance and mislabelling on random forest performance for large area land cover classification using the ensemble margin. *ISPRS Journal of Photogrammetry and Remote Sensing*, 105, pp.155-168.

4.1. Introduction

Accurate spatially explicit classification maps are important sources of information for natural resource land managers and forest monitoring programs. Land management agencies typically monitor and report on large areas (i.e. regional or continental scale, covering millions of hectares) relying on the interpretation of large complex remotely sensed data, calibrated and validated using, typically, a limited amount of ground reference data (Lippitt et al., 2008). Studies have demonstrated the successful application of ensemble machine learning classifiers, such as Random Forests (RF), integrating remote sensing (satellite imagery) and ancillary spatial data, to improve supervised classification accuracy of forest and other natural environment land cover maps (Cutler et al., 2007; Mellor et al., 2013; Rodriguez-Galiano et al., 2012), for which conventional parametric statistical classification techniques might not be appropriate (Gislason et al., 2006). In ensemble classification, multiple (base) classifiers are constructed. From the ensemble, a final class is determined by, for example, averaging or a majority vote. In machine learning, the margin theory examines the proximity of data points to decision boundaries.

Margin theory is a means by which to understand and evaluate ensemble classification and can be used to estimate confidence in the classification outcome (Schapire et al., 1998). Such ancillary information is important, particularly when relying on satellite image derived maps for scientific inference (McRoberts, 2011). The characteristics of training data is a fundamental consideration when constructing any supervised classifier (including ensemble machine learning).

Learning from imbalanced training data (i.e. unevenly distributed data between classes) is a common problem (Japkowicz & Stephen, 2002). Machine learning algorithms, such as RF, are constructed to minimize the overall classification error rate and imbalanced training data can result in poor accuracy for minority classes (Chen et al., 2004). Furthermore, it is assumed that, in its implementation, the classifier is run using data drawn from the same distribution as the training data (Provost, 2000). In RF, decision trees are induced using bootstrap samples of training data (Breiman 2001) and in situations where training data includes only a minority of training data samples for a particular class (relative to other classes), it is likely that a bootstrap sample may include few or even no samples from this class and hence

fewer leaves describing the minority class, resulting in poor classification accuracy for the minority class prediction (Chen et al., 2004) as well as weaker confidence estimates (He & Garcia 2009).

The imbalance training data problem is common in large area natural resource applications using remote sensing (e.g. forest classification), whereby within reference data, rare land cover or forest classes may be under-represented relative to more abundant classes, due to the time and cost resource constraints of collecting enough representative training samples. Studies have shown balanced datasets improve overall classification compared to imbalanced data (Weiss and Provost, 2003; Estabrooks et al., 2004). Several techniques have been demonstrated to address the imbalance training data problem. These include down-sampling majority classes (Freeman et al., 2012) and weighting rare training observations more highly than common classes (Chen et al., 2004). Techniques involving over-sampling the minority class through replication of samples to match the quantity of majority class training samples (Ling & Li, 1998) and a combination of over-sampling (minority) and down-sampling (majority) training classes (Chawla et al., 2002) have also been explored.

Training data class mislabelling (or noise) (Sluban et al., 2013) is another important consideration in using bagging ensemble algorithms such as RF. This is an issue that often adversely affects machine learning algorithms (Guo, 2011). In large area remote sensing classification for forest monitoring programs, training data typically include ground-based (i.e. field data collection) (Lillesand and Kiefer, 1994) or data sampled from remote sensing imagery of a higher spatial resolution, such as very high resolution satellite imagery (e.g. Quickbird) or digital aerial photography (Wulder, 1998).

Deriving training data using manual and semi-automated mapping from high spatial resolution imagery are methods which are prone to a variety of sources of labelling error and bias. These sources include interpreter bias and inconsistency, spatial resolution (scale), geometric and radiometric variability, and error associated with temporal discontinuity between training data (i.e. aerial photography acquisition date or season) and satellite imagery used for classification (Morgan et al., 2010). Other training data labelling errors are associated with inconsistency of vegetation

classification methods, techniques and spatial resolution (Bradley and Friedl, 1996). In forest environments, common training data class mislabelling errors are caused by the similarity of forest types as their signatures appear in aerial photography (Delaney and Skidmore, 1998).

For their application in an operational setting (such as a large area forest monitoring program), it is important that machine learning classifiers are resilient to mislabelling in training data (Lippitt et al., 2008). Studies have demonstrated the relative resilience of bagging ensemble classifiers, such as RF, to training data noise (class mislabelling) (DeFries and Cheung-Wai Chan, 2000). In evaluating machine learning algorithms for land cover change mapping, Rogan et al. (2008) investigated the effect of artificially introduced training data noise to classification accuracy. Their study found the addition of 10% noise reduced accuracy of decision tree classifiers S-Plus and C4.5 by 7% and 20% respectively. In a land cover classification study, Rodriguez-Galiano et. al. (2012) found the RF classifier performance (overall classification error) to be relatively insensitive for up to 20% deliberately mislabelled training instances, above which error rate increased exponentially. Na et al. (2009) reported a reduction in RF overall accuracy by almost 50% associated with a 30% increase in the amount of artificial noise.

In this chapter, we examine how training data class imbalance and class mislabelling affect RF performance in the context of large area forest classification in an operational land management agency setting. This was achieved across diverse and complex forest ecosystems and topography, dominated by open canopy sclerophyll forests and woodland. We evaluate RF performance associated with training data characteristics through a new perspective involving ensemble margins. The magnitude of ensemble margin is usually interpreted as a measure of confidence in classification prediction and significant work has been published about bounding and reducing prediction error based on the classification margin (Schapire et al., 1998; Guo, 2011).

The nature of a training set can have a major impact on classification accuracy (Foody, 1999) and the margin ensemble can be used to understand how training data characteristics can affect classification outcomes. Foody (2002) emphasizes the need for more accuracy assessment information (including confidence measures) to be

provided with land cover and other remote sensing derived classification maps, to aid user interpretation and application. The value of very large area mapping is ultimately limited by poor quality accuracy assessment and reporting (Foody, 2002). In this study, we evaluate new ensemble margin statistics as a means of providing distinct information about margin distribution and classification prediction confidence and supplementing traditional measures of classification performance. Furthermore, we introduce a novel method for assessing classification uncertainty through the use of an ensemble margin weighted confusion matrix, that to the best of our knowledge is used for the first time in land cover classification using remote sensing and ancillary geospatial data.

4.2. Random Forests

Random Forests (RF) uses a bootstrap aggregation technique (bagging) (Breiman, 1996) to generate sub-sets of training data with which to build an ensemble of decision trees (base classifiers). The bagging process involves resampling the original training set with replacement, resulting in a greater diversity of decision trees, thereby improving classifier stability and accuracy. Moreover, in constructing trees, as some training data instances may be used more than once or not at all, correlation between trees is reduced, and as a result, RF is more robust to variations in input data and less sensitive to mislabeled training data or over-fitting (Pal, 2005; Rodriguez-Galiano et al., 2012) .

In constructing each decision tree, at each node (split) a randomly selected subset of model predictor variables are evaluated for partitioning the data into increasingly homogeneous subsets - the variable used to split the data is that which results in the greatest increase in data purity. Increasing the number of predictor variables selected for tree construction results in stronger individual decision trees, but with increased correlation between trees, model accuracy is reduced (Rodriguez-Galiano et al., 2012). Therefore, to minimize the generalization error, it is necessary to optimize this parameter, together with the number of decision trees in the ensemble. Tree building continues until there are no further gains in purity. A response variable can be predicted as an average (continuous variable classification) or model vote (categorical classification) among all decision trees. Sample Out of Bag (OOB) data

(that are not drawn in the bagging sample used for tree construction) are used to compute accuracies and error rates averaged over all predictions (Cutler et al., 2007).

4.3. Ensemble Margin

The margin function is an important concept in ensemble classifiers such as RF. The classical margin function (Schapire et al., 1998) is calculated for each known data instance, and ranging from -1 to +1, is the normalised difference between the votes assigned to the true class and those assigned to the most voted class that is different from the true class (Guo et al., 2011). An alternative margin function, which does not require the known class labels, is an unsupervised version of Schapire's margin. It combines the first and second most probable class labels under the model. Equation 1 shows how the unsupervised margin is calculated where V_{c1} represents the number of votes for the most voted class c_1 for instance x and V_{c2} the number of votes for the second most popular class c_2. The unsupervised margin ranges from 0 to 1. Instances close to class boundaries (margin values around 0, i.e. not redundant and which are located near decision boundaries) are the most informative for the classification task. In that case, the true class labels are not of significance. As such, the unsupervised margin may be more robust to class mislabelling (Guo, 2011). Hereafter, unless otherwise specified, the term margin is used to describe the unsupervised margin.

$$margin(x) = \frac{V_{c1} - V_{c2}}{\sum_{c=1}^{L}(V_c)} \quad (1)$$

This may be also be expressed as:

$$\frac{max_{c=1,\ldots,L}(V_c) - max_{c=1,\ldots,L \cap c \neq c_1}(V_c)}{T}$$

where *T* represents the size of the ensemble and *L* represents the number of classes.

The margin of a correctly classified instance should be as high as possible (close to 1) and the margin of a misclassified instance, as low as possible (close to zero). Indeed, the lower the margin of a misclassified instance, the greater the opportunity for improving the classification by the ensemble of base classifiers.

4.4. Study Site and Data

4.4.1. Study Area

The study area (Figure 1) comprises approximately 7.1 million hectares of public land covering about one third of the state of Victoria, south east Australia. The extents of Victoria are north 141°47'36" E 33°58'54" S, east 149°58'36" E 37°30'20" S, south 146°17'13" E 39°9'33" S and west 140°57'29" E 34°28'23" S. The study area includes two public land tenures – national parks and conservation reserves and multiple-use commercial State forests, and covers a wide range of ecosystem types and a high diversity of forest types and structures, dominated by open sclerophyll forests. The state is characterised by diverse topography and a range of climate zones.

4.4.2. Reference (Training and Test) Data

On-screen digital Aerial Photographic Interpretation (API) of 30-50 cm ground sample distance colour aerial photographs (photoplots), acquired between 2006 and 2010, was the source of classification reference data (used for model training and test validation data). Seven hundred and sixty-six 2×2 km photo-plots, systematically distributed across a state-wide random stratified grid (Figure 3-2), were used to produce land cover maps based on a land cover classification system (Mellor and Haywood, 2010) which included broad forest or land cover type, forest canopy height class (low, medium, tall) and canopy cover class (woodland, open, closed). Thresholds for forest canopy height and cover are shown in Figure 3-1. Land cover object delineation was undertaken by trained interpreters, using photo-plot information including crown shape, size and arrangement, colour and shadow. Forest delineation followed the Australian National Forest Inventory forest definition (greater than 20% crown cover, minimum two metre stand height) (Department of Agriculture Fisheries and Forestry, 2012) with a 0.5 ha minimum mapping unit applied to the land cover maps, based on the UNFAO forest definition (Food and Agriculture Organization of the United Nations, 2001). Ancillary GIS polygons representing the boundaries of clear-fell logging and wild fires were used to update land cover polygons to a summer season 2008/09 baseline. A comprehensive

description of the API land cover mapping method can be found in (Farmer et al., 2013).

For this study, land cover data were aggregated into three forest canopy cover classes (woodland, open, closed), shrub (woody vegetation not meeting the 2 m height threshold or 20 % crown cover) and other non-forest land cover, and binary forest/non-forest classes. Land cover polygons were converted into raster format aligned with the resolution of classifier predictor variables described in the following section. Test (validation) data were randomly selected, using simple random sampling, from the total reference data and set-aside prior to selection of training data for building RF models.

4.4.3. Predictor variables

Predictor data comprised multi-source remote sensing, topographic and climate variables. Landsat TM data are commonly used in studies for forest type discrimination and cover estimation (Boyd and Danson, 2005). In this study, a 6-band study-area Landsat TM mosaic comprising nineteen scenes acquired between February and March 2009, was used. The time of acquisition was designed to correspond with the training and test data (land cover mapping) and late summer conditions, where high sun angles reduce shadow and terrain effects and where spectral reflectance differences between overstory evergreen woody vegetation and seasonally dynamic understory vegetation is maximized (Mellor et al., 2013). Following Flood et al. (2013), Landsat TM scenes were standardized to surface reflectance to reduce sources of inter-scene variation associated with atmospheric conditions, topography, sensor location and sun elevation.

Where vegetation appears spectrally similar but has differing spatial patterns, textural indices derived from satellite imagery have been shown to improve classification performance (Kayitakire et al., 2006; Rodríguez-Galiano et al., 2011). For this study, a grey-scaled (8-bit) Normalized Difference Vegetation Index (NDVI) layer was generated using Landsat TM bands 3 and 4, from which were produced textural indices representing spatial variation in optical imagery (Haralick, 1979). First order texture measures of variance and diversity were generated for 3x3 and 5x5 cell

neighbourhood moving windows, designed to capture textural variance of the study area's forested ecosystems (Mellor et al., 2013).

Multi-temporal MODIS NDVI data was used to represent phenological variance in the study area over a calendar year. Twenty-three 250 m spatial resolution 16-day MODIS NDVI images from 2008, covering the study area, were extracted from an Australian national mosaic (Paget and King, 2008), from which one standard deviation raster was calculated using each multi-temporal image pixel stack – to represent seasonal variance in NDVI.

Topographic and biophysical information was included in the model as a means of capturing species-environmental relationships, which are central to predictive geographical modeling (Guisan and Zimmermann, 2000; Mellor et al., 2013). Bioclimatic maps support assumptions about the influence climate has on forest type distribution (Beaumont et al., 2005) and the composition of vegetation that can be expected to occur in areas with similar soils, climate and topography (Franklin, 1995). Following Gislason et al. (2006), topographic predictor variables of elevation (metres), slope and aspect (degrees) were included in the model, derived from a 1 second (~30 m) smoothed Digital Elevation Model (DEM) (CSIRO, 2011). The DEM was used to generate precipitation, temperature, radiation and moisture climate prediction surfaces using BIOCLIM in the ANUCLIM (v 5.1) software package (Houlder, 2001). A detailed description of the BIOCLIM modelling process can be found in (Beaumont et al., 2005).

Reference data vector polygons were converted into raster format and resampled, using the nearest neighbour method, together with the predictor variables, to align with the resolution and extent of the 30 m x 30 m Landsat TM raster cells.

4.5. Methods

4.5.1. Experiments

Five experiments were carried out, using binary (experiments 1 and 4) and multiclass (experiments 2, 3 and 5) models, to examine the effect of training data class imbalance and mislabelling on RF performance. Binary classification experiments used forest/non-forest categories. Multiclass experiments used canopy cover classes

(woodland, open and closed), together with two non-forest classes (shrub and other non-forest). For each experiment, two training data sizes (subsets) were established: optimal and critical (Table 4-1). The optimal subsets for binary and multiclass experiments were estimated by running multiple iterations of the RF classifier with a balanced distribution of training samples per class, increasing the training set size with each iteration. Classification accuracy typically increases with training data size until reaching an asymptotic level (constant accuracy) indicating that the optimal accuracy has been achieved. The optimal training size is the minimum training size leading to this maximum accuracy, hence it is optimal in terms of both classification accuracy and complexity. Critical subset sizes for binary and multiclass experiments were 1% and 5% of the optimal subsets respectively. To ensure stable accuracy estimates from critical training sets, ten iterations of each experiment were run, from which mean performance measures were calculated.

Table 4-1 Optimal and critical training and test set sizes used for binary and multiclass experiments

Classification	Training data (total samples)		Test data
	Critical	Optimal	
Binary (experiments 1 & 4)	100	10,000	2,500
Multiclass (experiments 2, 3 & 5)	5,000	100,000	25,000

Training data class imbalance experiments

An examination of the impact of class imbalance on RF model performance was undertaken on training sets for binary (experiment 1) and multiclass (experiments 2 and 3) classification. For the binary imbalance experiment (1), for each RF model, balance as a ratio of forest to non-forest training samples was adjusted while maintaining the same total number of training samples. For the multiclass (forest cover) classification experiments (2 and 3), an initial RF model was generated using a class-balanced distribution of training data. Canopy cover classes from the forest super class (comprising woodland, open and closed forest classes) with the highest and lowest producer's accuracies (omission errors) were identified from a confusion matrix generated with test data. Hereafter these are referred to as *best* class (i.e. easiest class) and *worst* class (i.e. hardest class).

The first multiclass imbalance experiment (2) adjusted the ratio of *best* to *worst* class training samples in each RF model. The second multiclass imbalance experiment (3) involved generating imbalance in the training data samples by increasing the proportion of the *worst* class while simultaneously decreasing the proportion of the *best* class by the same amount. This is a sensible strategy that can outperform a balanced distribution at least in terms of per class accuracies as will be shown later by our experiment results. Hence, balancing training data is not always the best strategy especially if the classification task involves classes of differing complexity. For every iteration of both multiclass imbalance experiments, the number of samples representing the remaining classes was kept constant, thereby maintaining the same total number of multiclass training samples for each experiment.

For the binary classification, imbalance experiment (1) involved adjusting balance as a ratio of forest to non-forest training samples (10:90, 25:75, 50:50 (balanced), 75:25 and 90:10) while maintaining the same total number of samples (10,000 and 100 for optimal and critical sizes respectively). For the multiclass classification experiments (2 & 3), an initial RF model using a class-balanced training data distribution, determined the closed canopy cover and open canopy cover classes to be the *best* (easiest) and *worst* (hardest) of the forest cover super class. These classes were adjusted in the multiclass imbalance experiments. Firstly by changing the ratio of open to closed cover class samples (10:90, 25:75, 50:50 (balanced), 75:25 and 90:10), while maintaining the same total number of samples including each of the remaining classes (woodland, shrub and other non-forest) for both optimal (100,000 samples) and critical (5,000 samples) cases. In the second multiclass imbalance experiment, the proportion of open and closed cover classes in each model was adjusted by increasing the proportion of the Open (worst, or most difficult) class (5%, 10%, 25%, 50%, 75% and 90%), while simultaneously decreasing, by the same proportion, the number of Closed (best, or easiest) class training samples (Table 4-2).

Table 4-2 Training set sizes for each class for multiclass imbalance (experiments 2 and 3)

Class	Training set size per class	% Increase in *worst* class and decrease in *best* class					
		5	10	25	50	75	90
Optimal							
Woodland	20,000	20,000					
Open (worst)	20,000	21,000	22,000	25,000	30,000	35,000	38,000

Closed (best)	20,000	19,000	18,000	15,000	10,000	5,000	2,000
Shrub	20,000			20,000			
Non-forest	20,000			20,000			
Critical							
Woodland	1,000			1,000			
Open (worst)	1,000	1,050	1,100	1,250	1,500	1,750	1,900
Closed (best)	1,000	950	900	750	500	250	100
Shrub	1,000			1,000			
Non-forest	1,000			1,000			

4.5.2. Training data class mislabelling experiments

For the binary classification mislabelling experiment (4), training data mislabelling was undertaken by randomly re-assigning a proportion of forest instances as non-forest and non-forest instances as forest. While introducing artificial noise into binary class training data is a straightforward process, it is not always the case for multiclass problems. Studies typically use random class-label switching to simulate noise in a classification (Rogan et al., 2008; Guo, 2011; Rodriguez-Galiano et al., 2012; Sluban et al., 2014). We propose an alternative approach, designed to replicate realistic real-world operator misclassification (mislabelling) of reference data instances (Lowell et al., 2005), that results in a more reliable analysis of noise effects on a supervised classifier performance. In the multiclass mislabelling experiment (4), a preliminary RF model was built using a balanced distribution of samples of the multiclass training data. A confusion matrix derived from the OOB data (not used in the bootstrap training samples) was used to determine, for each class c_i, the class to which it was most frequently misclassified l_i, i.e. the most frequent error class predicted by the model from the OOB data. For the multiclass classification, starting with a training data set with more or less "real" noisy labels whose amount is unknown in practice, the introduction of artificial noise in class labels is performed by mislabelling a proportion of each class c_i to l_i ($c_i \rightarrow l_i$). For this experiment, iterations of the multiclass model were run, mislabelling an increasing proportion of each class c_i to l_i each time. Based on the results (confusion matrix) of a *clean* balanced training set, these were as follows:

- Woodland → Open

- Open → Closed
- Closed → Open
- Shrub → Woodland
- Non-forest → Woodland

4.5.3. Random forest model parameters

The *randomForest* package (Liaw and Wiener, 2002) in R (R Development Core Team, 2011) was used to build the RF models. For all models, the default number of randomly sampled variables as candidates for each decision tree node split was used (equal to the square root of the total number of predictor variables). For binary classification experiments, RF decision tree ensembles were constructed with 100 trees, a typical moderate size ensemble (Tsoumakas et al., 2009). For the more complex multiclass case, more trees were used in each ensemble (150). Assignment of class was determined by the majority of votes from all decision trees in the ensemble, a standard approach for combining the decisions of multiple component learners.

4.5.4. Random forest model performance evaluation

2,500 samples (1,250 per class) were used as test data for binary classification experiments, and 25,000 (5,000 per class) for multi-class experiments (Table 2). Test data were used to calculate overall and per-class accuracies and Kappa coefficient. Kappa is a measure of accuracy adjusted for chance agreement (Carletta, 1996). Kappa coefficient ranges from 0 to 1.0, with 1.0 representing 100% better agreement than by chance alone.

We introduce three ensemble margin descriptive statistics (mean, dominant (frequency) mode, entropy) – measures to analyse the effect of imbalance and mislabelling on RF. Besides, the classification uncertainty is assessed through the use of a novel weighted confusion matrix based on ensemble margin. These measures were calculated from unsupervised margin values (equation 1) of each model. Cumulative frequency distribution curves of correct and misclassified instance

margins were also used to illustrate and compare confidence rates between different models and experiments.

Mean margin

A margin criterion measuring a strong ensemble classifier is one which maximises the mean margin of correctly classified samples and minimises the mean margin of misclassified samples. This normalised measure, ranging from -1 (weakest classifier) to +1 (strongest classifier), is defined as follows (equation 2):

$$\mu = \frac{(n_c\ \mu_c) - (n_m\ \mu_m)}{n_c + n_m} \tag{2}$$

where n_c is the number of correctly classified instances, n_m is the number of misclassified instances and μ_c and μ_m are mean margins for correctly and misclassified instances respectively.

Dominant margin

The dominant margin is based on the mode of the margin histogram (margin value with highest frequency). The margin mode is calculated as follows:

1. Group margin values into bins (10 bins were used in this study): [0,0.1[, [0.1,0.2[, …, [0.9,1]
2. Calculate margin bins histogram
3. Detect the peak of the histogram (the bin with the highest frequency).

A normalised measure of the dominant margin is calculated from two margin modes, one for the margin distribution of correctly classified instances and another for misclassified instances (equation 3).

$$M = \frac{n_c M_c - n_m M_m}{n_c + n_m} \tag{3}$$

where M_c and M_m are the margin modes of correctly classified and misclassified instances respectively.

Margin entropy

Shannon's entropy (Shannon, 1948) was used to measure diversity and redundancy in margin distribution, applied on margin normalised frequency values. The margin entropy is estimated using the following steps:

1. Group margin values into bins
2. Calculate margin bins histogram
3. Normalise the histogram to determine bin (or coarse margin) probabilities for each margin bin
4. Calculate entropy on resulting bin probabilities using Shannon's formula (equation 4)

$$\mathrm{H} = -\sum_{i=0}^{n-1} [P(m_i) * log_2(P(m_i))] \;\; P(m_i) \neq 0 \qquad (4)$$

where P represents probability, m_i $(0 \leq i < n)$ a margin bin and n the number of margin bins.

The minimum value for margin entropy is 0 (lowest diversity). The entropy is maximum when underlying events are equiprobable, i.e. all margin frequencies are the same. For 10 margin bins, the maximum value of margin entropy is 3.32 (highest diversity such as random distribution). Unlike mean and dominant margins, the introduced margin entropy, which expresses diversity in ensemble models at data level, has to be high, but not at its maximum. Indeed, more diversity does not necessarily induce higher classification accuracy (Kuncheva and Whitaker, 2003). The complex relationship between diversity and ensemble accuracy is still not well understood, but diversity is recognised as an indispensable condition in designing effective ensemble classifiers (Kuncheva and Whitaker, 2003).

Margin-weighted confusion matrix

We introduce a novel Weighted Confusion Matrix (WCM) that uses the unsupervised margin of ensemble classifiers. It has been recommended that traditional accuracy estimates should be accompanied by confidence limits, which are rarely provided in published papers (Foody, 2004). This WCM, used together with a confusion matrix, provides a more thorough analysis of multiple classifier performance than a

traditional deterministic approach based solely on a confusion matrix. The Margin-Weighted Confusion Matrix (MWCM) measures the degree of certainty associated with the correctly and misclassified instances in the RF model. The MWCM is calculated as follows:

1. Calculate the unsupervised margin of each instance (Equation 1)
2. Assign to each instance its reference and RF model predicted class labels
3. For every reference-to-predicted class combination (e.g. woodland predicted as woodland, woodland predicted as open, etc.) calculate the sum of margins and the count of instances
4. Populate each confusion matrix cell with the sum of margins divided by the count of instances in the cell (normalisation).

The ideal MWCM is an identity matrix (i.e. $L \times L$ square matrix with ones on the main diagonal and zeros elsewhere), and the closer a model MWCM to this configuration, the stronger the underlying ensemble classifier. The classification confidence of correctly classified instances (MWCM main diagonal cells) should be as high as possible (close to 1). Conversely, the confidence of misclassified instances (MWCM non-diagonal cells) should be as low as possible (close to 0).

4.6. Results and Discussion

4.6.1. Effect of training data imbalance on RF performance

Experiment 1: Binary imbalance

The effect of training data class imbalance on binary classification performance is shown in Table 4-3 and Table 4-4, and in Figure 4-1 and Figure 4-2. Overall accuracy was highest in the balanced case for both optimal (overall accuracy 91.16 %, Kappa 82.32 %) and critical (overall accuracy 84.61 %, Kappa 69.22 %) cases. Results demonstrate that imbalance increases per class accuracy in favour of the majority class, and at the expense of the minority class (Table 4-3), as fewer training instances of the minority class are selected in each bootstrap sample used in tree construction. This effect is more pronounced in the critical case, where very few training samples between highly imbalanced minority cases (e.g. 90% forest to 10%

non-forest) results in large differences in majority and minority class accuracies (59.9 % difference in the critical case versus 30.2 % difference in the optimal case). Nevertheless, the balanced case provides the best (more balanced) pairwise per-class accuracies for both optimal and critical training set sizes. The minimum accuracy per class for an optimal training set size is 90% (80.45% Kappa) versus 83.36% (70.16% Kappa) for the best imbalanced case performance (25:75).

Mean margins, measuring the overall classification confidence, are also highest where training data are balanced (Figure 4-2). Confusion Matrices (CM) and Margin Weighted Confusion Matrices (MWCM), comparing optimal and critical balanced and imbalanced (25:75) experiments, are shown in Table 4-4. The outcome of both class label prediction and underlying uncertainties has significantly evolved in favour of the majority class (non-forest). The non-forest accuracy increased by 4.88% while the forest accuracy decreased by 8.9% (almost a factor of 2). The MWCM shows an increase in non-forest classification confidence of 7% and a decrease in the certainty of misclassifying non-forest as forest by the same amount. However, the class imbalance led to poor classification confidence for the minority class: loss in classification confidence of 14% and increase in uncertainty associated with forest to non-forest confusion of 12%. Hence, the RF performance loss for the minority class (forest) is twice as big as the gain for the majority class (non-forest) in both classification accuracy and associated confidence. These class imbalance effects are again, more pronounced in the critical case (Table 4-4). Unsupervised margin cumulative distribution curves (Figure 4-3) for balanced training data, illustrate the higher degree of uncertainty associated with the correctly classified instances for critical experiment and the lower certainty associated with misclassified instances for the optimal case. Furthermore, the critical margin distribution of correctly classified instances exhibits a significantly higher proportion of low margins (<0.5). Meanwhile, the critical margin distribution of misclassified instances exhibits a lower proportion of low margins. Both reflect poorer behavior of the underlying RF model.

The margin entropy values (equation 4) are higher for critical than for optimal size (e.g. 3.02 vs. 2.61 for the balanced case) (Table 4-3). This is expected as less data increases uncertainty leading to greater diversity. The interpretation of this concept has to be carefully addressed. For instance, for the critical case, the margin entropy

value for ratio 50:50 is already high: 3.02 (maximum at around 3.32 for 10 margin bins). It increases to 3.09 for ratio 25:75. In the meantime, the mean margin decreases (Figure 4-2), reflecting an increase in lower margin frequencies. The increase in margin entropy suggests an initially higher proportion of higher margins at ratio 50:50 which becomes more balanced at 25:75. From ratio 25:75 to 10:90, the entropy decreases while the mean margin continues to decrease. This means that the behaviour is reversed: the proportion of lower margins has become dominant with respect to higher margins. Among the five tested ratios, 25:75 results in maximum entropy (and hence maximum diversity) but it is the balanced case (50:50) which results in maximum accuracy. This is consistent with entropy theory outlined in section 4.6.

Table 4-3 RF model performance results for binary classification imbalance (experiment 1)

Balance, ratio Forest to Non-Forest training samples					
	10:90	25:75	50:50	75:25	90:10
Overall Kappa (%)					
Optimal	60.8	78.24	**82.32**	77.44	66.4
Critical	30.74	62.34	**69.22**	61.02	37.26
Margin entropy					
Optimal	**2.85**	2.78	2.61	2.54	2.52
Critical	2.97	**3.09**	3.02	2.93	2.70
Optimal					
Per-class producer Kappa (%)					
Forest	45.02	70.16	84.28	92.19	**95.18**
Non-forest	**93.6**	88.43	80.45	66.76	50.98
Critical					
Per-class producer Kappa (%)					
Forest	18.68	52.46	77.32	87.66	**93.12**
Non-forest	**86.89**	76.82	62.66	46.8	23.29

Table 4-4 Binary imbalance confusion matrices and margin-weighted confusion matrices for evenly balanced and imbalanced training data in optimal and critical cases

		Observed (test) data			
		Confusion Matrix		Margin-weighted confusion matrix	
	Forest to Non-forest training sample balance	Forest	Non-Forest	Forest	Non-Forest
	Optimal				
	50:50 (Balanced)				

	Forest	1149	127	**0.80**	0.49
	Non-forest	101	1123	0.37	**0.81**
	25:75				
	Forest	1054	70	**0.66**	0.42
	Non-forest	196	1180	0.49	**0.88**
	Critical				
Prediction	50:50 (Balanced)				
	Forest	1095	237	**0.70**	0.47
	Non-forest	155	1013	0.39	**0.71**
	25:75				
	Forest	842	102	**0.53**	0.37
	Non-forest	408	1148	0.47	**0.82**

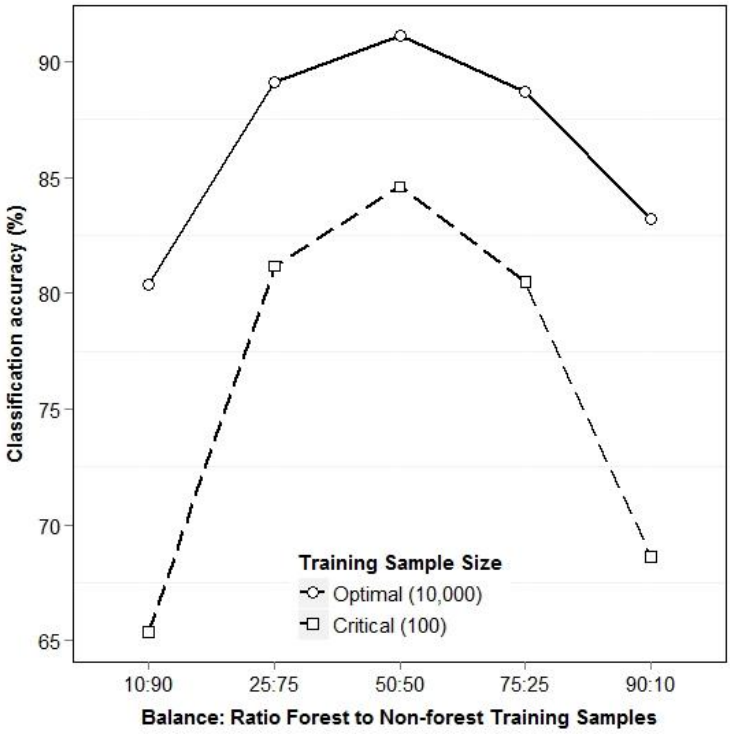


Figure 4-1 Effect of binary class imbalance on overall classification accuracy

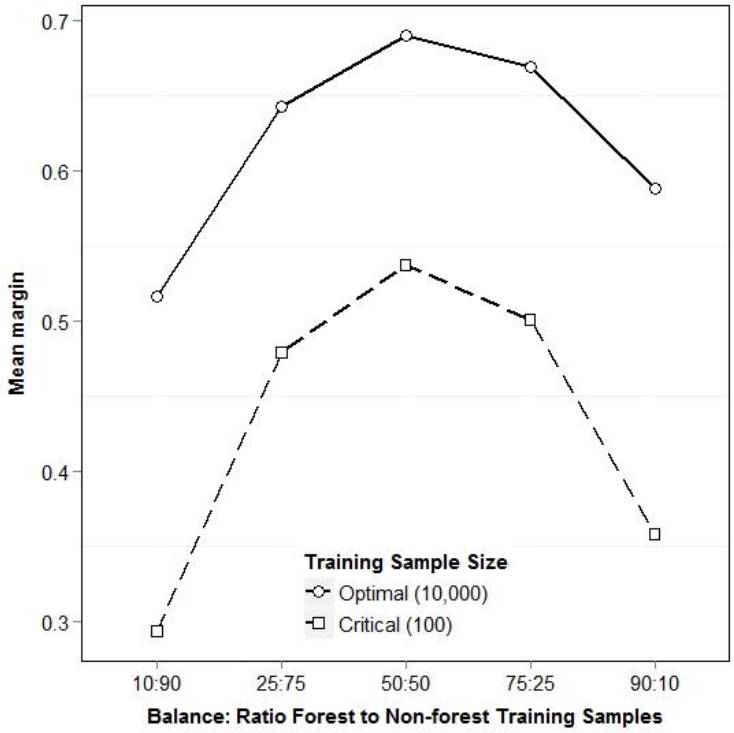


Figure 4-2 Effect of binary class imbalance on mean margin

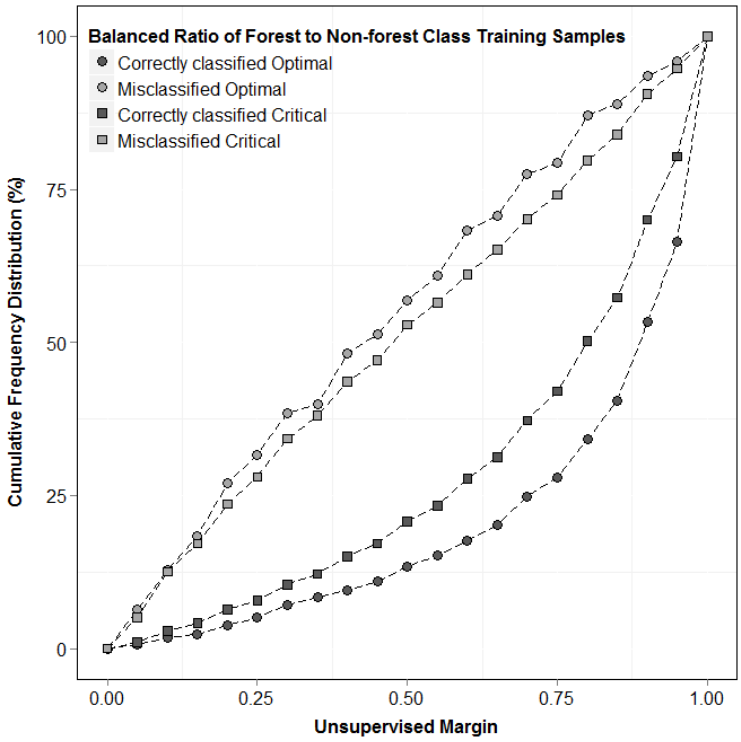


Figure 4-3 Binary classification unsupervised margin cumulative frequency distribution curve, comparing correctly and misclassified instance confidence, for optimal and critical training sizes.

Multiclass imbalance (experiments 2 and 3)

For the multiclass model (experiment 2), in the balanced optimal case, per class (producer) accuracies were highest for the closed canopy cover (89 %), shrub (88.9 %) and Non-forest (87.8 %) classes. Increasing the proportion of the most difficult class (open canopy cover) whilst simultaneously decreasing the proportion of the easiest class (closed canopy cover), improved class accuracy and Kappa for the most difficult (open cover) class (Table 4-5). Compared to the balanced class case, a 25 % increase in open class samples and decrease in closed class samples resulted in 7.5 % gain in producer accuracy for the open class, but a 9 % reduction in producer accuracy for the closed class, leading to more balanced pairwise (Open/Closed) per-class accuracies. However, despite this introduced imbalance, overall accuracy was only marginally affected (1.2 % less than the balanced model). Similarly, this imbalance resulted in only a minor reduction in per class accuracy for the shrub and non-forest classes (1.4 % and 0.3 % respectively for a 25 % imbalance). A 10% increase in open class samples and decrease in closed class samples led to the best

performance (overall producer accuracy). Indeed, this resulted in an increase in the minimum accuracy per class by 1.22 % (1.56 % Kappa) compared to the balanced case, in the optimal experiment (Table 4-5). The same ratio also led to optimal per class accuracies in the critical case. Two-class imbalance (between hardest and easiest classes) had only a minor effect on overall multiclass classification accuracy up to 25 % (particularly in the critical case), above which accuracy drops steeply with increasing imbalance (Figure 4-4). Mean margins show a similar pattern (though less distinct), whereby certainty in the model (i.e. high margins associated with correctly-classified samples together with low margins associated with misclassified samples), drops steadily above 25 % imbalance (Figure 4-5)).

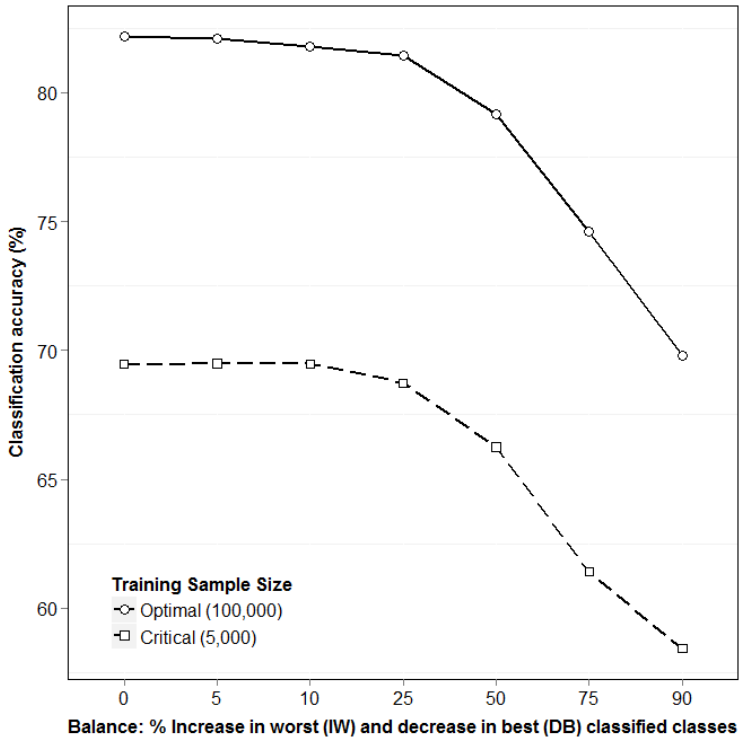


Figure 4-4 Effect of multiclass imbalance on overall multiclass classification accuracy

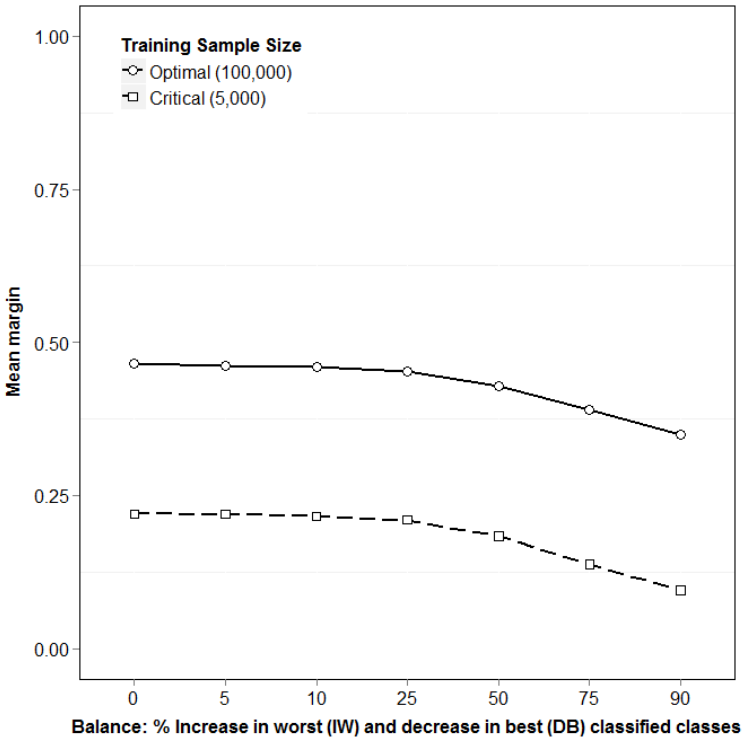


Figure 4-5 Effect of multiclass imbalance on mean margin

Table 4-5 RF model performance results for optimal and critical multiclass classification imbalance experiments

	% Increase worst class (open) and Decrease best class (closed)						
	Balanced	5	10	25	50	75	90
				Overall Kappa (%)			
Optimal	**78.3**	77.64	77.25	76.78	73.96	68.26	62.23
Critical	**62.61**	61.85	61.84	60.91	57.81	51.77	48.04
				Margin entropy			
Optimal	3.24	3.24	3.23	3.24	3.24	**3.25**	**3.25**
Critical	3.19	3.19	3.17	3.17	3.17	3.24	**3.30**
				Optimal			
				Per-class producer Kappa (%)			
Woodland	**69.04**	68.52	67.6	65.89	63.8	60.93	59.97
Open	66.04	66.47	68.21	73.73	81.86	87.52	**88.99**
Closed	**85.96**	83.21	81.72	75.48	59.66	34.97	13.21
Shrub	86.12	**85.5**	84.69	84.48	83.06	83.24	83.33
Non-forest	**84.76**	84.74	84.08	84.42	83.5	83.39	82.84
				Critical			
				Per-class producer Kappa (%)			
Woodland	**49.12**	46.72	46.79	44.78	41.78	38.5	37.46
Open	45.54	47.17	49.36	56.74	68.6	77.34	**79.88**
Closed	**72.42**	70.12	67.25	58.27	38.54	14.3	1.66

Shrub	**70.53**	69.54	69.77	69.16	69.18	68.74	68.56
Non-forest	**76.13**	75.99	75.89	75.52	75.44	75.43	75.35

Margin-weighted confusion matrices (Table 4-6) for multiclass model class imbalance, show the degree of certainty associated with correctly classified and misclassified samples. For each class, the highest normalised margin scores are associated with correctly classified instances (i.e. main diagonal cells). Margins for the most difficult classes (open cover, 0.46 and woodland, 0.5) are lower than the closed cover (0.57), shrub (0.75) and non-forest (0.77) classes. A 50% increase and decrease in proportion of open and closed classes improved model classification certainty in the open class by (0.46 to 0.53) and reduced the certainty of the misclassification of open class predicted as closed cover (0.27 to 0.19). Conversely, the same degree of imbalance reduced certainty in the correct classification of the closed class (0.57 to 0.43). However, an associated *increase* in the certainty of misclassifying closed as open (between balanced and 50% imbalanced), was less (3%) than the decrease (8%) in certainty of misclassifying open as closed. The latter is closer to a potential shift of the RF ensemble from incorrect to correct majority decisions, thereby resulting in a change in the outcome of misclassified instances. The results of experiment 2 demonstrate that introducing a *sensible* imbalance can improve ensemble classifier performances.

Table 4-6 Multiclass imbalance confusion matrices and margin-weighted confusion matrices for optimal case (balanced and 50% imbalanced)

		Confusion Matrix					Margin-weighted confusion matrix				
		Observed (test) data									
		Evenly balanced training samples									
		Woodland	Open	Closed	Shrub	Non-forest	Woodland	Open	Closed	Shrub	Non-forest
Prediction	Woodland	**3714**	449	121	261	265	**0.50**	0.20	0.17	0.21	0.28
	Open	512	**3559**	335	164	129	0.18	**0.46**	0.20	0.19	0.22
	Closed	245	784	**4434**	110	63	0.18	0.27	**0.57**	0.18	0.21
	Shrub	236	73	49	**4379**	168	0.32	0.16	0.13	**0.75**	0.40
	Non-forest	293	135	61	86	**4375**	0.32	0.34	0.28	0.23	**0.77**
		50% increase in Open canopy samples and 50% reduction in Closed canopy samples									
		Woodland	Open	Closed	Shrub	Non-forest	Woodland	Open	Closed	Shrub	Non-forest
Prediction	Woodland	**3478**	234	139	230	237	**0.47**	0.22	0.19	0.20	0.27
	Open	937	**4370**	1430	291	226	0.23	**0.53**	0.23	0.23	0.24
	Closed	61	220	**3282**	27	21	0.18	0.19	**0.43**	0.15	0.20
	Shrub	225	58	64	**4369**	179	0.33	0.14	0.18	**0.75**	0.37
	Non-forest	299	118	85	83	**4337**	0.30	0.35	0.29	0.24	**0.77**

Unsupervised margin cumulative frequency distribution curves, associated with correctly classified and misclassified instances, comparing balanced and imbalanced training data are shown in Figure 4-6 and Figure 4-7. Curves shifting toward the lower right corner of the plot indicate a higher degree of certainty in correctly classified instances and curves moving towards the upper-left indicate a decrease in the certainty of misclassified instances. Figure 4-6 and Figure 4-7 demonstrate that increasing the degree of imbalance (90% increase worst/decrease best) (Figure 4-7) results in a higher degree of divergence from the balanced case, relative to 50% imbalance (Figure 4-6). The margin distribution pair (correct and misclassified) associated with the evenly balanced case exhibits better behaviour than the distribution pair related to the 50% imbalance, the latter located entirely within the

'leaf-like' pattern of the evenly balanced pair of curves (Figure 4-6). However, a different pattern is shown in Figure 4-7 comparing balanced and 90% imbalanced cases. The even leaf-shaped pairwise margin distribution only partially contains the imbalanced pairwise margin distribution curves. While the even distribution associated with misclassified instances behaves significantly better than its imbalanced counterpart, the balanced distribution related to correctly classified samples exhibits lower classification confidence than the imbalanced case. This is due to the fact that only 16 % of the closed samples have been correctly classified in the extremely imbalanced case (versus 89% for the balanced case). Meanwhile, the accuracy of the open class increased by more than 20% relative to the balanced case. Both imbalanced results have a positive impact on the correctly classified imbalanced margin distribution. On the other hand, the very high misclassification rate of the closed class in the imbalanced case adversely affects the imbalanced misclassified margin distribution.

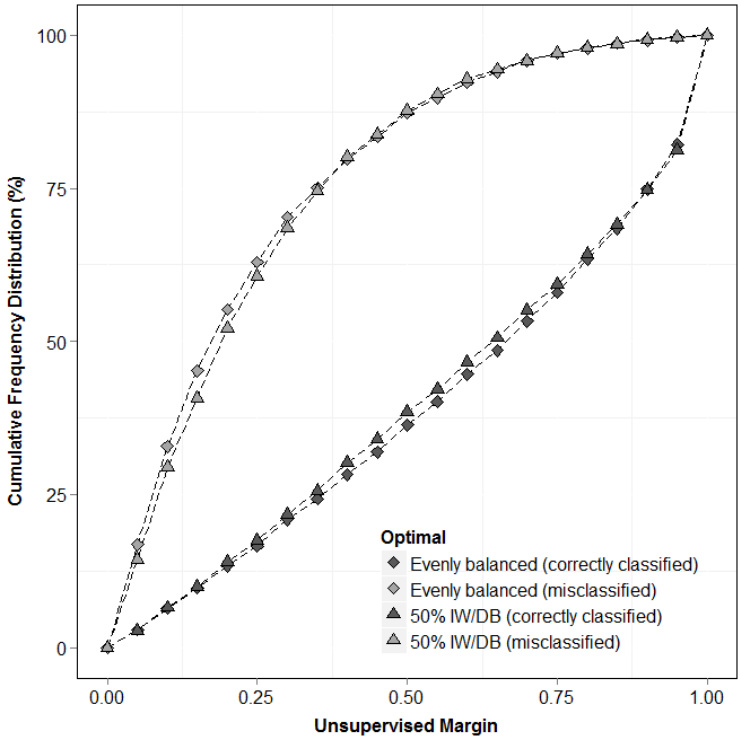


Figure 4-6 Unsupervised margin cumulative frequency distribution curves associated with correctly and misclassified instances, comparing balanced versus 50% increase/decrease open/closed

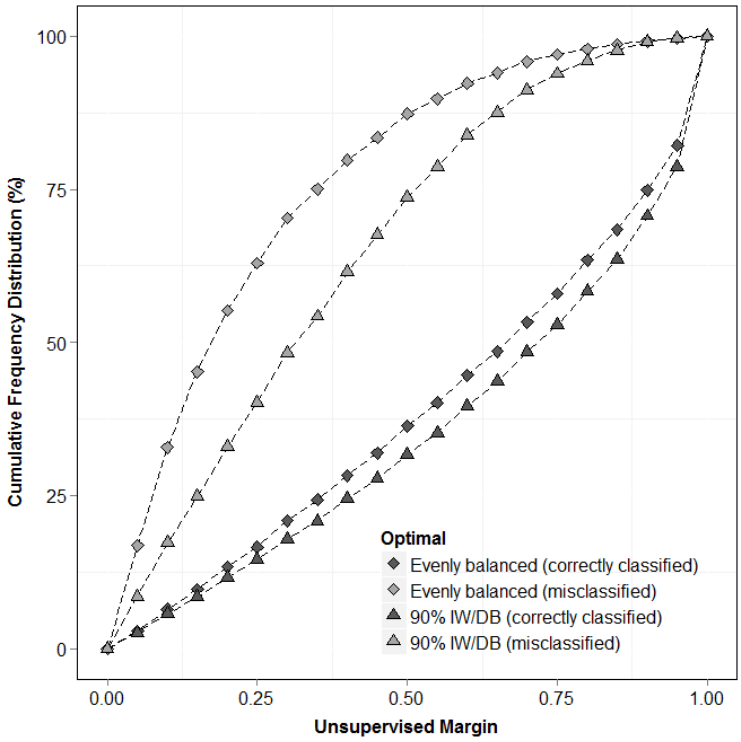


Figure 4-7 Unsupervised margin cumulative frequency distribution curves associated with correctly and misclassified instances, comparing balanced versus 90% increase/decrease open/closed

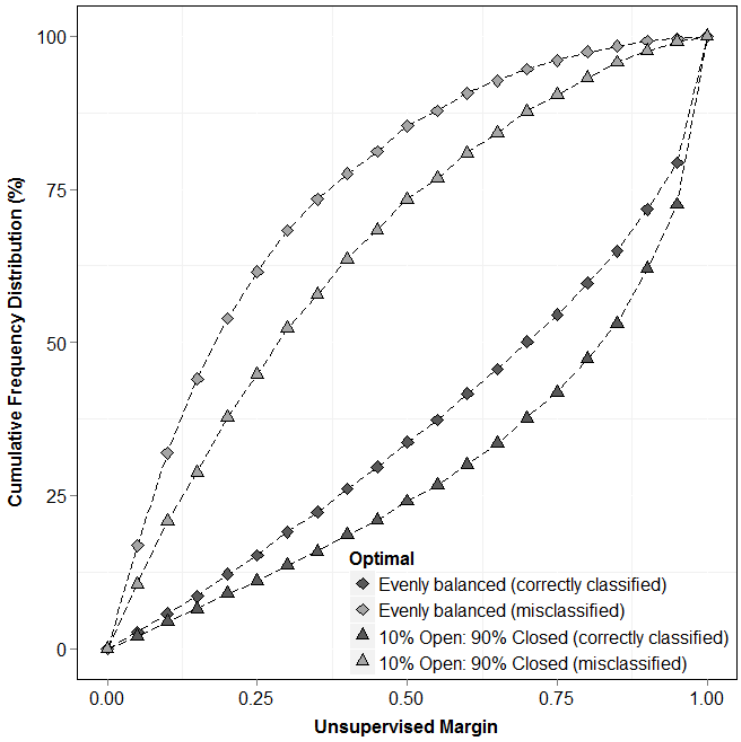


Figure 4-8 Unsupervised margin cumulative frequency distribution curves, comparing balanced and ratio-imbalanced (10 open: 90 closed) for optimal cases

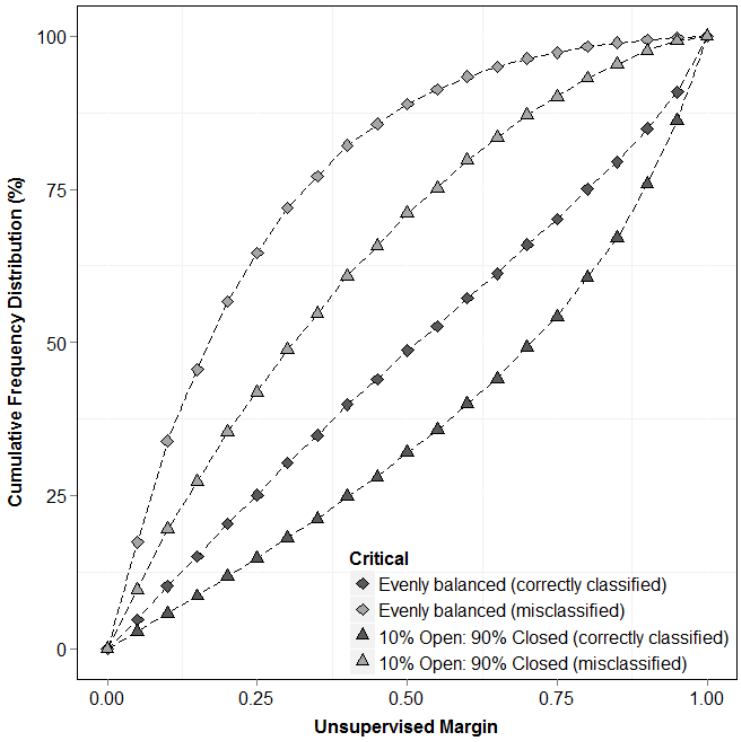


Figure 4-9 Unsupervised margin cumulative frequency distribution curves, comparing balanced and ratio-imbalanced (10 open: 90 closed) for critical cases

Comparing pairwise margin distribution curves for the other multiclass imbalance experiment (ratio of open to closed training instances) shows that, in the critical case (Figure 4-9), the divergence of curves from evenly balanced and imbalanced experiments (both correctly classified and misclassified instances) is more pronounced than in the optimal case (Figure 4-8).

4.6.2. Effect of training data mislabelling on RF performance

Binary mislabelling (experiment 4)

Results of the binary classification class mislabelling experiment are shown in Table 4-7 and Figure 4-10 and Figure 4-11. For the optimal case, the impact of the introduction of mislabeled data (i.e. forest class mislabeled as non-forest and non-forest as forest) on overall classification accuracy was negligible (only a 2.8% reduction from 0% mislabeled to 25% mislabeled). For the smaller (critical) training size, the reduction in overall accuracy was higher (a 6.9% accuracy reduction from 0% mislabeled to 25% mislabeled). The classifier uncertainty statistic (mean margin) showed a greater reduction associated with an increasing number of mislabeled

training instances. Mean margin decreases with each incremental increase in the proportion of mislabeled training instances for both optimal and critical cases (Figure 4-10 and Figure 4-11). Results showing this decline in overall model certainty is supported by a reduction in the margin mode (Table 4-7). For the optimal case, with each increase in the proportion of mislabeled instances, the dominant margin reduced from 0.86 to 0.30 for 25% mislabeled instances. Loss in classification confidence is more pronounced for the critical case.

Table 4-7 RF model performance results for optimal and critical binary class mislabelling experiments

% of Mislabeled training instances (per class)							
	0%	2.5%	5%	10%	15%	20%	25%
Overall Kappa (%)							
Optimal	**81.04**	80.96	81.12	80.08	79.12	78.32	75.36
Critical	**68.42**	67.34	67.26	64.59	63.13	56.65	54.64
Margin entropy							
Optimal	2.67	2.91	3.07	**3.11**	**3.11**	3.04	2.90
Critical	3.18	3.23	3.26	**3.28**	3.26	3.10	2.93
Dominant margin							
Optimal	**0.86**	0.85	0.77	0.67	0.58	0.48	0.30
Critical	**0.79**	0.70	0.70	0.61	0.52	0.34	0.18
Per-class producer Kappa (%)							
Optimal							
Forest	82.96	**83.09**	83.11	82.66	80.87	80.05	78.37
Non-forest	**79.2**	78.94	79.22	77.66	77.45	76.66	72.57
Critical							
Forest	**70.41**	69.22	68.79	66.22	66.94	56.45	56.9
Non-forest	**66.53**	65.57	65.79	63.04	59.73	56.84	52.55

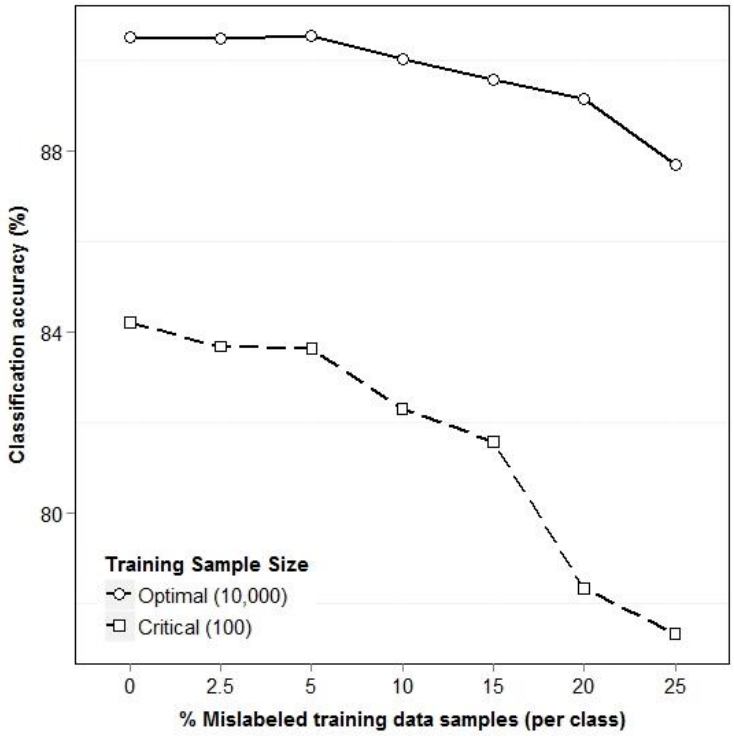


Figure 4-10 Effect of class mislabelling on binary classification overall accuracy

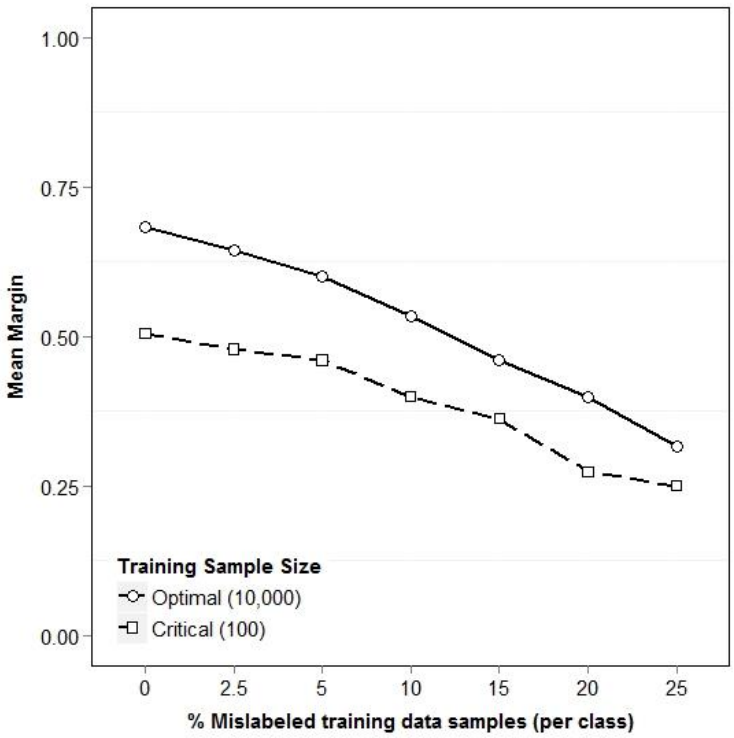


Figure 4-11 Effect of class mislabelling on binary classification mean margin

Multiclass mislabelling (experiment 5)

Between *clean* training data and 25% class mislabelling, overall classification accuracy was reduced by 6.6% for the optimal case and 7.2% for the critical case (Table 4-8). However, associated with this was about a 55% decrease in the mean margin for both optimal and critical cases, which indicates that while training data *noise* has only a minor impact on accuracy relative to the amount of mislabeled training data, it does have a strong influence on classification uncertainty. Dominant margin results for the optimal case support this - from 10% to 25% mislabelling, the dominant margin drops from 0.52 to 0.18 (Table 4-8).

Figure 4-12 and Figure 4-13 show change in classification accuracy and mean margin with mislabeled training instances. By selecting a random subset of training instances to build each decision tree in the ensemble, as well as randomly selecting the features involved in data partitioning at each tree node, RF is robust to noise. This is further demonstrated by comparison to the accuracy of a classifier comprising only a single decision tree constructed using all available training instances (optimal case) and not a bootstrapped sample (reduced by about a third in standard bagging) (Figure 4-12). Results of the single tree classifier show a lower overall accuracy (17% less than the optimal ensemble case) and a steeper reduction in overall accuracy associated with the increasing proportion of mislabeled training instances. Despite a dramatic reduction in training data, the accuracy curve of RF for the critical case behaves significantly better than the single curve. This again, highlights the capability of ensemble classifiers, especially RF, in handling multiple mislabelling classification problems.

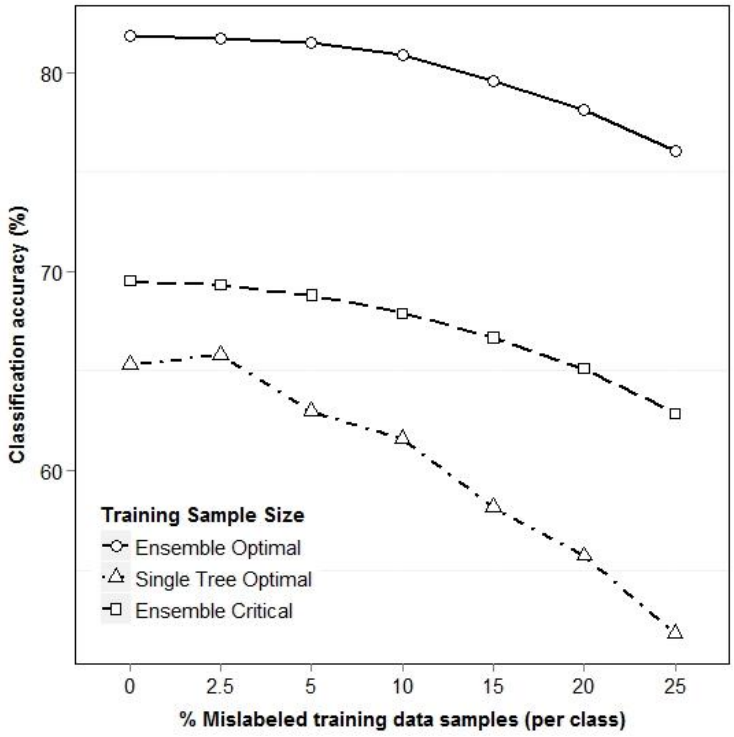


Figure 4-12 Effect of class mislabelling on multiclass classification overall accuracy

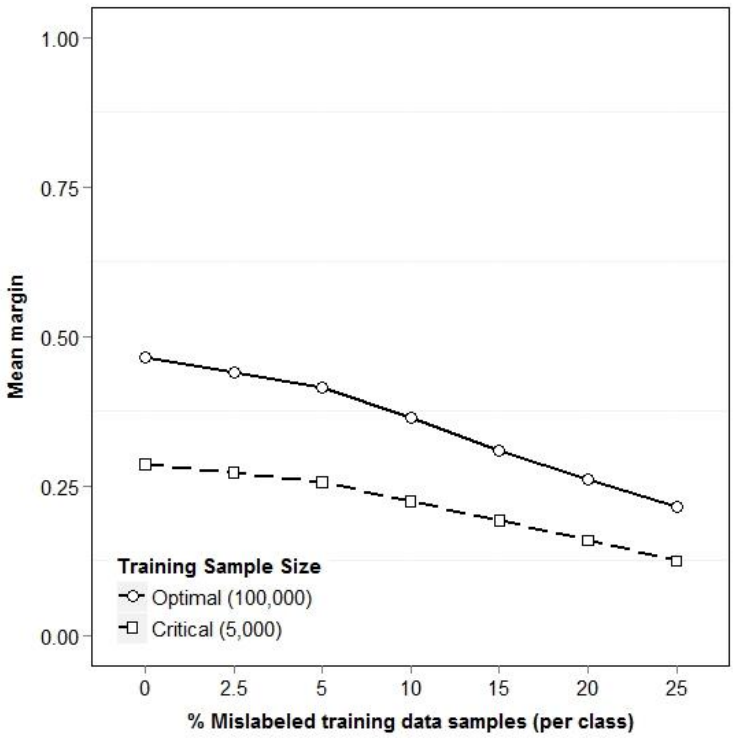


Figure 4-13 Effect of class mislabelling on multiclass classification mean margin

Table 4-8 RF model performance results for optimal and critical size multiclass class mislabelling experiments

	% of Mislabelled training instances in each class						
	0	2.5	5%	10%	15%	20%	25%
	Overall Kappa (%)						
Optimal	**78.3**	77.1	76.8	76	74.4	72.7	70.1
Critical	**62.61**	61.7	61	59.9	58.4	56.4	53.6
	Margin entropy						
Optimal	3.24	3.3	**3.31**	3.25	3.12	2.96	2.79
Critical	**3.23**	3.22	3.18	3.09	2.95	2.79	2.61
	Dominant margin						
Optimal	**0.78**	0.77	0.68	0.52	0.43	0.26	0.18
Critical	**0.65**	0.02	0.02	0.02	0.02	0.02	0.01

Margin weighted confusion matrices (optimal case, Table 4-9, and critical case, Table 4-10) demonstrate the increase in per-class uncertainty (lower classification confidence values) associated with correctly classified instances (main diagonals) with an increasing proportion of mislabeled instances. For the critical case in particular, where a training set is small, the addition of even low levels of class mislabelling can greatly increase ensemble diversity and the likelihood of more randomness in the classifier outcome. At the 25% mislabelling level, both optimal and critical cases have similar margins. Margin entropy values (Table 4-8) are initially (0% artificial mislabelling) very high (close to maximum entropy), reflecting a high diversity in the large area forest data. The margin entropy results should be interpreted with caution. In the optimal case, the margin entropy increases from the original (0% mislabelling) to 5% mislabelling. An increase in entropy with noise is an expected behavior, as uncertainty increases with noise. In the meantime, the mean margin decreases (Figure 4-13), suggesting an increase in lower margins to achieve a more balanced distribution between higher and lower margins, and consequently an increase in entropy. Then, with increasing mislabeled instances (i.e. more noise), entropy decreases as lower margins become over-represented (as illustrated by the decreasing mean margin) inducing an increasing imbalance in margin frequencies.

Table 4-9 Multiclass mislabelling confusion matrices and margin weighted confusion matrices for optimal case

	Confusion Matrix	Margin-weighted confusion matrix
	Observed (test) data	

		5% mislabeled training instances per class									
		Woodland	Open	Closed	Shrub	Non-forest	Woodland	Open	Closed	Shrub	Non-forest
Prediction	Woodland	**3658**	439	104	287	300	**0.45**	0.18	0.14	0.21	0.26
	Open	619	**3590**	414	173	139	0.17	**0.42**	0.17	0.17	0.19
	Closed	253	763	**4370**	108	62	0.19	0.25	**0.50**	0.15	0.20
	Shrub	202	70	45	**4355**	168	0.31	0.13	0.13	**0.67**	0.37
	Non-forest	268	138	67	77	**4331**	0.30	0.32	0.26	0.23	**0.69**
		25% mislabeled training instances per class									
		Woodland	Open	Closed	Shrub	Non-forest	Woodland	Open	Closed	Shrub	Non-forest
Prediction	Woodland	**3352**	359	90	658	634	**0.29**	0.13	0.11	0.15	0.17
	Open	1088	**3677**	782	251	154	0.15	**0.26**	0.12	0.12	0.12
	Closed	240	825	**4051**	104	63	0.12	0.14	**0.28**	0.10	0.11
	Shrub	138	35	22	**3935**	146	0.19	0.10	0.10	**0.38**	0.23
	Non-forest	182	104	55	52	**4003**	0.19	0.19	0.19	0.17	**0.39**

Table 4-10 Multiclass mislabelling confusion matrices and margin weighted confusion matrices for critical case

		Confusion Matrix					Margin-weighted confusion matrix				
		Observed (test) data									
		5% mislabeled training instances per class									
		Woodland	Open	Closed	Shrub	Non-forest	Woodland	Open	Closed	Shrub	Non-forest
Prediction	Woodland	**2850**	693	180	552	461	**0.31**	0.18	0.12	0.19	0.21
	Open	979	**2872**	813	338	169	0.15	**0.27**	0.15	0.13	0.14
	Closed	505	1219	**3869**	294	115	0.19	0.25	**0.40**	0.18	0.15
	Shrub	285	67	58	**3662**	305	0.30	0.10	0.11	**0.63**	0.37
	Non-forest	381	150	79	154	**3950**	0.28	0.30	0.25	0.21	**0.61**
		25% mislabeled training instances per class									

		Woodland	Open	Closed	Shrub	Non-forest	Woodland	Open	Closed	Shrub	Non-forest
Prediction	Woodland	**2576**	602	158	921	971	**0.24**	0.15	0.10	0.16	0.16
	Open	1540	**3092**	1354	503	242	0.14	**0.21**	0.13	0.12	0.11
	Closed	446	1179	**3409**	271	102	0.13	0.16	**0.24**	0.12	0.11
	Shrub	194	30	28	**3208**	254	0.20	0.09	0.07	**0.36**	0.24
	Non-forest	243	98	50	96	**3432**	0.18	0.20	0.17	0.15	**0.36**

Results of the binary (experiment 4) and multi-class mislabelling (experiment 5) experiments support previous research demonstrating RF resistance to mislabelling (Rodriguez-Galiano et al., 2012; Rogan et al., 2008). While results of these experiments are encouraging for noise resistance, ensemble margin statistics (mean and dominant margins) and MWCM reveal mislabelling to have a large effect on classification uncertainty.

Results on the affect of training data characteristics on random forest performance inform the design and implementation of large area land cover classification for natural resource management. The study's findings highlight issues for consideration in the design, training data collection and RF model construction phases of a classification. Furthermore, these are issues which inform where best to allocate limited resources in building a robust and accurate RF classification. While results demonstrate that balanced distributions of training data achieve the greatest overall classification accuracy and certainty, introducing imbalance favouring more difficult to classify classes, can be used to boost per class accuracy without compromising a classification's overall accuracy. Results from mislabelling experiments emphasize the importance of training data labelling accuracy in large area classifications. While the accuracy of random forests is relatively robust to noise in training data, associated model certainty is less so. As such, assuming that increasing that amount of training data collected leads to an increase in the proportion of mislabelled training data instances, it is important to consider an appropriate balance between the amount of training data (per class), class mislabelling, classification accuracy and uncertainty.

4.7. Conclusion

Results from this study provide important insights into the behaviour of the RF ensemble classifier that should provide a guide to the design of an operational implementation in other large area settings, particularly across complex, dynamic and heterogeneous environments. Measures of accuracy and confidence reveal the degree of influence that training data imbalance and mislabelling have on overall and per-class classification performance. The binary and multiclass land cover classification experiments showed the relevance of the introduced ensemble margin criteria and margin weighted confusion matrix for the investigation of both imbalance and mislabelling problems in ensemble classification.

Across large areas with spectrally similar and noisy land cover classes, a degree of training class mislabelling is inevitable. Our findings reveal that while traditional confusion matrices (derived either from independent validation data or an Out-Of-Bag (OOB) sample) can show reasonable classification performance, classification certainty can be significantly reduced, especially where the amount of training data is limited. While previous studies have shown classification to perform better with balanced datasets (Estabrooks et al., 2004; Freeman et al., 2012), we demonstrate that deliberately imbalancing classes can be used to improve the classification and performance of more challenging classes, without significantly compromising overall and other per-class classification results. Given the costs of training data collection (ground-based collection or from high resolution remote sensing data), in an operational setting, optimising a classification involves balancing the total amount, class distribution and labelling accuracy of training data. For example, prioritising rare or more difficult classes over those that are more common or easy to classify. And, where resources limit the amount of available training data, the quality (i.e. correct labelling) becomes a more important consideration.

Future research will investigate methods to address imbalance and mislabelling problems and using margin statistics and the MWCM to evaluate their success.

Chapter 5. Exploring Diversity in Ensemble Classification: Applications in Large Area Land Cover Mapping

Based on the peer-reviewed published article:

Mellor, A. and Boukir, S., 2017. Exploring diversity in ensemble classification: Applications in large area land cover mapping. *ISPRS Journal of Photogrammetry and Remote Sensing*, *129*, pp.151-161.

5.1. Introduction

Across a broad range of applications, ensemble classification systems (also known as multiple or committee classifiers) have been shown to produce better results than single expert systems (Polikar, 2006) and achieve reduced generalization error (Tumer and Ghosh, 1996; Opitz and Maclin, 1999). In remote sensing application areas, such as ecology and natural resource management, ensemble classifiers, like Random Forests (RF) (Breiman, 2001), have become increasingly popular. Incorporating remote sensing data and ancillary continuous and categorical biophysical spatial data, RF has been applied in a variety of large area land cover (Rodriguez-Galiano et al., 2012) and forest attribution studies, including biomass (Baccini et al., 2008), canopy height (Wilkes et al., 2015), canopy cover (Mellor et al., 2015) and species (Evans and Cushman, 2009; Dalponte et al., 2013). The RF classifier builds an ensemble of decision trees (known as base classifiers or ensemble members) and assigns classification through voting or averaging among these ensemble members.

Diversity between ensemble members is considered a key factor affecting overall classification performance (Kuncheva and Whitaker, 2003; Ham et al., 2005; Melville and Mooney, 2005; Kapp et al., 2007). Ensemble classifiers which achieve higher overall classification rates are those in which misclassified instances (errors) made by ensemble members are uncorrelated (Banfield et al., 2005; Elghazel et al., 2011). Ensemble classifiers are often more accurate than their component (base) classifiers, and diversity is greater, if errors made by ensemble members are uncorrelated (Hansen and Salamon, 1990; Díez-Pastor et al., 2015) and more uniformly distributed (Banfield et al., 2005). While ensemble diversity has been studied in the fields of information science and machine learning, to the best of our knowledge, the relationship between ensemble diversity and classification performance has not been actively explored in remote sensing. Gaining a greater insight into the role of diversity in ensemble classification is important, not least because of the increasing popularity of ensemble classifiers, such as random forests in this field (Belgiu and Drăguţ, 2016). Moreover, while advances in remote sensing science and technology (such as new sensors and image analysis techniques) seek to address land cover mapping (classification) error, the availability of suitable

reference (training and test) data is a fundamental requirement in supervised image classification (Foody et al., 2016). Training and test data are also expensive (Pflugmacher et al., 2012), and as such, there are significant benefits to designing classifiers which make more efficient use of training data, such as reducing class information redundancy and maximizing the application of training data for classes which are rarer or more difficult to classify.

In this paper, we explore the relationship between ensemble diversity and classification performance in the context of large area land cover classification across complex forest ecosystems and topography, using remote sensing and ancillary spatial data. We focus on the relationship between ensemble diversity and ensemble margin, two fundamental theories in ensemble learning. Applying the RF classifier, we evaluate different ways of inducing diversity in ensemble classification to improve classification performance and efficiency, and reduce training data redundancy. The main novelty of our work is on *boosting* diversity by targeting lower margin training samples (which represent class decision boundaries or more difficult or rarer classes) in the learning process. We also propose a new empirical analysis that explores the influence of tree pruning, and decision tree depth, on diversity, which leads to a better understanding of RF classifier performance. The findings of this work may be used to inform training data collection strategies and to design more efficient classification. Key concepts used in the paper are introduced in sections 5.2 through 5.4. Section 5.5 describes the study area and data, and experiments, results and discussion are included in sections 5.6 through 5.7.

5.2. Random Forests

Random forests (Breiman, 2001) is a popular ensemble classifier (Belgiu and Drăguț, 2016), which generates decision trees using sub-sets of bootstrap-aggregated training data (sampling with replacement), otherwise known as bagging. These decision trees represent diverse base classifiers, which are combined into an ensemble. In addition to bagging, diversity is induced through the random selection of a sub-set of input (explanatory or predictor) variables which are evaluated for partitioning data at each decision tree node (Elghazel et al., 2011). A response variable is predicted as a modal vote (for categorical data) or average (for continuous variables) among the ensemble

decision trees. Studies have reported that the number of variables randomly sampled to split training data at decision tree nodes does not affect classification rates (and other RF performance measures) (Cutler et al., 2007).

5.3. Ensemble Margin

As demonstrated in previous chapters of this research, the margin provides a measure of confidence in ensemble classification (Guo et al., 2011; Mellor et al., 2014, 2015) and is an important concept in ensemble methods (Schapire and Freund, 1998). The ensemble margin is calculated as the difference between the number of votes assigned to different classes by the base classifiers in an ensemble. The unsupervised version of Schapire's margin (equation 1) of a sample x is the difference between the number of votes (respectively V_{c1} and V_{c2}) assigned to the first and second most popular classes (respectively c_1 and c_2), normalised by the number of base classifiers (T) in the ensemble, regardless of true class labels (Guo and Boukir, 2013). It has been used in large area remote sensing classification as an ancillary measure of random forest classifier performance (Mellor et al., 2014, 2015).

$$margin(x) = \frac{V_{c1} - V_{c2}}{T}, \quad 0 \leq margin\,(x) \leq 1 \tag{1}$$

Correctly classified training instances with high margin values (i.e. close to 1) represent instances located away from class decision boundaries and can contain a high degree of redundant information in a classification problem. Conversely, training instances with low margin values (i.e. close to 0) are located near decision boundaries and are more informative in a classification task. Unlike Schapire's margin (Schapire and Freund, 1998), which is supervised and calculated as the difference between votes assigned to the true class and those assigned to the most voted class that is different from the true class, class labels in the unsupervised margin (Guo and Boukir, 2013) (applied in this study) are not of significance. As such, the unsupervised margin may be more robust to noise (e.g. incorrect class labels) (Guo, 2011). The mean margin (equation 2) is a descriptive statistic for the ensemble margin, calculated from the unsupervised margin values (equation 1), which can be used as a confidence measure for model performance (Mellor et al.,

2014, 2015). This measure ranges from -1 (weakest ensemble classifier) to +1 (strongest ensemble classifier).

$$\mu = \frac{(n_c\,\mu_c) - (n_m\,\mu_m)}{n_c + n_m}\,, -1 \leq \mu \leq 1 \qquad (2)$$

where n_c is the number of correctly classified instances, n_m is the number of misclassified instances, μ_c and μ_m are mean margins for correctly and misclassified instances respectively.

5.4. Ensemble diversity

Ensemble diversity is important for majority vote accuracy and aims at decreasing the probability of identical errors (correlation between ensemble members). While it is accepted that diversity improves overall ensemble classification performance, there is no general agreement on how it should be quantified or dealt with (Kapp et al., 2007), nor is there a widely perceived concept of diversity or theoretical framework which supports the development of methods to capture diversity among classifiers (Bi, 2012). A review by Kuncheva and Whitaker (2003) compared ten measures of pairwise and non pairwise diversity, finding most to be highly correlated. In pairwise measures, the diversity values between all pairs of classifiers are initially calculated. The overall diversity measure value is then computed as the mean of all pairwise values. Unlike pairwise measures, non-pairwise measures are calculated by counting a statistical value of all ensemble classifiers to measure the whole diversity. Therefore they generally run much faster than pairwise measures (Guo, 2011). Diversity can be measured at the output (prediction) level, the input (training data) level and at the structure or parameter level (Guo and Boukir, 2014). In this study, we measure diversity at the output level (i.e. diversity among the class labels assigned across each of the base classifiers in the ensemble), using *KW (Kohavi and Wolpert) variance* (Kohavi and Wolpert, 1996), a popular non-pairwise diversity measure, which can be expressed as equation (3) (Kapp et al., 2007).

$$KW = \frac{1}{NT^2}\sum_{j=1}^{N} t\,(x_j)\,(T - t(x_j))\,, 0 \leq KW \leq 0.25 \qquad (3)$$

where diversity increases with KW variance, T is the size of the ensemble of classifiers, $t(x_j)$ is the number of classifiers that correctly recognise sample x_j, and N represents the number of samples.

The minimum value for KW diversity is 0 (lowest diversity), which occurs when all the T ensemble members correctly classify all of the samples (overall accuracy of 100% and mean margin μ of 1), or conversely, when all of the T ensemble members misclassify all of the samples (overall accuracy of 0% and negative mean margins μ ranging from -1, in binary classification, to 0). KW Diversity is maximised (KW = 0.25) when half of the T ensemble members correctly classify each of the samples, and mean margin μ ranges from 0 to 0.5 (in the case of binary classification). In this case, underlying events are equiprobable i.e. the probability of an instance being correctly classified and misclassified are the same, such as in random prediction.

A good diversity measure would have the ability to find the extent of diversity among classifiers and estimate the improvement or deterioration in accuracy of individual classifiers when they have been combined (Bi, 2012). An optimal ensemble classifier achieves the right balance between the accuracy of base classifiers and the diversity of the ensemble. Over-fitting can occur if diversity is too low and there is too much correlation between base classifiers. Too much diversity however, can reduce the accuracy of the ensemble. For example, an ensemble classifier with random prediction has the highest diversity but the lowest accuracy. This accuracy-diversity trade-off will be investigated in this study. An emphasis is placed on analysing the relationship between diversity and ensemble margin which play a key role in majority vote performance.

5.5. Study Area and Data

The experiments study area covers about seven million hectares of diverse dry-sclerophyll dominated public forests in Victoria, Australia. This area is characterised by varied topography and a range of climate zones. Classification predictor variables include remote sensing data (Landsat TM and MODIS), derived texture indices, elevation, slope, aspect and biophysical climate data. Landsat TM data - frequently applied in studies for forest type mapping and canopy cover assessment (e.g. Boyd and Danson, 2005) - comprises a mosaic of nineteen scenes, captured between

February and March 2009, coinciding with the time of training and test data land cover mapping. High sun angles during the summer period of Landsat data acquisition minimised shadow and terrain artifacts in the imagery, and enhanced spectral differences between overstorey evergreen vegetation and more seasonally dynamic understory vegetation (Mellor et al., 2013). Landsat TM scenes were processed to standardised surface reflectance (Flood et al., 2013), reducing inter-scene variation due to atmospheric conditions, topography, sun angle and sensor location. A single standard deviation raster surface was extracted from an annual twenty-three image multi-temporal stack of 16-day MODIS NDVI mosaics (Paget and King, 2008) - this was used to represent phenological variance over a calendar year across the study area.

To characterise vegetation regions which can appear spectrally similar, but have different spatial patterns, textural indices were included as variables in the model. Texture indices have been shown to improve classification performance (Kayitakire et al., 2006; Rodríguez-Galiano et al., 2011). First order texture measures of variance and entropy (Haralick, 1979) were generated for 3x3 and 5x5 cell neighbourhood moving windows, from a grey-scaled (8-bit) Landsat TM derived Normalized Difference Vegetation Index (NDVI). Textural indices were designed to capture textural variance of the study area's forested ecosystems (Mellor et al., 2013).

Topographic and biophysical data were used in the classifier to capture species-environmental relationships, which are key information to geographical modeling (Guisan and Zimmermann, 2000). Vegetation composition is expected to occur in locations with similar soils, topography and climate (Franklin, 1995), and bioclimatic maps provide information about the climatic influence on the distribution of different forest types (Beaumont et al., 2005). Elevation, slope and aspect data were derived from a 30m Digital Elevation Model (DEM) (CSIRO, 2011). The DEM was also used to generate precipitation, temperature, radiation and moisture climate prediction surfaces using BIOCLIM in the ANUCLIM (v 5.1) software package (Houlder, 2001) - a description of the BIOCLIM process can be found in Beaumont et al., (2005).

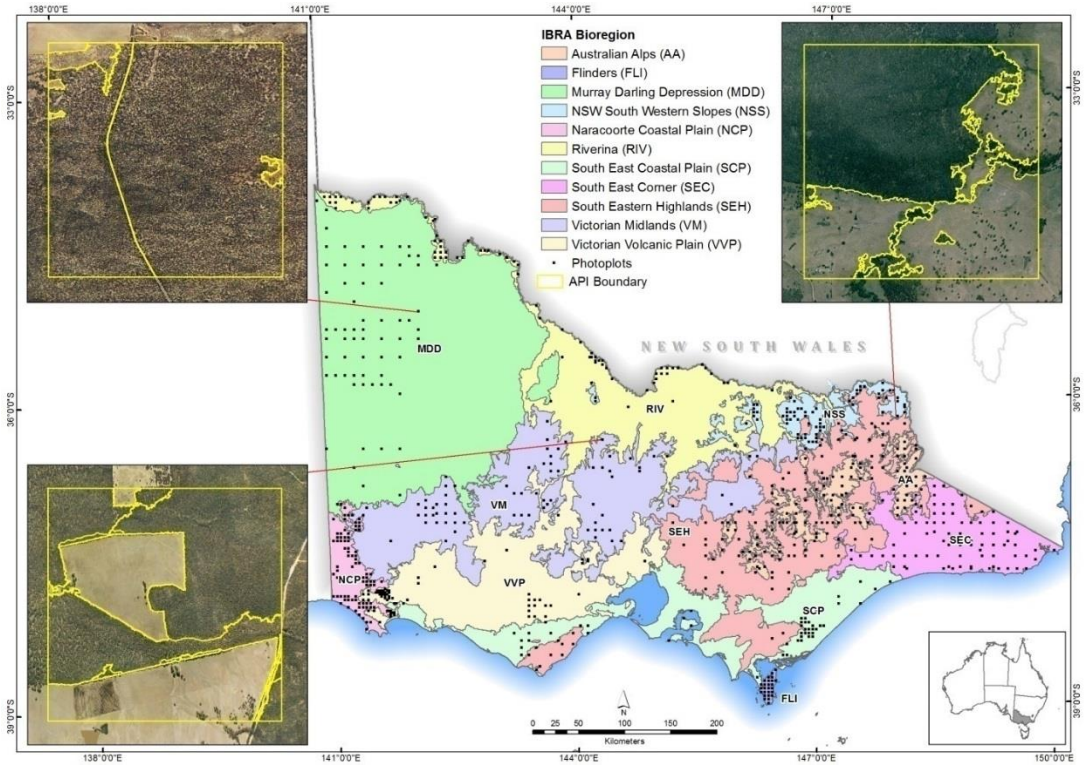
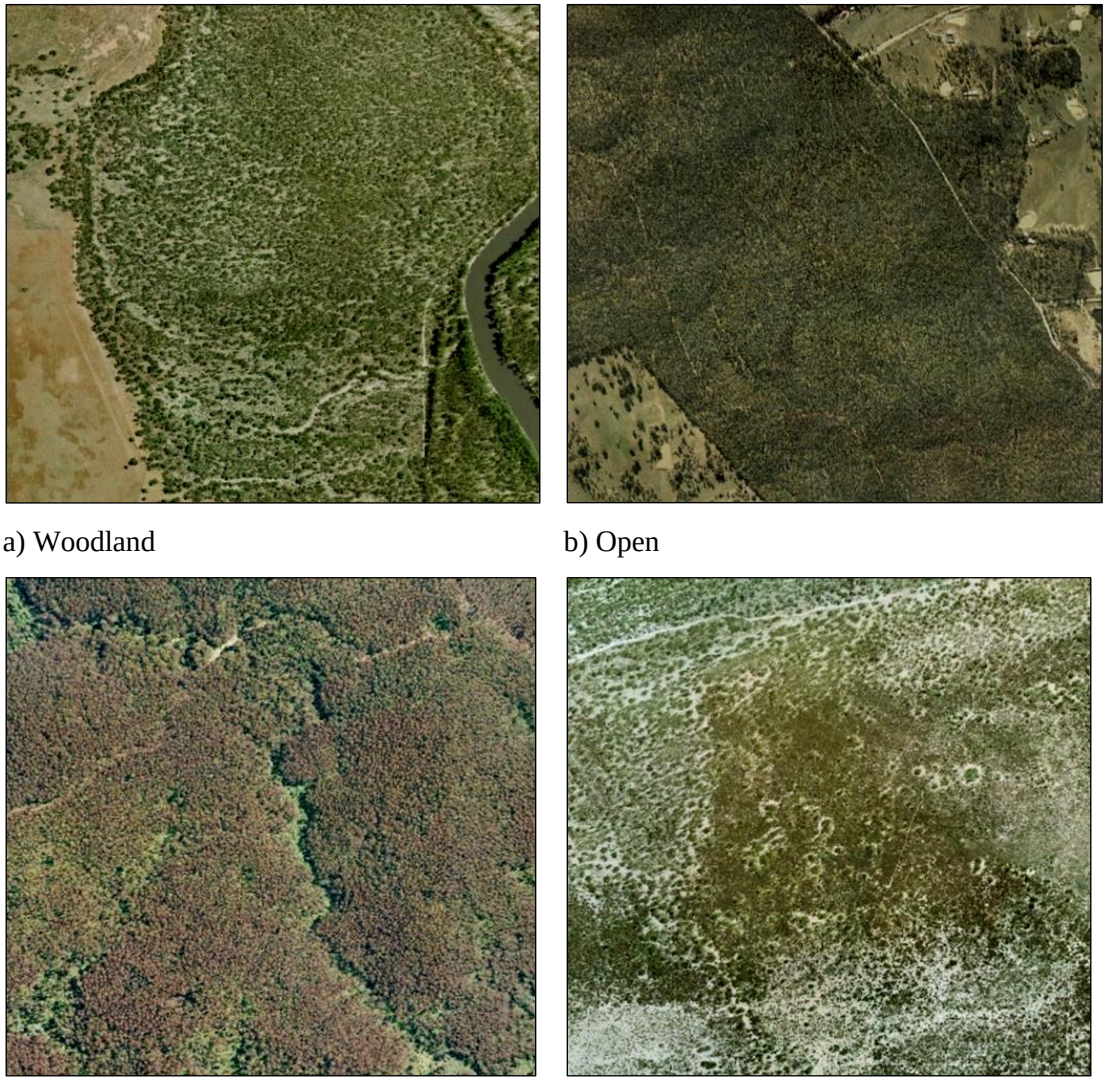


Figure 5-1 Study area map: Victorian Interim Biogeographic Regionalisation for Australia (IBRA Bioregions) and Aerial Photographic Interpretation (API) land cover maps.

Classification reference (training and test) data were derived from seven hundred and sixty-six 2×2 km digital aerial photograph interpreted (API) land cover maps, systematically distributed across a state-wide random stratified grid (Figure 5-1) from imagery acquired between 2006 and 2010. Trained interpreters delineated land cover classes based on information which included crown-shape, colour, shadow and size. A land cover classification system was applied based on Mellor and Haywood, (2010), which included broad forest or other land cover types, three forest canopy height classes (low, medium and tall) and three canopy cover classes (woodland, open and closed). The forest definition applied followed the Australian National Forest Inventory (Department of Agriculture Fisheries and Forestry, 2012), whereby forest is defined as having a greater than 20% crown cover and a minimum stand height of two metres. A half hectare minimum mapping unit was also applied to land cover maps, following UNFAO forest definition (Food and Agriculture Organization of the United Nations, 2001). A detailed description of the land cover reference data methodology can be found in Farmer et al. (2013).

For this study, land cover data were aggregated into three broad canopy cover classes (woodland, open, closed) and two non-forest classes (shrub and non-forest). Examples of canopy cover classes in aerial photography are shown in Figure 2. Land cover polygons were converted to raster and combined with the classification predictor variables. Following Mellor et al. (2015), reference data were divided into training and test subsets, comprising 100,000 (20,000 per class) and 25,000 (5,000 per class) samples respectively.



a) Woodland b) Open

c) Closed d) Shrub

Figure 5-2 Aerial photography examples of forest canopy cover used in the multiclass classification (a) Woodland, 20-50% canopy cover; b) Open, 51-80% canopy cover; c) Closed, >80% canopy cover; d) Shrub (land cover dominated by woody vegetation shrub species, up to 2 m in height). Scale various around 1:25,000

5.6. Experiments

Three experiments were performed using the RF algorithm and assessed using measures of overall and per-class accuracies, Kappa coefficient, ensemble margin and KW diversity. The experiments were designed to explore the influence of, and relationship between, ensemble diversity and classification performance. The main originality of this empirical analysis lies in how the ensemble margin is explicitly involved in the learning process, to induce greater diversity in the ensemble and influence its performance. The *randomForest* package (Liaw and Wiener, 2002) in R (R Development Core Team, 2011) was used to build the RF models and run experiments. Following our previous work (Mellor et al., 2015), 150 base classifiers (decision trees) were used in each experiment. Training data were used to calculate unsupervised margin values then mean margin. Test data were used to calculate RF model overall and per-class accuracies, Kappa statistic and KW diversity. Overall accuracy was first calculated for each individual ensemble base classifier before being combined to calculate ensemble accuracy, ensemble margin and KW diversity for the ensemble. To more clearly illustrate results, all diversity values were normalised, to range from 0 to 1. Calculated Kappa coefficients (Carletta, 1996) also range from 0 to 1.

5.6.1. Experiment 1: Influence of the number of predictor variables on diversity and margin

The number of variables randomly sampled as candidates to partition training data at each decision tree node (hereafter referred to as *mtry* from the *randomForest* R package) was adjusted to evaluate the parameter's effect on classification performance and diversity. For this experiment, starting with two, *mtry* was increased (in single increments) for each RF ensemble model, up to 17 (the maximum number of predictor variables available). Classification accuracy, Kappa statistic, mean margin and KW diversity were calculated for each ensemble.

5.6.2. Experiment 2: Training margins and high diversity data selection

The second experiment constitutes the major contribution of this exploration of ensemble diversity - by investigating a new means of inducing diversity in ensemble learning. This consists of emphasizing the role of lower margin samples in the learning process at the expense of highest margin samples, the latter having the least influence on diversity and ensemble classification performance. For this experiment (Figure 5-3), the unsupervised margin (equation 1) was first calculated for each training data instance as the difference between the maximum number of decision tree (ensemble member) votes assigned to a class minus the number of votes assigned to the second most voted for class, by the ensemble. Percentile distributions were then calculated from the unsupervised margin values of the training set. RF classifications were run on sub-sets of the original training set using only training instances in the bottom (lowest margins) and top (highest margins) 50th, 60th, 70th, 80th and 90th percentiles to build RF models, as well as all training instances. For each ensemble, the mean of the individual ensemble members overall and per-class accuracy and Kappa statistic, the ensemble overall and per-class accuracy and Kappa statistic, and KW diversity, were calculated. These results were compared to ensemble classifiers generated using random subsets (50%, 60%, 70%, 80% and 90%) of all available training instances.

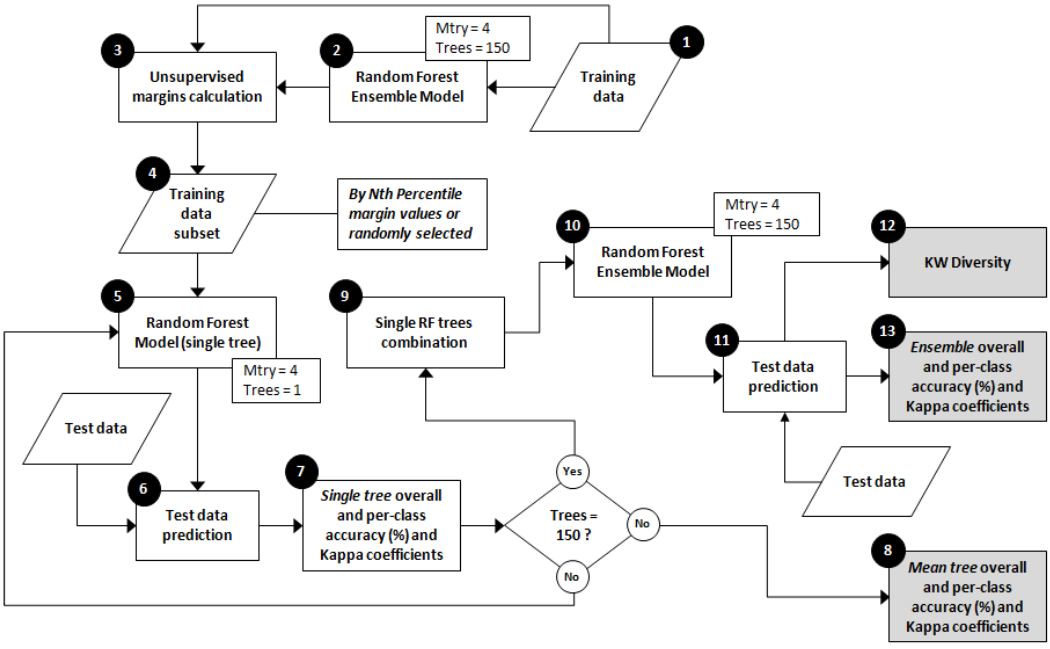


Figure 5-3 Flow chart illustrating training margins experiment (2)

5.6.3. Experiment 3: Influence of the minimum node size on diversity

The last original empirical analysis aims to investigate the influence of tree pruning (and therefore decision tree depth) on diversity for a better understanding of ensemble performance in general, and RF performance in particular. The minimum node size is a model parameter used to control the minimum size of terminal nodes in each decision tree, and therefore, the depth of decision trees. By default in the RF package (and the other experiments applied in this study), the minimum node size is set to 1. In this experiment (Figure 5-4), the minimum node size was increased for each RF ensemble model (from 1 up to 250) and ensemble and mean base classifier accuracies, Kappa statistics and diversity were calculated for each.

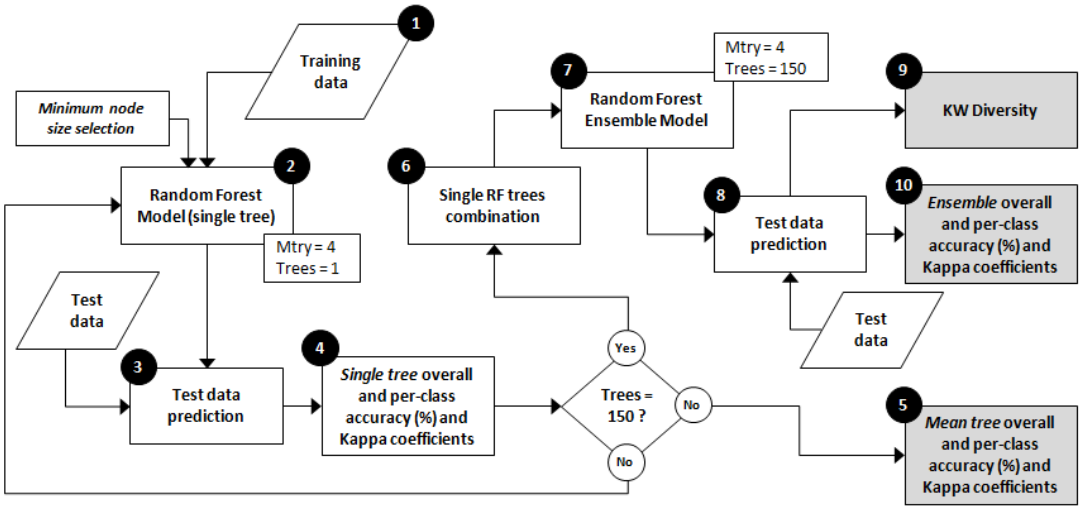


Figure 5-4 Flow chart illustrating minimum node size experiment (3)

5.7. Results and Discussion

5.7.1. Influence of the number of predictor variables on diversity and margin

Figure 5-5 and Table 5-1 show the results of experiment 1. These results show that diversity decreases as the number of predictor variables selected for decision tree splitting (*mtry*) increases. Indeed, the fewer the variables assessed for node splitting, the greater the amount of introduced uncertainty and the higher the diversity achieved (as shown by the mean margin). Increasing the number of predictor

variables assessed at each node split increases classification confidence (Guo and Boukir, 2014).

Table 5-1 Mean tree, ensemble accuracies (%) and Kappa statistic results for the number of predictor variables experiment

Mtry	Mean Tree Accuracy (%)	Mean Tree Kappa	Ensemble Accuracy (%)	Ensemble Kappa
2	65.48	0.57	80.95	0.76
3	67.11	0.59	81.82	0.77
4	68.13	0.60	82.21	0.78
5	68.65	0.61	82.43	0.78
6	69.21	0.61	82.93	0.78
7	69.39	0.62	82.88	0.78
8	69.74	0.62	83.00	0.79
9	69.87	0.62	83.11	0.79
10	70.11	0.63	83.14	0.79
11	70.13	0.63	83.21	0.79
12	70.33	0.63	83.10	0.79
13	70.34	0.63	82.94	0.79
14	70.48	0.63	82.92	0.79
15	70.58	0.63	82.89	0.79
16	70.65	0.63	82.70	0.79
17	70.63	0.63	82.84	0.79

The ensemble and mean individual decision tree classification accuracies increase marginally with increasing *mtry*. Above an *mtry* value of 5, overall ensemble and mean base classifier accuracies are stable (83.0%, 0.79 Kappa, and 70.2%, 0.63 Kappa respectively). Note that a standard RF model would use 4 node split variables (*mtry* = $\sqrt{17}$), which, applied here, does not result in the highest overall ensemble classification accuracy.. Overall classification accuracy and Kappa coefficient by *mtry* are shown in Table 5-1.

While the mean single tree accuracy is reduced with less variables (and uncertainty is higher), the difference between overall (ensemble) and single tree accuracies is greater for 2 variables than for the maximum 17 variables (15.5% and 12% respectively). This illustrates how a loss in tree accuracy and uncertainty associated with a low number of variables is compensated for by higher diversity which influences classification performance.

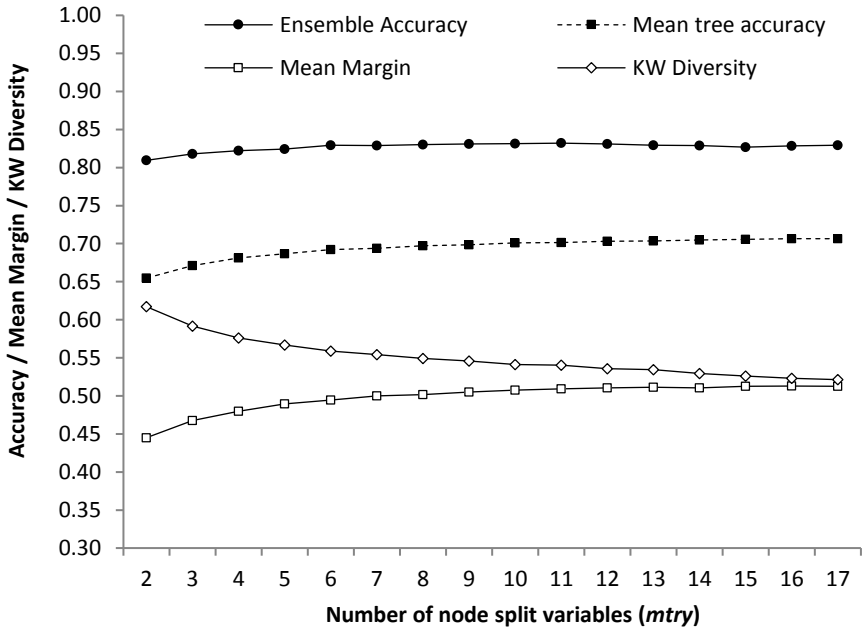


Figure 5-5 Ensemble and mean base classifier accuracies, mean margin and KW diversity plotted against mtry

5.7.2. Training margins and high diversity data selection

Figure 5-6 to Figure 5-8 show results from experiment 2, the mean base classifier accuracy, ensemble accuracy and normalised KW diversity as a function of training set size, selected by training instances in the bottom (lowest margins) and top (highest margins) 50th to 90th percentiles, and randomly selected training instances (equivalent proportions of the total training set). The x-axis on Figure 5-6 to Figure 5-8 ranges from 50 to 100, and represents the margin percentile (in the case of margin-based training data selection), and the proportion of the training set size (in the case of random training data selection). For example, the bottom 50th margin percentile training data sub-set is the same size as the randomly sampled 50% training set.

Table 5-2 shows mean tree and ensemble accuracies (%) and Kappa results for the training margin experiments. Lower margin models (using training samples with margin values in the *bottom* 50th, 60th, 70th, 80th and 90th percentiles) result in lower mean decision tree accuracies compared to higher margin models (using training samples with margin values in the top 50th to 90th percentiles) (Figure 5-6). This is especially true when comparing the top and bottom training instance margin models in the 50th to 70th percentile range.

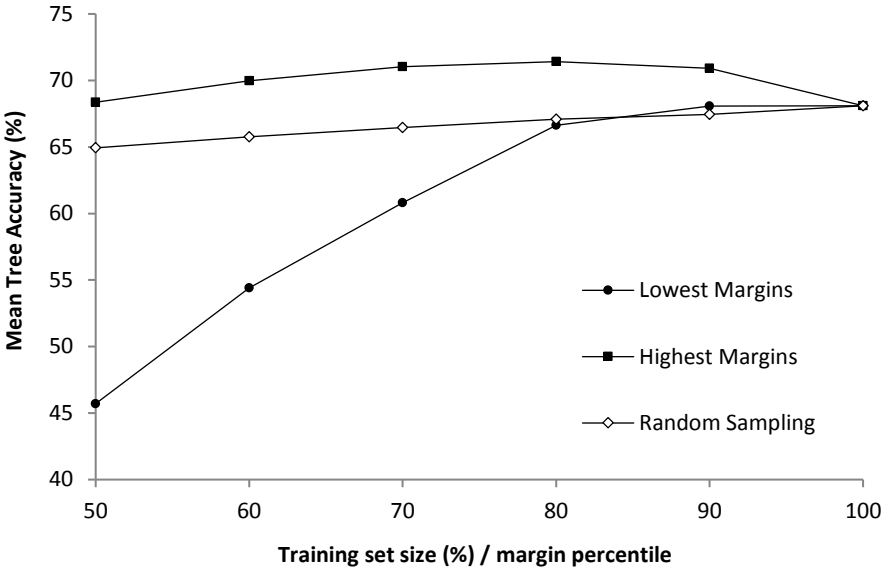


Figure 5-6 Mean tree accuracy as a function of training set size by lowest and highest unsupervised margins, and random sampling

Highest margin generated models (50th to 90th percentiles) exhibit the highest mean tree accuracy (Figure 5-6), but apart from the 50th margin percentile case, return the poorest ensemble accuracies compared to equivalent training set size models from bottom margin percentiles and random sampling (Figure 5-7). It is worth highlighting that for the 70th lowest margin percentile, the overall accuracy achieved is the same as that of the entire training set. Hence, the 30% highest margin samples that have been discarded from the training set are redundant. Redundancy not only slows down the training task, it also weakens bagging performance, affecting the rarer and most difficult classes. The lowest margin training sample selection approach minimises data redundancy.

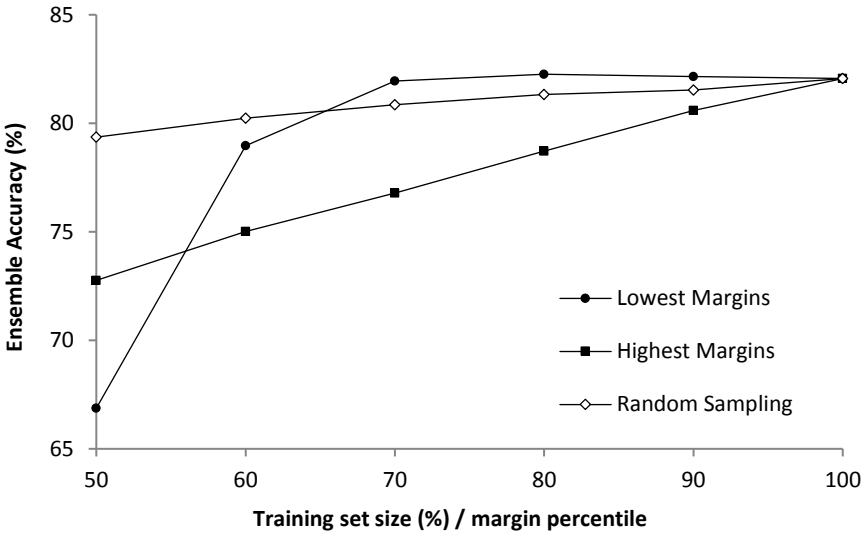


Figure 5-7 Ensemble accuracy as a function of training set size by lowest and highest unsupervised margins, and random sampling

Models generated from training instances in the bottom 70th, 80th and 90th margin percentiles achieve the best ensemble accuracy (Figure 5-7). Figure 5-8 shows that low margin sampling models also exhibit the highest diversity (close to maximum diversity for the 50th lowest percentile) compared to random and highest margin sampling models. Diversity for lowest margins and random sampling converge at the 90th lowest percentile and 90% training set size models. The strength of the RF ensemble bagging approach to induce diversity is underscored by the relative stability of the mean tree (Figure 5-6) and ensemble accuracy curves (Figure 5-7) for random sampling models by training set size, even when only half of the training data are used, particularly in comparison to the low and high margin sampling cases. Indeed, bootstrap sampling (Efron and Tibshirani, 1994) is a robust and effective approach that is suitable for small datasets.

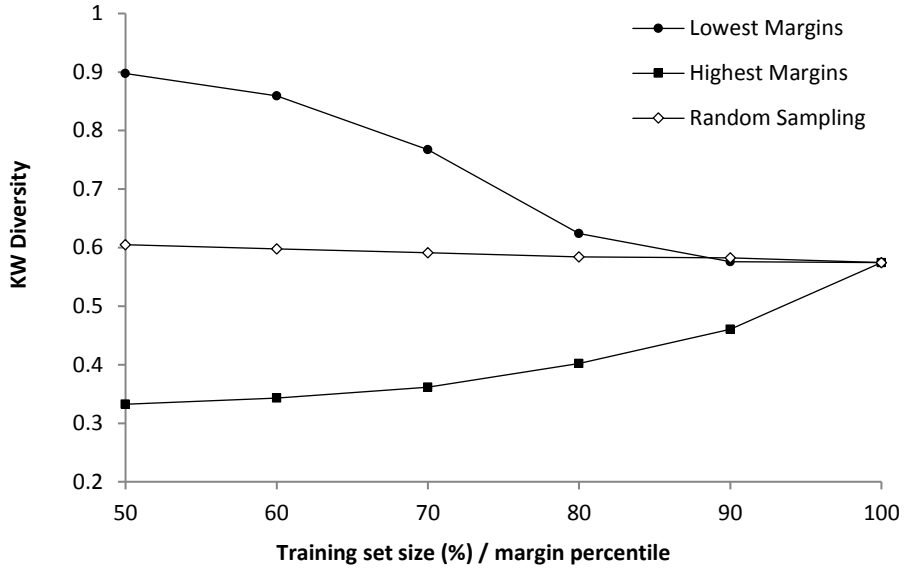


Figure 5-8 Ensemble KW diversity as a function of training set size by lowest and highest unsupervised margins, and random sampling

These results, comparing two opposite margin sampling strategies, show that targeting lower margin training data (which represent samples closer to class boundaries and/or more difficult than higher margin samples) is a means of inducing further diversity among decision trees in an ensemble classifier. The low margin sampling selection strategy (targeting more class decision boundary, difficult and rare class examples) while decreasing mean tree accuracy, demonstrates improved ensemble performance induced by the underlying increase in ensemble diversity.

Table 5-2 Mean tree and ensemble accuracies (%), and Kappa statistic results for the training margin experiments

Margin Percentile Bottom	Mean Tree Accuracy (%)	Mean Tree Kappa	Ensemble Accuracy (%)	Ensemble Kappa
50th	45.71	0.33	66.87	0.60
60th	54.41	0.44	78.96	0.73
70th	60.80	0.52	81.94	0.78

80th	66.64	0.58	82.26	0.78
90th	**68.08**	**0.60**	**82.15**	**0.78**
Top				
50th	68.37	0.60	72.76	0.66
60th	69.98	0.62	75.01	0.68
70th	71.04	0.64	76.78	0.71
80th	**71.41**	**0.64**	78.72	0.73
90th	70.92	0.63	**80.59**	**0.76**
Random Sampling (%)				
50	64.93	0.56	79.36	0.74
60	65.76	0.57	80.24	0.75
70	66.45	0.58	80.85	0.76
80	67.08	0.59	81.32	0.77
90	67.45	0.60	81.54	0.77
100	**68.10**	**0.60**	**82.06**	**0.78**

The effect of low margin sampling is even more pronounced when looking at ensemble accuracy results for only the open canopy class (the most challenging class to classify) (Figure 5-9). Unsurprisingly, this class returns its highest accuracy (74%) in the bottom 50th percentile margins model and its lowest accuracy (53%) in the top 50th percentile margins model. Furthermore, there is a greater than 5% increase in accuracy between lowest margin and random sampling for 50% training set size. Indeed, open canopy has the highest proportion of low margin samples (Figure 5-11). Consequently, as any hard or rare class, it is favoured by an approach which favours the selection of lower margin training data. This strategy reduces data redundancy and increases information significance (e.g. class decision boundary instances are more informative). Therefore, it designs stronger classifiers with an increased capability for handling hard or rare classes.

Classes which are more challenging to predict, such as the open canopy class, may be more commonly misidentified (as woodland or shrub for example) than more easily distinguishable forest canopy classes (e.g. the closed canopy class - which has the lowest proportion of low margin samples among the forest canopy classes - Figure 5-11). Reducing the dominance of highest margin instances in the training dataset may be a strategy to increase ensemble diversity, whereby bagging samples used to construct each decision tree are themselves more diverse, through the inclusion of more instances close to class decision boundaries and more hard class examples.

However, an important reduction in the proportion of higher margin instances in the training set would affect the ensemble classifier performance on easier classes, such as closed canopy, whose loss in accuracy is about 10% in the bottom 50th percentile margin model (Figure 5-10), while this model allows the hardest class (open canopy) to achieve its highest accuracy. This poor ensemble per-class accuracy is associated with relative training data imbalance for the pair closed/open canopies of about 40%-60% (Figure 5-11) - an increase of 10% for the hardest class and a decrease of 10% for the easiest class compared to the balanced case, as well as a reduction in training set size of half of the original set. This result is consistent with the pairwise (open/closed canopies) class imbalance experiment results, involving random sampling, reported in chapter 4. A trade-off in the proportion of low and high margin training samples will benefit harder classes while maintaining, or even improving, the classification performance of easier classes. As Figure 5-10 shows, from the 60th lowest margin percentile, the ensemble accuracy is increased slightly for the closed canopy class, compared to using all of the training data.

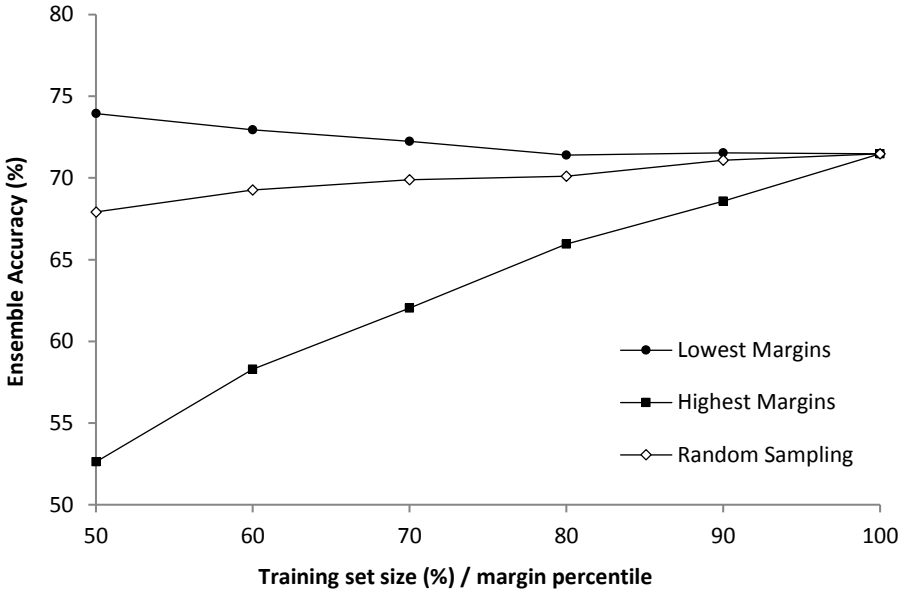


Figure 5-9 Ensemble accuracy for the open canopy class as a function of training set size by lowest and highest unsupervised margins, and random sampling

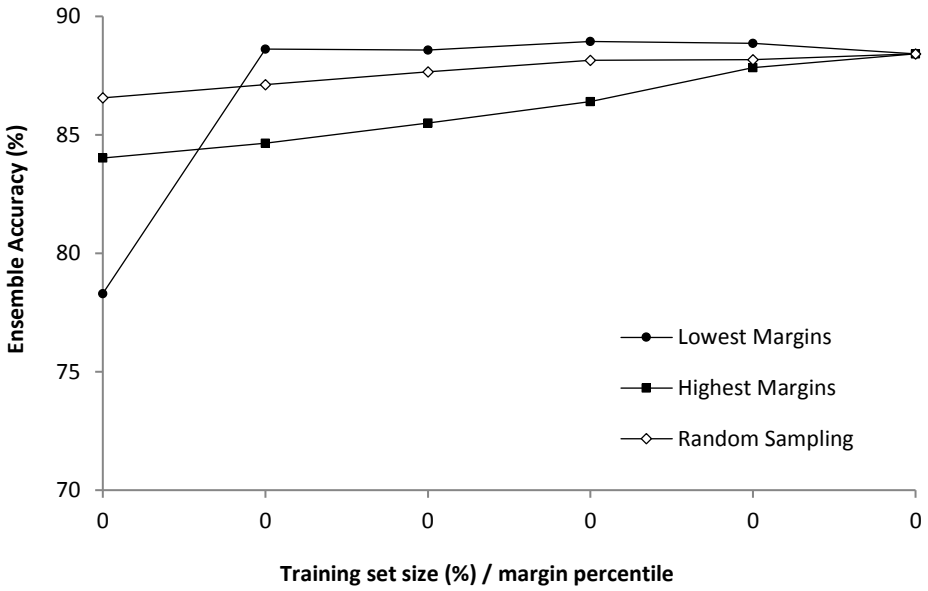


Figure 5-10 Ensemble accuracy for the closed canopy class as a function of training set size by lowest and highest unsupervised margins, and random sampling

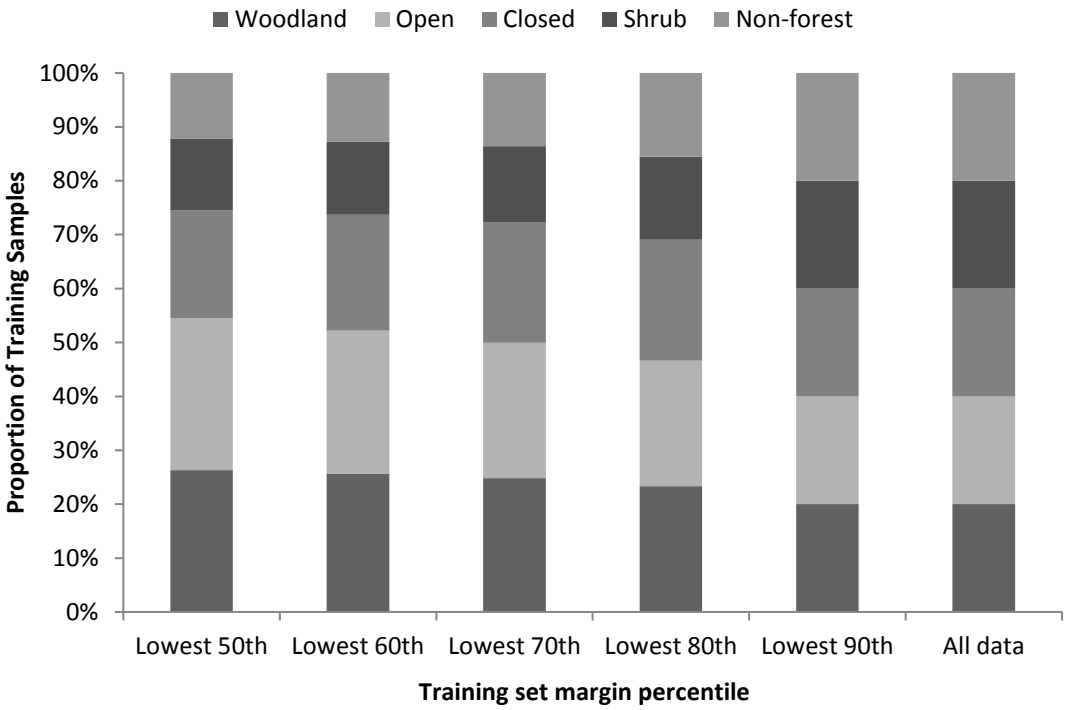


Figure 5-11 Proportion of training samples by class and lowest unsupervised margins by percentile

5.7.3. Influence of the minimum node size on diversity

Results of the minimum node size experiment (Figure 5-12 and Table 5-3) reveal ensemble accuracy to be highest where decision trees are grown to their greatest depth (minimum node size of 1), such as in RF ensembles which use unpruned trees. Decreasing ensemble diversity is associated with lower ensemble accuracy and increasing minimum node size (shallower decision trees). Mean tree accuracy is relatively stable for minimum node size under 50. Hence, the loss in ensemble accuracy in this range is mainly due to the loss in diversity. A minimum node size over 50 also affects mean tree accuracy and therefore induces a steeper drop in ensemble accuracy. Indeed, the generalisation error can be attributed to the combination of the precision of base classifiers and a relative diversity between them (Kapp et al., 2007). While these results demonstrate the relationship between diversity across decision trees and ensemble accuracy, deeper trees mean more complex decision rules which can result in overfitting - particularly if trees are permitted to split down to a single observation.

Table 5-3 Mean tree and ensemble accuracies (%) and Kappa statistic results for the minimum node size experiment

Minimum node size	Mean Tree Accuracy (%)	Mean Tree Kappa	Ensemble Accuracy (%)	Ensemble Kappa
1	68.08	0.62	**82.18**	**0.78**
7	67.83	**0.63**	81.64	0.77
15	68.08	0.62	80.60	0.76
30	**68.50**	0.62	78.90	0.73
50	68.37	0.62	76.99	0.71
100	67.17	0.62	74.06	0.68
250	65.14	0.62	71.06	0.64

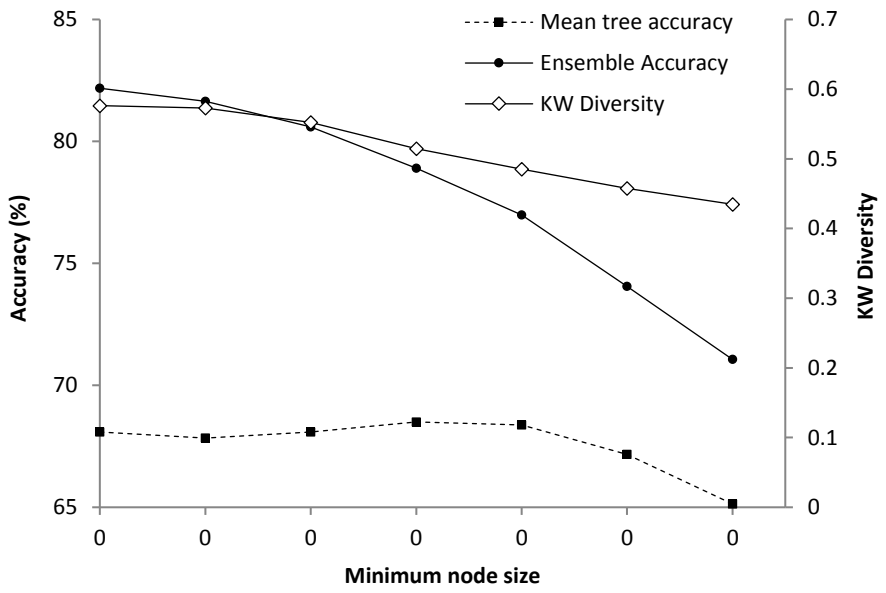


Figure 5-12 Ensemble and mean base classifier accuracies and KW diversity as a function of minimum node size

5.8. Conclusion

The results of these experiments provide insights into the relationship between ensemble diversity and classification performance, in a large area classification problem context using the random forest ensemble classifier. Investigating the effect of the number of decision tree splitting variables on classification and performance showed how lower single tree classification performance (both accuracy and uncertainty) associated with fewer splitting variables is compensated for by higher ensemble diversity, influencing ensemble classification performance. Targeting lower margin training samples (which represent class decision boundaries or more difficult or rarer classes), is a way to increase uncertainty and consequently induce diversity in ensemble learning - a strategy which reduces data redundancy and increases the significance of training information. In the context of large area remote sensing classification, where reference data can be expensive and time-consuming to collect, the margin-based selection of training samples is a way to optimise ensemble classification design, boost efficiency and reduce reference data resource and processing costs. Exploring the influence of tree pruning (through the variation of minimum node size) on classification performance, demonstrated that unpruned decision trees (minimum node size of 1) achieve both the highest single tree classification accuracy and the highest diversity among ensemble members, two ingredients for optimal ensemble classification performance. This result partly explains the superiority of random forests, which use unpruned trees, over other tree-based ensembles such as boosting and bagging, which involve tree pruning.

The findings of this study may inform the design of training data collection strategies and ensemble classification design and parameterisation. Future research will investigate the combined use of ensemble diversity and ensemble margin, two key concepts in ensemble learning, to guide RF training data selection for improved learning and better large area land cover mapping performance.

Chapter 6. Sensitivity of forest Landscape Pattern Indices to training data characteristics in the Random forest classifier

Based on a paper in preparation:

Mellor, A., Haywood, A., Jones, S., Bellman, C. and Boukir, S., (preparation). Sensitivity of forest Landscape Pattern Indices to training data characteristics in the Random forest classifier

6.1. Introduction

Forest fragmentation – the division of forest habitat into smaller and isolated fragments – is considered a significant threat to biodiversity (Haddad et al., 2015) resulting in the geographic and genetic isolation of populations, limiting flora and fauna interactions, interfering with pollination, seed dispersal, wildlife migration and breeding. The effects of forest fragmentation are a function of the number of and distance between forest patches, as well as the amount of edge habitat within each forest patch (Gergel, 2007; Uuemaa et al., 2009). Increased exposure along forest fragment edges as fragmentation increases beyond natural disturbance, leads to long-term changes in structure and function of habitat which remains (Haddad et al., 2015). Fragmentation of forests is an indicator (1.1.c) under Criterion 1 (Conservation of Biological Diversity) of The Montreal Process Criteria and Indicators for the Conservation and Sustainable Management of Temperate and Boreal Forests (Montréal Process Working Group, 2015). The fragmentation of forests indicator describes the loss of forest cover as well as the spatial configuration of that loss. Measures of forest fragmentation and spatial configuration of forest patches, hereafter referred to as Landscape Pattern Indices (LPIs), are used in a number forest monitoring and reporting at national and sub-national scales (e.g. Food and Agriculture Organization of the United Nations, 2015; McRoberts and Liknes, 2002; The State of Victoria Department of Environment and Primary Industry, 2014; Vermont Department of Forests Parks and Recreation, 2015).

LPIs provide quantitative measures for the analysis of landscape structure and composition, including forest fragmentation (Shao et al., 2001). Studies have demonstrated a number of factors which influence the characterisation and computation of landscape indices, including spatial resolution, scale and minimum mappable unit (Wu et al., 1997; Riitters et al., 2000; Shen et al., 2004; Lechner et al., 2012), the distribution, size, and shape of patches in a landscape, and their alignment to remote sensing sensor (Lechner et al., 2009). Methods use to examine and quantify landscape configuration and derive measures of fragmentation, strongly influence the outcome of spatial analysis (Lechner et al., 2013).

Error is always present in the classification of image pixels into land cover classes, and previous research has shown that classification accuracy is not always a good

indicator for the accuracy of landscape pattern characterisation (Langford et al., 2006; Lechner et al., 2013). Moreover, the propagation of error from classification into landscape pattern analysis is of critical importance in landscape ecology (Shao and Wu, 2008). Moreover, classification errors can lead to significant errors and variation in classification map derived landscape pattern indices (Hess, 1994). The sensitivity of landscape pattern indices to classification error - whether resulting from scale-dependent factors such as pixel size or minimum mappable unit (Shen et al., 2004), classification method applied or from mislabelled training data - needs consideration given the potential implications of their application in resource management decision making (Kleindl et al., 2015).

While the sensitivity of landscape metrics to the scale of analysis is reasonably well understood (Shen et al., 2004; Shao and Wu, 2008), their sensitivity to classification error is less known. However, some studies have shown particular metrics (such as mean patch size and patch density) to be more sensitive to classification error than others (e.g. Wickham et al., 1997).

LPIs can be categorised into five groups: area, shape, isolation/proximity, contagion/interspersion and diversity (McGarigal and Marks, 1995). Within these groups, a range of metrics have been used to quantify landscape structure for different land cover classes, including patch area, patch density, patch size, patch variability, amount of edge, shape complexity, core area, nearest neighbour, diversity and contagion and interspersion among patches (Butler et al., 2004). A review by Betts (2000) showed the most commonly applied metrics to be percentage habitat cover, the distribution of patch sizes, edge effects and landscape configuration.

Remote sensing classification is used routinely to generate spatially explicit thematic land cover products, at a range of spatial and temporal scales, from which to measure fragmentation and calculate LPIs. And ensemble machine learning classifiers, like Random Forests (RF) (Breiman, 2001), are now popular techniques for generating land cover maps using remote sensing and ancillary spatial data (Pal, 2005; Mellor et al., 2013; Stefanski et al., 2013; Du et al., 2015b; Belgiu and Drăguţ, 2016). The Random Forest (RF) classifier (Breiman, 2001) uses bootstrap aggregated (bagging) sampling of training data to construct decision trees (base classifiers) which represent a set of diverse base classifiers, which combined into an ensemble are used to assign

a class (prediction) to a response variable through voting (in the case of categorical data) or by averaging (for continuous variables).

Previous studies have demonstrated RF classifier's resistance to mislabelled training data (Rodriguez-Galiano et al., 2012; Rogan et al., 2008), but that associated classification uncertainty is present at even low amounts of mislabelling (Mellor et al., 2015). Given the cited claims about RF robustness to noise, and the inevitability of training class mislabelling in any large area land cover classification setting, it is important to understand the sensitivity of landscape pattern indices to error associated with mislabelled training data. Research has also demonstrated that targeting training data selection on the basis on proximity to class decision boundary is a means to affect per class and overall classification performance.

The specific objectives of this study were to examine, through two experiments, the relationship and sensitivity of LPIs, calculated from RF binary classification forest cover maps, to 1) different rates of mislabelled training data and 2) training data sampling based on the class boundary (i.e. low and high margin training data margin selection strategies). The results from this analysis will provide information to guide the use of LPIs for reporting in forest and monitoring and reporting of forest fragmentation.

6.2. Study Areas

Experiments were applied in two study areas (Figure 6-1) representing contrasting different degrees of forest habitat connectivity and configuration. Study area 1 (Naringal) is located in south west Victoria, Australia. Covering 43,000 hectares (extents -38.29 dd north, 142.85 dd east, -38.49 dd south and 142.63 dd west), the Naringal study area is a highly fragmented agricultural landscape, dominated by grazed pastures, with forest limited to small patches connected by linear forest strips along creeks and road reserves.

Remnant native vegetation in the Naringal study area include heathland, dry forests, herb rich and riverine woodlands, riparian scrub, riverine grass and coastal scrub. The Naringal study area is almost exclusively (98%) private land tenure, it includes about 800 hectares of mostly linear riparian national park and conservation reserve

land. Since European settlement, the Naringal study area and surrounding landscape has experienced significant loss (more than 90%) of its original forest and ongoing fragmentation and isolation of remaining patches (Bennett, 1990)

Study area 2 (Newstead) is located in west central Victoria, about 75 km north west of Melbourne and covers 44,000 hectares (extents -37.06 dd north, 144.24 dd east, -37.26 dd south and 144.02 dd west). Approximately one-third (12,000 ha) of the Newstead study area is public land, comprising mostly large contiguous forested areas of multiple-use commercial State forest tenure (5,000 ha), National park and conservation land (5,000 hectares) and other public land (1,500 ha). The Newstead study area includes the townships of Newstead, Guildford and part of Castlemaine. Native vegetation types in the study area include Box-iron bark forest and herb-rich and riverine woodlands.

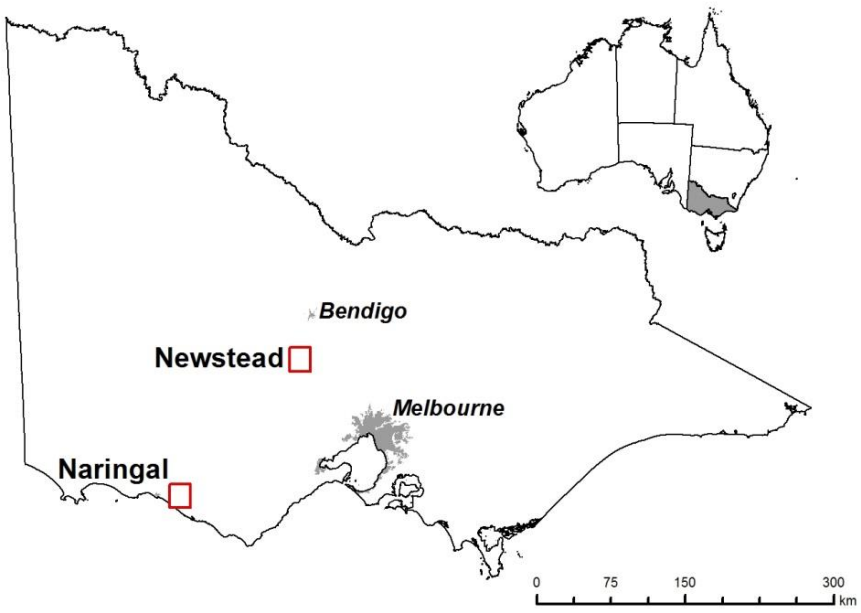


Figure 6-1 Study areas map

6.3. Data

6.3.1. Reference Data

Forest cover reference (training and test) data used to construct and validate RF models, were derived from Aerial Photographic Interpretation of seven hundred and sixty six 2 x 2 km photo-plots, systematically distributed on a random stratified grid

across the State of Victoria. A detailed description of the reference data sources and methodology can be found in (Farmer et al., 2013; Mellor et al., 2013).

For this study, land cover classes were aggregated into a binary forest and non-forest classes. Following Australian National Forest Inventory's definition, forest is defined as having a crown cover greater than 20% and a minimum two metre stand height (Department of Agriculture Fisheries and Forestry, 2012). A half hectare minimum mapping unit was applied to forest cover maps, following the UNFAO forest definition (Food and Agriculture Organization of the United Nations, 2001). API maps from which reference (training and test) data were sampled, were limited to Bioregions containing forest with the most similar structural characteristics to the two study area sites.

6.3.2. Feature variables

Feature variables, comprised remote sensing data - Landsat TM standardised to surface reflectance (Schmidt et al., 2013); Landsat NDVI derived texture indices and Tasselled cap features (Crist and Cicone, 1984); elevation, slope, aspect (Farr et al., 2007) and biophysical climate data (Houlder, 2001).

6.3.3. Geospatial Database

Google Earth Engine - a cloud-based online platform which combines public remote sensing and geospatial data with large computational facility designed for parallel processing geospatial data (Hansen et al., 2013; Google Earth Engine Team, 2015) was used to source and pre-process feature variable input data.

Reference data were imported into Google Earth Engine (Google Earth Engine Team, 2015) as vector polygons, converted to raster and resampled to align with feature resolution. Reference and feature variable data were extracted into two forest and non-forest datasets, comprising about 1 million randomly sampled candidate forest and non forest pixels. Feature variables were extracted over the extents of the two study area sites (Figure 1). All data were imported into R (R Development Core Team, 2011) for RF model construction and evaluation and LPI generation.

6.4. Random forest

The randomForest package (Liaw and Wiener, 2002) in R (R Development Core Team, 2011) was used to build RF models in each experiment. The R package SDM Tools (VanDerWal and Falconi, 2014) was used to calculate landscape shape indices (class statistics) from forest cover class in each forest/non-forest map. The default number of randomly sampled predictor variables (parameter *mtry*) was used as candidates for each decision tree splitting node (equal to the square root of the total number of predictor variables). RF ensembles were constructed with 100 trees (base classifiers). Assignment of class was determined by the majority of votes from all decision trees in the ensemble, a standard approach for combining the decisions of multiple component learners.

6.5. Ensemble margin

Margin theory is a machine learning concept which explores data proximity to decision boundaries. It is a means of understanding ensemble classification (such as RF) and of estimating confidence in classification outcomes (Schapire et al., 1998; Mellor et al., 2015). The ensemble margin of a training data instance is the difference between the number of class votes to which it is assigned by decision trees in an ensemble classifier. For example, in a binary classification problem, with an ensemble containing 100 trees, a training instance (of Class A) assigned 60 decision tree votes to Class A and the remaining 40 votes to class B, would have a margin score of 20 (or 0.2, normalised by the total number of decision trees). Correctly classified training instances with high margin values (i.e. close to or equal to 1), where there is strong consensus among all decision trees, typically represent training instances located away from class decision boundaries. Training instances with low margin values, which are located closer to class decision boundaries may offer more information to a classification problem, unlike high margin values which may contain a high degree of redundant information.

The unsupervised Schapire's margin, a variant show by equation 1 (Guo and Boukir, 2013) for a sample x, is the difference between the number of votes (respectively V_{c1} and V_{c2}) assigned to the first and second most voted for classes (respectively c_1

and c_2), normalised by the number of decision trees (or base classifiers) in the ensemble (T), regardless of the true class label.

$$margin(x) = \frac{V_{c1} - V_{c2}}{\sum_{c=1}^{L}(V_c)} \quad (1)$$

True class labels are not considered in the unsupervised margin and so this measure may be more robust to noise (Guo, 2011).

Previous chapters and Mellor et al., (2015) and Mellor and Boukir (2017) have shown that using the ensemble margin is an effective training data sampling technique that can be used to increase the significance of particular training samples in a classification problem, such as deliberately targeting samples close to classification decision boundaries and reducing the proportion of redundant high margin training samples - both means of increasing the significance of training information, boosting classification efficiency and improving global classification model performance (as well as the performance of more difficult classes).

6.6. Landscape Pattern Indices

Six commonly applied LPIs in the categories of area, shape and aggregation were evaluated in this study (Table 6-1).

Table 6-1 Description of Landscape Pattern Indices (LPIs)

Landscape Pattern Index	Description	Category
Number of (forest) patches	Total number of patches in the forest class category in the landscape	Area
Class (forest) area (ha)	Total area (in hectares) of the forest class category in the landscape	Area
Area weighted mean patch size (ha)	The sum of all forest patches in the landscape multiplied by the proportional abundance of the of the patch (i.e. patch area divided by the sum of all patch areas).	Area
Edge Density	Ratio of total edges (number of cells at patch boundary) to total area (all cells) (m/ha)	Shape
Area weighted mean fractal dimension	Measure of patch shape complexity (mean fractal dimension of patches weighted by patch area)	Shape
Percentage of like adjacencies	The frequency with which different pairs of patch types (including like adjacencies between the same patch type) appear side-by-side on the map (measures the degree of aggregation of patch types)	Aggregation

6.7. Experiment 1: Margin-based training data sampling

Using all available training data (50,000 forest and 50,000 non-forest samples), an initial RF model was constructed from which the unsupervised margin values for all training data samples was calculated (equation 1). Percentiles were then calculated from the unsupervised margin values of the total training set, and RF models constructed using 20,000 training instances (10,000 per class) randomly selected from margins in the bottom 40th, 50th, 60th, 70th, 80th and 90th percentiles, and all available training data regardless of margin score. For each model, a random sub-set of 10,000 samples (5,000 per class) was drawn from the master training set as validation (test) data with which to calculate the overall and per-class accuracy of each RF model.

Each RF model was applied to create forest/non-forest land cover maps in each study area. A minimum mapping unit of 0.5 hectares was applied to the forest cover maps to meet the forest cover definition (FAO, 2000; The State of Victoria Department of Environment and Primary Industry, 2014) and remove classification noise, by first resampling cells to 28m, grouping together horizontally, vertically and diagonally contiguous forest and non-forest classified cells, and reclassifying cells in groups of less than six (from either forest to non-forest, or non-forest to forest).

Landscape Pattern Indices (Table 6-1) were calculated from the post-processed forest/non-forest classified maps. The process of randomly selecting training samples by margin percentile, constructing RF models, creating forest/non-forest prediction maps and calculating LPIs, was repeated to generate 30 sets of LPI results using training data drawn (in 10 percentile ranges) from the bottom and top 40th to 90th percentiles, and randomly sampled.

6.8. Experiment 2: Training data mislabeling

A second experiment examined the relationship between the proportion of mislabeled training instances used to construct RF ensemble models, and derived Landscape shape indices. Training data sub-sets comprising 20,000 samples (10,000

per class) were randomly drawn from the master training set and for each sub-set, training data instances were randomly re-assigned their class label at proportions of 5 per cent of the training sample (i.e. 1,000 out of 20,000 samples), increasing the proportion of mislabeling at 5 per cent increments up to 30 per cent mislabeling. Following steps outlined in Experiment 1 (above), RF models were constructed, forest classification maps generated (and minimum mapping unit applied), from which LPIs were calculated for each study site.

6.9. Analysis of sensitivity of experiments

A simple linear model (equation 2) and linear model with quadratic function (equation 3) were fitted to the results to test whether there was an overall trend (sensitivity) between the margin-based training data selection (experiment 1) or mislabeling (experiment 2) and the derived accuracy and landscape shape indices, and if so (and significant), whether this trend was linear or curve linear.

$$Y = a + bX \tag{2}$$

$$Y = a + bX + cX^2 \tag{3}$$

where Y represents the Landscape Pattern Index, a the intercept, b the slope and X is the training margin percentile selection.

6.10. Results and discussion

6.10.1. Experiment 1 results

Table 6-2 and Table 6-3 show results of the first experiment (margin-based training data selection) for the two study areas. For each LPI and study area, results show the nature of the overall trend (relationship) – linear (L), curve linear (CL) or Not Significant (NS) as well the associated p-value, between LPI and training margins percentile.

Table 6-2 Nature of the trend between margin-based training selection (40th to 90th percentile and random sampling) and LPIs for the Naringal Study area. Curve Linear (CL), Linear (L) or Not-significant (NS).

Landscape Pattern Index	**Trend**	**p-value**
Number of forest patches	CL	0.005
Total Area of Forest	NS	0.167
Area weighted mean patch size	NS	0.212
Edge density	CL	0.001
Area weighted mean fractal dimension	NS	0.113
Percentage of like adjecencies	L	0.028
Overall accuracy	CL	2.85E-105
User accuracy forest	CL	2.24E-83
User accuracy non-forest	CL	3.70E-79

Table 6-3 Nature of the trend between margin-based training selection (40th to 90th percentiles and random sampling) and LPIs for the Newstead Study area. Curve Linear (CL), Linear (L) or Not-significant (NS).

Landscape Pattern Index	**Trend**	**p-value**
Number of forest patches	CL	0.010
Total Area of Forest	NS	0.881
Area weighted mean patch size	CL	0.004
Edge density	NS	0.931
Area weighted mean fractal dimension	NS	0.170
Percentage of like adjecencies	NS	0.757
Overall accuracy	CL	2.85E-105
User accuracy forest	CL	2.24E-83
User accuracy non-forest	CL	3.70E-79

Results for the Naringal study site, show no significant linear or curve linear relationship between the RF classifiers trained with samples selected based on the

ensemble margin (bottom 40th to 90th percentile and random sampling) and, the total area of forest, the area weighted mean patch size or the area weighted mean fractal dimension. The number of forest patches (Figure 6-2) and edge density (Figure 6-3) LPIs exhibit a curve linear relationship with margin percentile. Edge density is lowest for maps classified with RF models constructed with training data randomly sampled from the bottom 90th percentile by ensemble margin value. Bottom 90th percentile models include a greater representation of training instances close to decision boundaries compared to models built using training data selected in the lower percentile range (e.g. 40th percentile), in which there are a greater proportion of training samples further from decision boundaries.

Figure 6-2 illustrates a trend at the Naringal Study Site in which the lowest margin sampling strategy (the bottom 40th percentile) produces classified forest maps with the fewest forest patches. This LPI increases (curve linearly) as training data are sampled from a greater range of margin values (i.e. up to the 90th percentile, and random sampling).

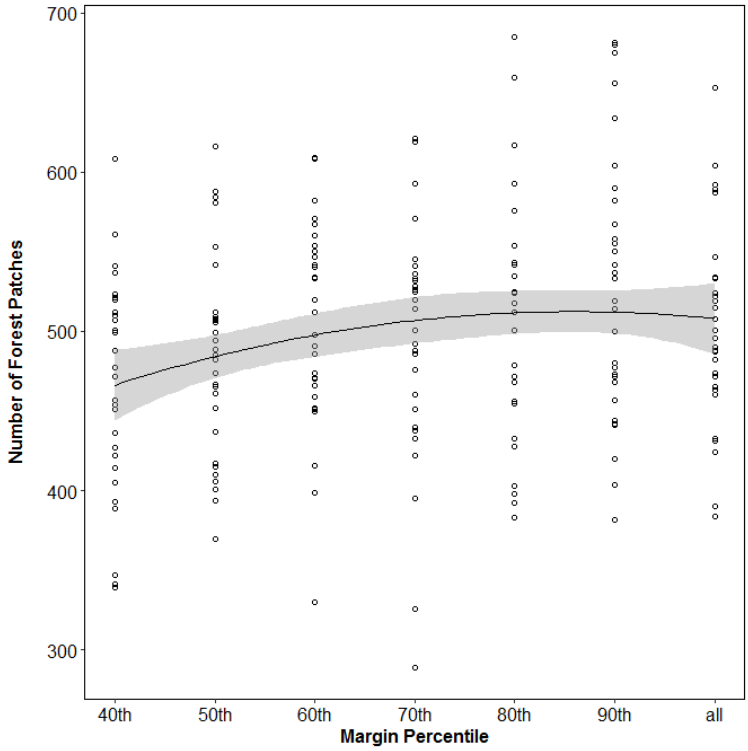


Figure 6-2 Scatter plot showing curve linear trend between Number of Forest Patches and training data sampling margin percentile (Naringal)

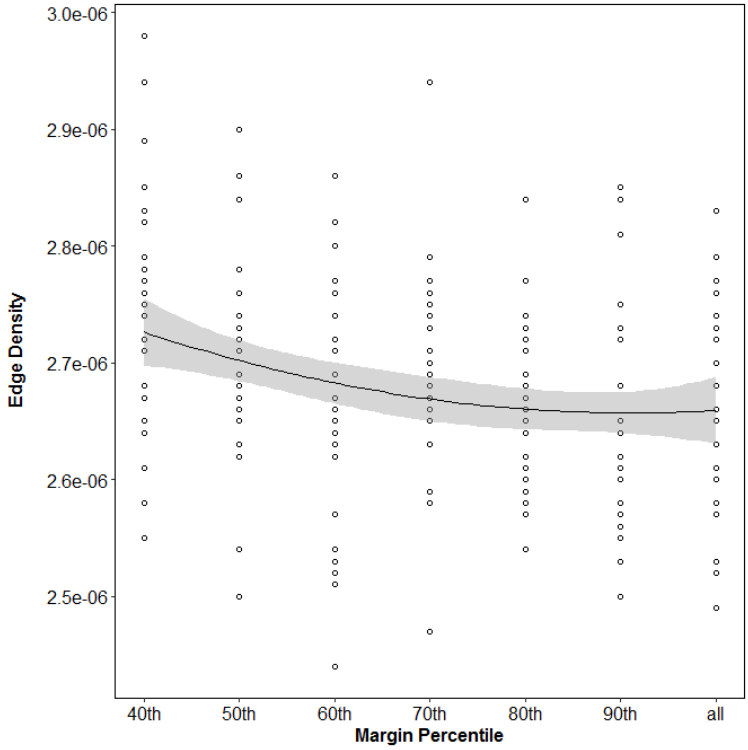


Figure 6-3 Scatter plot showing curve linear trend between Edge Density and training data sampling margin percentile (Naringal)

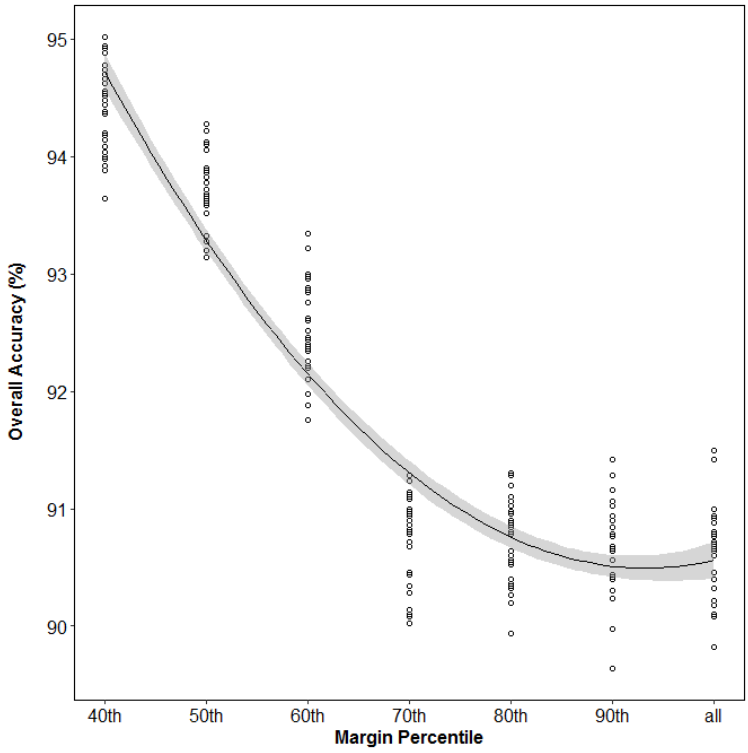


Figure 6-4 Scatter plot showing curve linear trend between overall model accuracy and training data sampling margin percentile

The percentage of like agencies LPI is also sensitive to margin-based training sample selection (with a linear trend). In the Narginal landscape - like the edge density metric - the percentage of like agencies (a measure of landscape heterogeneity) decreases with the proportion of higher margin training samples.

The total area of forest metric shows no significant linear or curve linear trend (Figure 6-5) with the margin percentile sampling. Figure 6-5 also highlights the relatively high variance in calculated forest area for the different iterations of the model run for each margin percentile. There was an average difference of 676 ha between the minimum and maximum total forest area calculated among the margin percentile sampling levels, with the lowest and lowest variance in the 40th and 80th margin percentile sampling respectively.

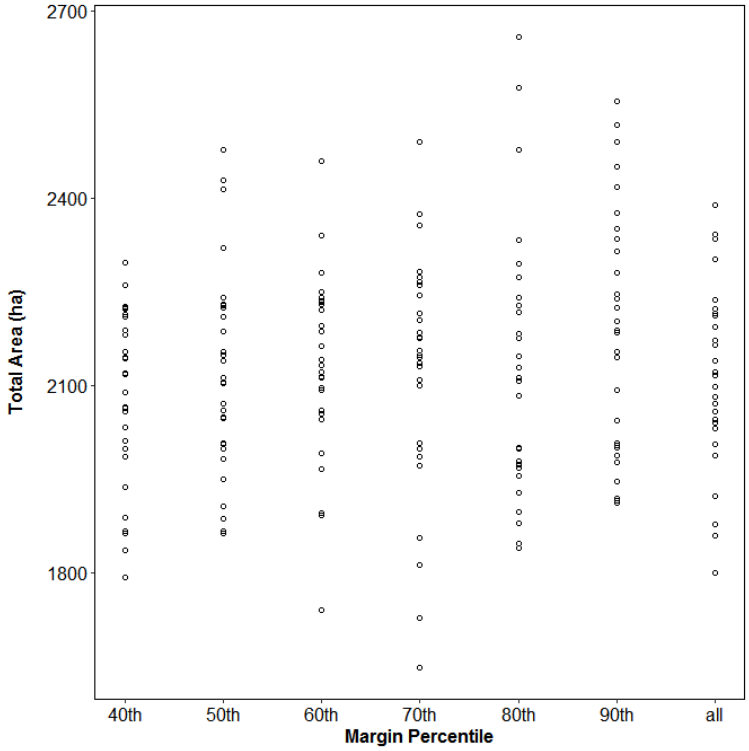


Figure 6-5 Scatter plot showing no significant relationship between Total area of forest and training data sampling margin percentile (Naringal)

Figure 6-6 and Figure 6-7 shows example forest extent maps generated from training data sampled in the bottom 40th percentile margin, and training data sampled in the bottom 90th percentile margin.

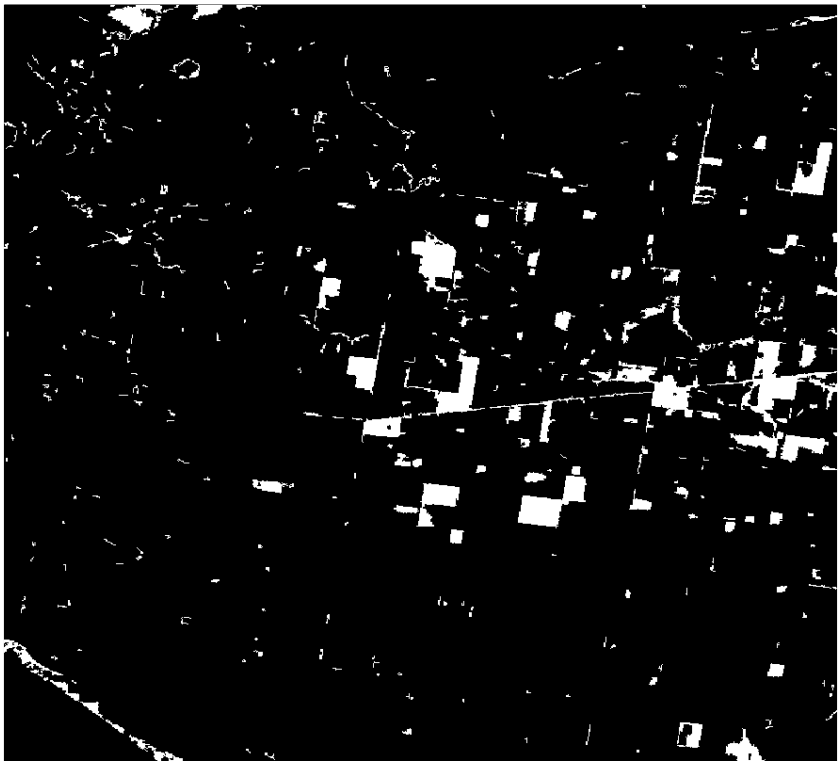


Figure 6-6 Naringal forest extent map from classification training data sampled from the 40th percentile margin values

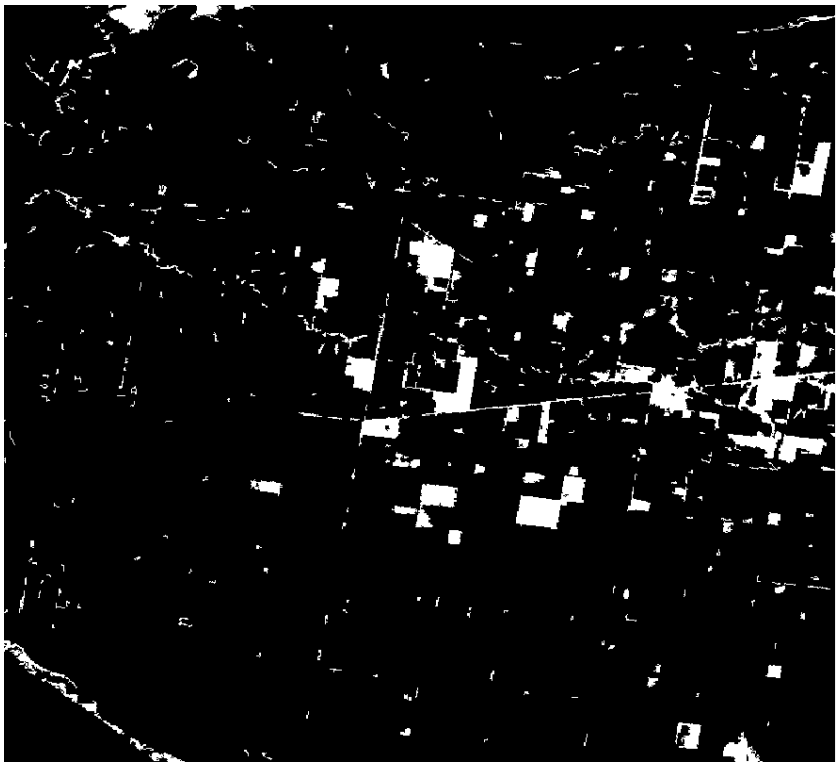


Figure 6-7 Naringal forest extent map from classification training data sampled from the 90th percentile margin values

Previous research has demonstrated that targeting lower margin training samples (which are closer to class decision boundaries) is an effective technique to increase the diversity and performance of an ensemble classifier like RF (Mellor and Boukir, 2017). A training set sampled from the bottom 90th margin percentile contains a higher greater proportion of training instances located away from decision boundaries, and therefore more redundant and less significant information in the classifier.

Overall classification accuracy (calculated from set-aside test data), exhibits a strong curve linear relationship with training margin. Highest overall accuracy is achieved with RF models generated using lowest margin (bottom 40th percentile) training data (an average overall accuracy of 94.4% over 30 iterations), which drops to an average 90.6% for the bottom 80th and 90th percentiles, and random sampling models.

The classification of smaller and fragmented patches of forest and as well as linear strips along roadside and riperian vegetation, common across the Naringal site, improves with a higher proportion of low margin training instances in the RF model (i.e. training data close to forest-non-forest decision boundaries). As the proportion of higher margin training instances increases in the RF models, the edge density LPI, representing the ratio of total (forest patch) edges and total area, falls (curve linearly). These results show that for the Naringal site, low margin model (bottom 40th percentile) classified forest extent is slightly less fragmented, compared to higher margin models (bottom 90th percentile or random sampling). The curve linear increasing trend in the total number of forest patches, also shows fragmentation of forest cover increases with the training margin percentile.

In contrast, in Newstead study site (which has a greater proportion of contiguous forest and larger forest patches compared to the Naringal site), the area weighted mean patch size exhibits a curve linear relationship with margin percentile and the number of forest patches.

In the Newstead study area, results show only area weighted mean patch size, number of forest patches LPIs and the accuracy metrics, have a significant curve linear relationship with margin-based training data selection. While the number of forest patches had a positive relationship with margin percentile (Figure 6-8), there was no associated increase in the total area of forest for the Newstead study area.

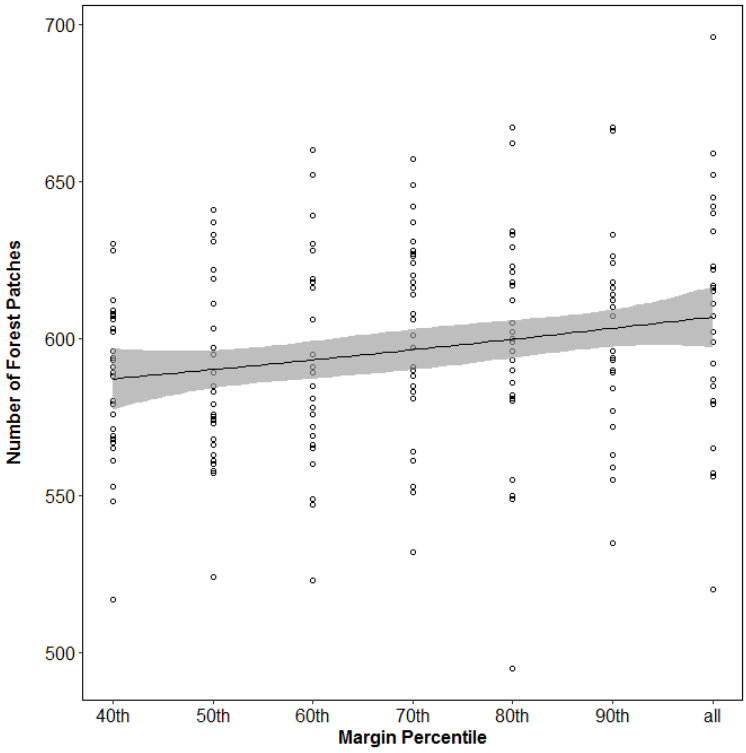


Figure 6-8 Scatter plot showing curve linear relationship between the number the forest patches and training data sampling margin percentile (Newstead).

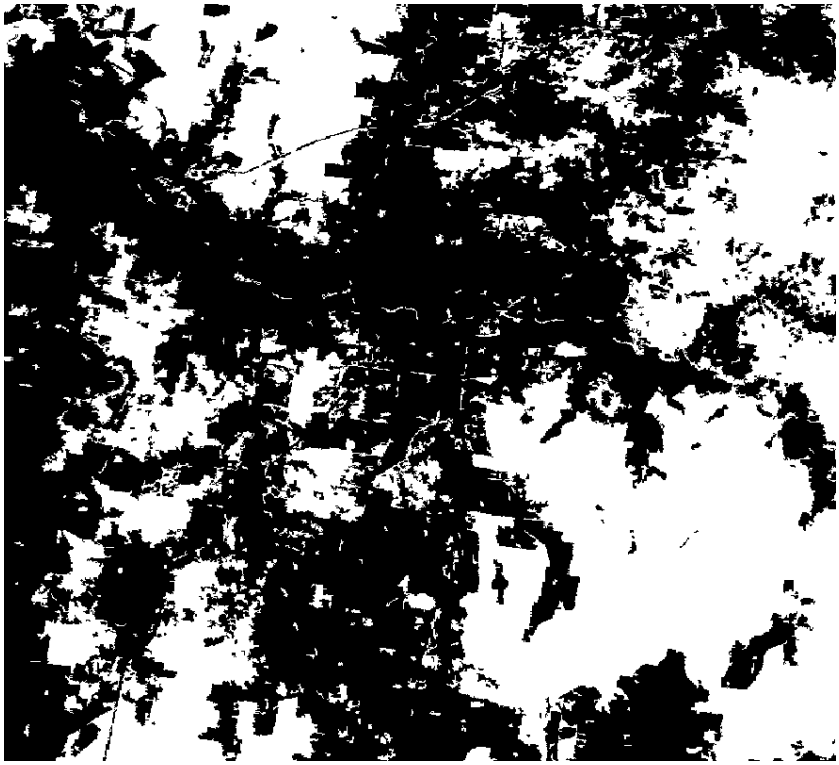


Figure 6-9 Newstead forest extent map from classification training data sampled from the 90th percentile margin values

Results of experiment 2 (training data mislabeling) at the Naringal site, show all LPIs to have a significant curve linear relationship with increasing rate of training data mislabeling (Table 6-4) and for all but Total area of forest and edge density, at the Newstead site (Table 6-5).

Table 6-4 Nature of the trend between mislabeled training data (from 0% up to 30%) and LPIs for the Naringal Study area. Curve Linear (CL), Linear (L) or Not-significant (NS).

Landscape Pattern Index	**Trend**	**p-value**
Number of forest patches	CL	5.14E-20
Total Area of Forest	CL	2.80E-08
Area weighted mean patch size	CL	1.17E-31
Edge density	CL	0.002
Area weighted mean fractal dimension	CL	1.17E-31
Percentage of like adjecencies	CL	1.23E-55
Overall accuracy	CL	1.06E-135

User accuracy forest	CL	7.80E-112
User accuracy non-forest	CL	1.42E-106

Table 6-5 Nature of the trend between mislabeled training data (from 0% up to 30%) and LPIs for the Newstead Study area. Curve Linear (CL), Linear (L) or Not-significant (NS).

Landscape Pattern Index	**Trend**	**p-value**
Number of forest patches	CL	1.18E-36
Total Area of Forest	NS	0.109
Area weighted mean patch size	CL	0.001
Edge density	NS	0.978
Area weighted mean fractal dimension	CL	2.00E-21
Percentage of like adjecencies	CL	1.00E-69
Overall accuracy	CL	4.96E-106
User accuracy forest	CL	4.47E-100
User accuracy non-forest	CL	4.60E-95

6.11. Conclusion

This investigation contributes to the understanding of the sensitivity of LPIs to training data characteristics used in machine learning classification. The study's findings demonstrate that LPIs can have strong sensitivity to training data selected on the basis of the ensemble margin and proximity to class decision boundaries. Although the accuracy of forest extent maps produced using the RF algorithm are generally insensitive to low to moderate levels of training data mislabelling (Rodriguez-Galiano et al., 2012; Mellor et al., 2015), the study's findings indicate LPIs have a high degree of sensitivity to training data quality (even at low rates of mislabelling). As such, forest and other land cover mapping applications need to consider the implications of training data quality used in the generation of LPIs through RF classification.

Chapter 7. Thesis Synthesis and Conclusions

Satellite-based (remote sensing) earth observing technology has long been recognised as a critical source of large area land cover maps and information, used for a variety of natural resource management applications. Over the past decade, machine learning algorithms have become increasingly popular techniques for classifying remote sensing data, superseding traditional parametric classification algorithms, due to improved performance and their ability to address complexity inherent in many large and heterogeneous landscapes. In the remote sensing literature, the Random Forests (RF) ensemble classifier stands out as an increasingly popular ML technique - with an average 30% increase in published articles citing Random forests in remote sensing since 2010.

The overarching objective of this thesis was to examine cited advantages of the RF classifier in the context of large area land cover classification problems. The research also explored the utility of ensemble learning as a means to design more efficient classification systems which use reference data more efficiently and effectively.

7.1. Research Questions

Research Question 1: How do training data characteristics of class imbalance and class mislabelling affect RF performance?

A series of binary and multiclass forest cover classification experiments presented in Chapter 4, provide insight into the behaviour of the RF ensemble classifier and the degree of influence training data imbalance and mislabelling can have on classification performance.

Class-balanced classification models for binary and multi-class experiments - for both optimal (large) and critical (small) training dataset sizes - provided highest overall classification accuracies and associated measures of confidence. However, results of multiclass classification imbalance experiments showed that careful and deliberate imbalancing of training data is an effective means to improve the performance of challenging (or difficult classes), that does not appreciably compromise overall or other per-class classification results. A 10% decrease in the number of training samples in the easiest class (closed forest canopy cover) coupled

with a 10% increase in the hardest class (open canopy) achieved the best classification performance result.

Previous studies have shown key features of the RF classifier make the performance of this algorithm relatively robust to training data mislabelling (or noise) - including bagging used to select random sub-sets of training data to construct decision trees, as well as the random selection of features used to partition training data at decision tree nodes. Indeed, results of training data multiclass mislabelling experiments in chapter 4 showed that compared to clean (i.e. not deliberately mislabelled) training data, applying a 25% mislabelling rate to training samples only resulted in reductions of 6.6% and 7.2% in overall classification accuracies for optimal and critical training set sizes respectively. However, an associated 55% decrease in the mean margin for these mislabelling experiments showed that while class mislabelling effects on reduced classification accuracy is not considerable, even low mislabelling rates can strongly influence rates of classification confidence (uncertainty).

Research Question 2: What is the relationship between ensemble diversity and classification performance?

This research question sought to examine the degree of influence that RF ensemble diversity has on classification performance, and to understand how diversity can be controlled or induced to improve RF classification effectiveness and efficiency. While the complex relationship between diversity and ensemble classification performance is not yet fully explored nor understood, diversity is recognised as an essential condition for designing high performing ensemble classifiers, such as RF.

Building on ensemble margin theory introduced in chapter 4, research chapter 5 examined the theme of ensemble diversity and its association with RF performance in a large area land cover classification setting. Results provide insights into the trade-off between ensemble classification accuracy and diversity, and through the ensemble margin, demonstrate how inducing diversity by targeting lower margin training samples is an effective means of achieving better classifier performance for more difficult or rarer classes and reducing information redundancy in classification problems. For the most difficult canopy cover class (open forest), targeting low margin training samples in the bottom 50th percentile returned its highest accuracy

(74%), compared to training data selection in the top 50th percentile (only 53% accuracy). Moreover, results showed a more than 5% improvement in accuracy using lowest margin training data sampling, compared to random sampling selection of training data. This chapter emphasised the importance of the trade-off between accuracy and RF ensemble member diversity, whereby over-fitting can be the result of too little diversity, and poor accuracy, the result of too much diversity.

Research question 3: What is the relationship between training data characteristics (used to construct RF ensemble classification models) and Landscape Pattern Indices calculated from the derived RF forest prediction maps?

The final research chapter (6), builds on the concepts introduced in chapters 4 and 5, by examining the relationship between training data characteristics used to construct RF ensemble models and landscape pattern indices derived from classified forest extent maps. LPIs provide quantitative measures for analysing land landscape structure and composition, including forest fragmentation. Through two experiments, this study examined the sensitivity of LPIs to increasing rates of training mislabelling, and also how training data sampling strategies (using the margin to sample training data on the basis of proximity to class decision boundary), affect LPI measures. The study revealed a high degree of sensitivity to training data mislabelling, even at low mislabelling rates, regardless of whether the landscape comprises highly fragmented forest cover or is characterised as more contiguous forest configuration.

7.2. Summary

Previous studies have cited the importance of various reference data attributes used in supervised land cover classification problems. Land cover map accuracy is sensitive to the quality of reference data (Foody et al., 2016). A number of key training data themes relating to the sensitivity of RF classifier in remote sensing have, and continue to be explored, ranging from sampling design (Colditz, 2015) and sample size (Stumpf and Kerle, 2011; Deng and Wu, 2013; Du et al., 2015a), to training data class imbalance (as explored in this research and Jin et al. (2014), and training data noise (mislabelling) e.g. this thesis, and Rodriguez-Galiano et al. (2012).

Generating sufficient supervised classification training data is a time consuming, expensive and subjective task (Lippitt et al., 2008; Ghimire et al., 2012). Moreover, training data-hungry machine learning techniques for large area land cover classification problems, requires large samples of unbiased representative reference data (Egorov et al., 2015) which account for within and between class heterogeneity, are of suitable accuracy, geographic coverage and align with remote sensing imagery acquisition/capture time or seasonal conditions. These training data challenges may be exacerbated in jurisdictions containing remote and inaccessible regions, or in resource poor environments.

The novel exploration and application of the unsupervised ensemble margin in large area remote sensing classification in this thesis, provides some insight into the behaviour of RF with respect to class imbalance and mislabelling. Moreover, through the unsupervised margin, the work presents a useful means to evaluate the relative contribution of individual training samples to the learning process and boost classification performance. This technique could be applied to design more efficient RF classifiers and reduce the generation and use of wasteful "information redundant" training data and focus sampling on areas and classes which have a greater influence on the outcome of an ensemble RF classifier.

Three key developments have facilitated the uptake of machine learning algorithms like RF, in the field remote sensing. These include the increasing availability of remote sensing data - associated with both an increasing number of satellite sensors, and open data policies (Wulder et al., 2012). Open-source implementations of machine learning algorithms which allow classifiers like RF to be readily automated with a set of user defined adjustable parameters (Rodriguez-Galiano et al., 2012) which make algorithms relatively straightforward to apply by relatively inexperienced users (Qi et al., 2006). Several implementations of the RF classifier are now available, including the most popular randomForest (Liaw and Wiener, 2002) in the statistics package R (R Core Team, 2013), as well as implementations in Python, such as *scikit learn Ensemble forest* (scikit-learn developers, 2016) and through the Machine Learning Tool Kit (MILK) (Coelho, 2017) and *Fast random forest* in the WEKA Environment. The increased performance and availability of low-cost computing is also facilitating the uptake of machine learning in remote sensing. For example, cloud computing – the practice of using a network of internet

hosted, remotely accessed servers to store, manage and process data – provides significant opportunities to address the challenge of large scale and data and processing-hungry remote sensing applications (Sugumaran et al., 2015). Cloud-computing offers relatively low-cost and scalability advantages in data storage and processing compared to traditional high powered, user owned computing clusters.

7.3. Future research

An extension of the research presented in this thesis, will be to extend the application of the ensemble margin as a means to improve RF stability and performance. This would include evaluating the application of the ensemble margin to inform up and down-sampling of class imbalanced training datasets. In addition, the link between ensemble diversity and machine learning performance in remote sensing classification has not been fully explored. Further research in this area could look at ways to induce diversity in ensemble algorithms such as RF to build more robust classifiers, that are, for example, more robust to noise in training data and predictor variables. Techniques to promote and artificially induce ensemble diversity could include through training data sampling strategies, decision tree construction techniques, and class switching (the deliberate introduction of class mislabelling).

References

Amit, Y. and Geman, D. (1997) 'Shape Quantization and Recognition with Randomized Trees', *Neural Computation*, 9(7), pp. 1545–1588. doi: 10.1162/neco.1997.9.7.1545.

Armston, J. D. (2009) 'Prediction and validation of foliage projective cover from Landsat-5 TM and Landsat-7 ETM+ imagery', *Journal of Applied Remote Sensing*, 3(1), p. 33540. doi: 10.1117/1.3216031.

Austin, M. P. and Meyers, J. a. (1996) 'Current approaches to modelling the environmental niche of eucalypts: implication for management of forest biodiversity', *Forest Ecology and Management*, 85(1–3), pp. 95–106. doi: 10.1016/S0378-1127(96)03753-X.

Australian Surveying and Land Information Group (1990) *Atlas of Australian Resources (Volume 6 Vegetation)*. Canberra.

Baccini, A., Laporte, N. T. T., Goetz, S. J., Sun, M. and Dong, H. A. (2008) 'A first map of tropical Africa's above-ground biomass derived from satellite imagery', *Environ. Res. Lett.*, 3, pp. 1–9.

Banfield, R. E., Hall, L. O., Bowyer, K. W. and Kegelmeyer, W. P. (2005) 'Ensemble diversity measures and their application to thinning', *Information Fusion*, 6(1), pp. 49–62. doi: 10.1016/j.inffus.2004.04.005.

Beaumont, L., Hughes, L. and Poulsen, M. (2005) 'Predicting species distributions: use of climatic parameters in BIOCLIM and its impact on predictions of species' current and future distributions', *Ecological Modelling*, 186, pp. 250–269.

Behn, G., McKinnell, F. and Caccetta, P. (2001) 'Mapping forest cover, Kimberley Region of Western Australia', *Australian Forestry*, 64(2), pp. 80–87. Available at: http://www.freepatentsonline.com/article/Australian-Forestry/197666175.html (Accessed: 18 May 2012).

Belgiu, M. and Drăguţ, L. (2016) 'Random forest in remote sensing: A review of applications and future directions', *ISPRS Journal of Photogrammetry and Remote Sensing*, 114, pp. 24–31. doi: 10.1016/j.isprsjprs.2016.01.011.

Belward, A. S. and Skøien, J. O. (2015) 'Who launched what, when and why; trends in global land-cover observation capacity from civilian earth observation satellites', *ISPRS Journal of Photogrammetry and Remote Sensing*. International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS), 103, pp. 115–128. doi: 10.1016/j.isprsjprs.2014.03.009.

Bennett, A. F. (1990) 'Habitat corridors and the conservation of small mammals in a fragmented forest environment', *Landscape Ecology*, 4(2), pp. 109–122. doi: 10.1007/BF00132855.

Betts, M. (2000) *In Search of Ecological Relevancy: A Review of Landscape Fragmentation Metrics and Their Application for the Fundy Model Forest*. Natural Resources Canada.

Bhandari, S. (2011) *Monitoring Forest Dynamics using Time Series of Satellite Image Data in Queensland, Australia*. The University of Queensland.

Bi, Y. (2012) 'The impact of diversity on the accuracy of evidential classifier ensembles', *International Journal of Approximate Reasoning*, 53, pp. 584–607. Available at: http://www.sciencedirect.com/science/article/pii/S0888613X11001885 (Accessed: 6 October 2015).

Bivand, R. (2007) 'Using the R-GRASS Interface: Current Status', *OSGeo Journal*, 1, pp. 36–38. Available at: http://www.osgeo.org/files/journal/final_pdfs/OSGeo_vol1_GRASS-R.pdf.

Boyd, D. S. and Danson, F. M. (2005) 'Satellite remote sensing of forest resources: three decades of research development', *Progress in Physical Geography*, 29(1), pp. 1–26. doi: 10.1191/0309133305pp432ra.

Bradley, C. and Friedl, M. (1996) 'Identifying and Eliminating Mislabeled Training Instances', in *Proceedings of the thirteenth national conference on Artificial intelligence*. Amherst, MA, pp. 799–805. Available at: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Identifying+and+eliminating+mislabeled+training+instances#1 (Accessed: 23 March 2014).

Breiman, L. (1996) 'Bagging predictors', *Machine learning*, 24(2), pp. 123–140. Available at: http://www.springerlink.com/index/L4780124W2874025.pdf (Accessed: 22 February 2013).

Breiman, L. (2001) 'Random Forests', *Machine Learning*. Springer Netherlands, 45(1), pp. 5–32. doi: 10.1023/A:1010933404324.

Breiman, L. and Cutler, A. (2001) *Random Forests*. Available at: http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm.

Brus, D. (2010) 'Design-based and model-based sampling strategies for soil monitoring', in *19th World Congress of Soil Science*, pp. 32–34.

Butler, B. J., Swenson, J. J. and Alig, R. J. (2004) 'Forest fragmentation in the Pacific Northwest: quantification and correlations', *Forest Ecology and Management*, 189(1–3), pp. 363–373. doi: 10.1016/j.foreco.2003.09.013.

Calle, M. L. and Urrea, V. (2011) 'Letter to the editor: Stability of Random Forest importance measures.', *Briefings in bioinformatics*, 12(1), pp. 86–9. doi: 10.1093/bib/bbq011.

Carletta, J. (1996) 'Assessing agreement on classification tasks: The kappa statistic', *Computational Linguistics*, 22(2), pp. 249–254.

Carlotto, M. J. (2009) 'Effect of errors in ground truth on classification accuracy', *International Journal of Remote Sensing*, 30(18), pp. 4831–4849. doi: 10.1080/01431160802672864.

Chan, J. C.-W. and Paelinckx, D. (2008) 'Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery', *Remote Sensing of Environment*, 112(6), pp. 2999–3011. doi: 10.1016/j.rse.2008.02.011.

Chawla, N. V, Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. (2002) 'SMOTE : Synthetic Minority Over-sampling Technique', *Journal of Artificial Intelligence Research*, 16, pp. 321–357.

Chen, C., Liaw, A. and Breiman, L. (2004) *Using Random Forest to Learn Imbalanced Data*. Statistics Department, University of California at Berkeley. Available at: http://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf.

Chen, D. and Stow, D. (2002) 'The effect of training strategies on supervised classification at different spatial resolutions', *Photogrammetric Engineering & Remote Sensing*, 68(11), pp. 1155–1161. Available at: http://www.asprs.org/a/publications/pers/2002journal/november/2002_nov_1155-1161.pdf (Accessed: 19 October 2012).

Chen, J., Chen, J., Liao, A., Cao, X., Chen, L., Chen, X., He, C., Han, G., Peng, S., Lu, M., Zhang, W., Tong, X. and Mills, J. (2015) 'Global land cover mapping at 30 m resolution: A POK-based operational approach', *ISPRS Journal of Photogrammetry and Remote Sensing*. International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS), 103, pp. 7–27. doi: 10.1016/j.isprsjprs.2014.09.002.

Chutia, D., Bhattacharyya, D. K., Sarma, K. K., Kalita, R. and Sudhakar, S. (2016) 'Hyperspectral Remote Sensing Classifications: A Perspective Survey', *Transactions in GIS*, 20(4), pp. 463–490. doi: 10.1111/tgis.12164.

Clerici, N., Weissteiner, C. J. and Gerard, F. (2012) 'Exploring the Use of MODIS NDVI-Based Phenology Indicators for Classifying Forest General Habitat Categories', *Remote Sensing*, 4(6), pp. 1781–1803. doi: 10.3390/rs4061781.

Coburn, C. a. and Roberts, a. C. B. (2004) 'A multiscale texture analysis procedure for improved forest stand classification', *International Journal of Remote Sensing*, 25(20), pp. 4287–4308. doi: 10.1080/0143116042000192367.

Cochran, W. G. (1977) *Sampling Techniques*. 3rd edn. New York, USA: Wiley.

Coelho, L. P. (2017) *MILK: Machine Learning Toolkit*. Available at: https://pythonhosted.org/milk/.

Colditz, R. R. (2015) 'An evaluation of different training sample allocation schemes for discrete and continuous land cover classification using decision tree-based algorithms', *Remote Sensing*, 7(8), pp. 9655–9681. doi: 10.3390/rs70809655.

Coops, N. C., Gillanders, S. N., Wulder, M. a., Gergel, S. E., Nelson, T. and Goodwin, N. R. (2010) 'Assessing changes in forest fragmentation following infestation using time series Landsat imagery', *Forest Ecology and Management*. Elsevier B.V., 259(12), pp. 2355–2365. doi: 10.1016/j.foreco.2010.03.008.

Corcoran, J. M., Knight, J. F. and Gallant, A. L. (2013) 'Influence of multi-source and multi-temporal remotely sensed and ancillary data on the accuracy of random forest classification of wetlands in northern Minnesota', *Remote Sensing*, 5(7), pp. 3212–3238. doi: 10.3390/rs5073212.

Crist, E. P. and Cicone, R. C. (1984) 'A Physically-Based Transformation of Thematic Mapper Data---The TM Tasseled Cap', *IEEE Transactions on Geoscience and Remote Sensing*, GE-22(3), pp. 256–263. doi: 10.1109/TGRS.1984.350619.

CSIRO (2011) *One-second SRTM digital elevation model*. Available at: http://www.csiro.au/Outcomes/Water/Water-information-systems/One-second-SRTM-Digital-Elevation-Model.aspx.

Culbert, P., Pidgeon, A. and St-Louis, V. (2009) 'The impact of phenological variation on texture measures of remotely sensed imagery', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2(4), pp. 299–309. doi: 10.1109/JSTARS.2009.2021959.

Cummings, B. and Hardy, A. (2000) *Revision of the Interim Biogeographic Regionalisation for Australia (IBRA) and Development of Version 5.1*. Canberra.

Cutler, D., Jr, T. E. and Beard, K. (2007) 'Random forests for classification in ecology', *Ecology*, 88(11), pp. 2783–2792. Available at: http://www.esajournals.org/doi/abs/10.1890/07-0539.1 (Accessed: 13 September 2012).

Dalponte, M., Orka, H. O., Gobakken, T., Gianelle, D. and Naesset, E. (2013) 'Tree Species Classification in Boreal Forests With Hyperspectral Data', *IEEE Transactions on Geoscience and Remote Sensing*, 51(5), pp. 2632–2645. doi: 10.1109/TGRS.2012.2216272.

DeFries, R. and Cheung-Wai Chan, J. (2000) 'Multiple Criteria for Evaluating Machine Learning Algorithms for Land Cover Classification from Satellite Data', *Remote Sensing of Environment*, 74(3), pp. 503–515. doi: 10.1016/S0034-4257(00)00142-5.

DeFries, R. S. and Townshend, J. G. R. (1994) 'NDVI derived land cover classifications at a global scale', *International Journal of Remote Sensing*, 5, pp. 3567 – 3586.

Delaney, J. L. and Skidmore, A. K. (1998) 'Discrepancy or error in forest type classification', *Australian Forestry*. Taylor & Francis, 61(2), pp. 82–88. doi: 10.1080/00049158.1998.10674724.

Deng, C. and Wu, C. (2013) 'The use of single-date MODIS imagery for estimating large-scale urban impervious surface fraction with spectral mixture analysis and machine learning techniques', *ISPRS Journal of Photogrammetry and Remote Sensing*, 86, pp. 100–110. doi: 10.1016/j.isprsjprs.2013.09.010.

Department of Agriculture Fisheries and Forestry (2012) *Australia's forest at a glance*. Canberra.

Deppe, F. (1998) 'Forest Area Estimation Using Sample Surveys and Landsat MSS and TM Data', *Photogrammetric Engineering & Remote Sensing*, 64(4).

Dietterich, T. (2000) 'An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization', *Machine learning*, 40, pp. 139–157. Available at: http://www.springerlink.com/index/h045561mr84u1018.pdf (Accessed: 12 December 2011).

Díez-Pastor, J. F., Rodríguez, J. J., García-Osorio, C. I. and Kuncheva, L. I. (2015) 'Diversity techniques improve the performance of the best imbalance learning ensembles', *Information Sciences*, 325, pp. 98–117. doi: 10.1016/j.ins.2015.07.025.

DSE (2000) *VICGRID94 Map Projection*. Available at: http://www.dse.vic.gov.au/__data/assets/pdf_file/0012/117102/VICGRID94MapProjectionInformation.pdf.

Du, P., Samat, A., Waske, B., Liu, S. and Li, Z. (2015a) 'Random Forest and Rotation Forest for fully polarized SAR image classification using polarimetric and spatial features', *ISPRS Journal of Photogrammetry and Remote Sensing*, 105, pp. 38–53. doi: 10.1016/j.isprsjprs.2015.03.002.

Du, P., Samat, A., Waske, B., Liu, S. and Li, Z. (2015b) 'Random Forest and Rotation Forest for fully polarized SAR image classification using polarimetric and spatial features', *ISPRS Journal of Photogrammetry and Remote Sensing*, 105, pp. 38–53. doi: 10.1016/j.isprsjprs.2015.03.002.

Eckert, S. (2012) 'Improved Forest Biomass and Carbon Estimations Using Texture Measures from WorldView-2 Satellite Data', *Remote Sensing*, 4(4), pp. 810–829. doi: 10.3390/rs4040810.

Efron, B. and Tibshirani, R. J. (1994) *An Introduction to the Bootstrap*. Chapman & Hall/CRC.

Egorov, A. V, Hansen, M. C., Roy, D. P., Kommareddy, A. and Potapov, P. V (2015) 'Image interpretation-guided supervised classification using nested segmentation', *Remote Sensing of Environment*, 165, pp. 135–147. doi: http://doi.org/10.1016/j.rse.2015.04.022.

Elghazel, H., Aussem, A. and Perraud, F. (2011) 'Trading-Off Diversity and Accuracy for Optimal Ensemble Tree Selection in Random Forests', in Okun, O., Valentini, G., and Re, M. (eds) *Ensembles in Machine Learning Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg (Studies in Computational Intelligence), pp. 169–179. doi: 10.1007/978-3-642-22910-7.

Estabrooks, A., Jo, T. and Japkowicz, N. (2004) 'A Multiple Resampling Method for Learning from Imbalanced Data Sets', *Computational Intelligence*, 20(1), pp. 18–36. doi: 10.1111/j.0824-7935.2004.t01-1-00228.x.

European Space Agency (2016) *European Space Agency*. Available at: http://www.esa.int/Our_Activities/Observing_the_Earth/Copernicus/Overview4 (Accessed: 25 September 2016).

Evans, J. S. and Cushman, S. . (2009) 'Gradient modeling of conifer species using random forests', *Landscape Ecology2*, 24, pp. 673–683.

Fahsi, A., Tsegaye, T., Tadesse, W. and Coleman, T. (2000) 'Incorporation of digital elevation models with Landsat-TM data to improve land cover classification accuracy', *Forest Ecology and Management*, 128, pp. 57–64. Available at: http://www.sciencedirect.com/science/article/pii/S0378112799002728 (Accessed: 19 October 2012).

FAO (2000) *FRA 2000 on definitions of forest and forest change*, *FRA 2000 on definitions of forest and forest change*. Available at: http://www.fao.org/docrep/006/ad665e/ad665e06.htm.

Farmer, E., Jones, S., Clarke, C., Buxton, L., Soto-Berelov, M., Page, S., Mellor, A. and Haywood, A. (2013) 'Creating a large area landcover dataset for public land monitoring and reporting', in Arrowsmith, C. et al. (eds) *Progress in Geospatial Science Research*. Melbourne: Publishing Solutions, pp. 85–98.

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D. and Alsdorf, D. (2007) 'The Shuttle Radar Topography Mission', *Reviews of Geophysics*, 45(2), p. RG2004. doi: 10.1029/2005RG000183.

Flood, N., Danaher, T., Gill, T. and Gillingham, S. (2013) 'An Operational Scheme for Deriving Standardised Surface Reflectance from Landsat TM/ETM+ and SPOT HRG Imagery for Eastern Australia', *Remote Sensing*, 5(1), pp. 83–109. doi: 10.3390/rs5010083.

Food and Agriculture Organization of the United Nations (2001) *Global forest resources assessment 2000*.

Food and Agriculture Organization of the United Nations (2015) *State of Europe's Forests 2015*. Madrid. Available at: http://www.foresteurope.org/docs/fullsoef2015.pdf.

Foody, G. (1999) 'The significance of border training patterns in classification by a feedforward neural network using back propagation learning', *International Journal of Remote Sensing*, 20(18), pp. 3549–3562. Available at: http://www.tandfonline.com/doi/abs/10.1080/014311699211192.

Foody, G. (2004) 'Thematic map comparison: evaluating the statistical significance of differences in classification accuracy', *Photogrammetric Engineering & Remote Sensing*, 70(5), pp. 627–633. Available at: http://asprs.org/a/publications/pers/2004journal/may/2004_may_627-633.pdf (Accessed: 19 June 2013).

Foody, G. and Arora, M. (1997) 'An evaluation of some factors affecting the accuracy of classification by an artificial neural network', *International Journal of Remote Sensing*, 18(4), pp. 799–810. Available at: http://www.tandfonline.com/doi/abs/10.1080/014311697218764 (Accessed: 10 November 2012).

Foody, G. and Cutler, M. (2006) 'Mapping the species richness and composition of tropical forests from remotely sensed data with neural networks', *Ecological Modelling*, 195(1–2), pp. 37–42. doi: 10.1016/j.ecolmodel.2005.11.007.

Foody, G. M. (2002) 'Status of land cover classification accuracy assessment', *Remote Sensing of Environment*. Elsevier, 80(1), pp. 185–201. Available at: http://dx.doi.org/10.1016/S0034-4257(01)00295-4.

Foody, G. M. and Mathur, A. (2004) 'Toward intelligent training of supervised image classifications: directing training data acquisition for SVM classification', *Remote Sensing of Environment*, 93(1–2), pp. 107–117. doi: 10.1016/j.rse.2004.06.017.

Foody, G., McCulloch, M. and Yates, W. (1995) 'Classification of remotely sensed data by an artificial neural network: issues related to training data characteristics', *Photogrammetric Engineering & Remote Sensing*, 61(4), pp. 391–401. Available at: http://cat.inist.fr/?aModele=afficheN&cpsidt=3496683 (Accessed: 9 November 2012).

Foody, G., Pal, M., Rocchini, D., Garzon-Lopez, C. and Bastin, L. (2016) 'The Sensitivity of Mapping Methods to Reference Data Quality: Training Supervised Image Classifications with Imperfect Reference Data', *ISPRS International Journal of Geo-Information*, 5(11), p. 199. doi: 10.3390/ijgi5110199.

Franklin, J. (1995) 'Predictive vegetation mapping: Geographic modelling of biospatial patterns in relation to environmental gradients', *Progress in Physical Geography*, 19, pp. 474–499.

Freeman, E. A. and Moisen, G. (2008) 'PresenceAbsence: An R Package for Presence Absence

Analysis.', *Journal of Statistical Software*, 23(11), pp. 1–31. Available at: http://ddr.nal.usda.gov/handle/10113/35166 (Accessed: 31 May 2013).

Freeman, E. a., Moisen, G. G. and Frescino, T. S. (2012) 'Evaluating effectiveness of down-sampling for stratified designs and unbalanced prevalence in Random Forest models of tree species distributions in Nevada', *Ecological Modelling*. Elsevier B.V., 233, pp. 1–10. doi: 10.1016/j.ecolmodel.2012.03.007.

Gautier, L. (2012) 'Rpy'. Available at: http://rpy.sourceforge.net/.

Gergel, S. E. (2007) 'New Directions in Landscape Pattern Analysis and Linkages with Remote Sensing', in Wulder, M. A. and Franklin, S. E. (eds) *Understanding Forest Disturbance and Spatial Pattern*. Taylor & Francis, p. 246.

Ghimire, B., Rogan, J., Galiano, V. R., Panday, P. and Neeti, N. (2012) 'An Evaluation of Bagging, Boosting, and Random Forests for Land-Cover Classification in Cape Cod, Massachusetts, USA', *GIScience & Remote Sensing*, 49(5), pp. 623–643. doi: 10.2747/1548-1603.49.5.623.

Gislason, P., Benediktsson, J. and Sveinsson, J. (2006) 'Random Forests for land cover classification', *Pattern Recognition Letters*, 27(4), pp. 294–300. doi: 10.1016/j.patrec.2005.08.011.

Gomez, C., White, J. C. and Wulder, M. A. (2016) 'Optical remotely sensed time series data for land cover classification: A review', *ISPRS Journal of Photogrammetry and Remote Sensing*, pp. 55–72. doi: 10.1016/j.isprsjprs.2016.03.008.

Google Earth Engine Team (2015) *Google Earth Engine: A planetary-scale geo-spatial analysis platform*. Available at: https://earthengine.google.com.

GRASS Development Team (2012) 'Geographic Resources Analysis Support System (GRASS) Software'. Source Geospatial Foundation. Available at: http://grass.osgeo.org.

Graves, S. J., Asner, G. P., Martin, R. E., Anderson, C. B., Colgan, M. S., Kalantari, L. and Bohlman, S. A. (2016) 'Tree species abundance predictions in a tropical agricultural landscape with a supervised classification model and imbalanced data', *Remote Sensing*, 8(2). doi: 10.3390/rs8020161.

Green, G. M. and Sussman, R. (1990) 'Deforestation history of the eastern rainforests of Madagascar from satellite images', *Science*, 248, pp. 212–215.

Guisan, A. and Zimmermann, N. E. (2000) 'Predictive habitat distribution models in ecology', *Ecological Modelling*, 135(2–3), pp. 147–186. doi: 10.1016/S0304-3800(00)00354-9.

Guo, L. (2011) *Margin framework for ensemble classifiers. Application to remote sensing data*. University of Bordeaux, France.

Guo, L. and Boukir, S. (2013) 'Margin-based ordered aggregation for ensemble pruning', *Pattern Recognition Letters*, 34(6), pp. 603–609. doi: 10.1016/j.patrec.2013.01.003.

Guo, L. and Boukir, S. (2014) 'Ensemble margin framework for image classification', in *ICIP 2014, IEEE International Conference on Image Processing*, pp. 4231–4235. Available at:

http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7025859 (Accessed: 24 May 2015).

Guo, L., Chehata, N., Mallet, C. and Boukir, S. (2011) 'Relevance of airborne lidar and multispectral image data for urban scene classification using Random Forests', *ISPRS Journal of Photogrammetry and Remote Sensing*. Elsevier B.V., 66(1), pp. 56–66. doi: 10.1016/j.isprsjprs.2010.08.007.

Haddad, N., Brudvig, L., Clobert, J., Davies, K., Gonzalez, A., Holt, R., Lovejoy, T., Sexton, J., Austin, M., Collins, C., Cook, W., Damschen, E., Ewers, R., Foster, B., Jenkins, C., King, A., Laurance, W., Levey, D., Margules, C., Melbourne, B., Nicholls, A., Orrock, J., Song, D. and Townshend, J. (2015) 'Habitat fragmentation and its lasting impact on Earth's ecosystems', *Science Advances*, 1(2), pp. 1–9. Available at: http://advances.sciencemag.org/content/1/2/e1500052.short (Accessed: 12 April 2015).

Ham, J., Chen, Y., Crawford, M. M., Member, S. and Ghosh, J. (2005) 'Investigation of the Random Forest Framework for Classification of Hyperspectral Data', *IEEE Transactions on Geoscience and Remote Sensing*, 43(3), pp. 492–501.

Hansen, L. and Salamon, P. (1990) 'Neural network ensembles', *IEEE transactions on pattern analysis and Machine Intelligence*, 12(October), pp. 993–1001. Available at: http://www.computer.org/csdl/trans/tp/1990/10/i0993.pdf (Accessed: 1 February 2015).

Hansen, M. C. and Loveland, T. R. (2012) 'A review of large area monitoring of land cover change using Landsat data', *Remote Sensing of Environment*. Elsevier Inc., 122, pp. 66–74. doi: 10.1016/j.rse.2011.08.024.

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., Thau, D., Stehman, S. V., Goetz, S. J., Loveland, T. R., Kommareddy, A., Egorov, A., Chini, L., Justice, C. O. and Townshend, J. R. G. (2013) 'High-Resolution Global Maps of 21st-Century Forest Cover Change', *Science*, 342(6160), pp. 850–853. doi: 10.1126/science.1244693.

Haralich, R. M. (1979) 'Statistical and structural approach to texture', *Proc. IEEE*, 67(5), pp. 786–804.

Haralick, R. M. (1979) 'Statistical and structural approaches to texture', *Proceedings of the IEEE*, 67(5), pp. 786–804.

Haywood, A., Mellor, A. and Stone, C. (2016) 'A strategic forest inventory for public land in Victoria, Australia', *Forest Ecology and Management*. Elsevier B.V., 367, pp. 86–96. doi: 10.1016/j.foreco.2016.02.026.

Haywood, A. and Stone, C. (2011) 'Mapping eucalypt forest susceptible to dieback associated with bell miners (Manorina melanophys) using laser scanning , SPOT 5 and ancillary topographical data', *Ecological Modelling*. Elsevier B.V., 222(5), pp. 1174–1184. doi: 10.1016/j.ecolmodel.2010.12.012.

Haywood, A. and Stone, C. (2017) 'Estimating Large Area Forest Carbon Stocks — A Pragmatic Design Based Strategy', *Forests*, 8(99), pp. 1–14. doi: 10.3390/f8040099.

Haywood, A., Thrum, K., Mellor, A. and Stone, C. (2017) 'Monitoring Victoria's public forests: implementation of the Victorian Forest Monitoring Program', *Southern Forests*, In Press. doi:

10.2989/20702620.2017.1318344.

He, H. and Garcia, E. A. (2009) 'Learning from Imbalanced Data', *IEEE Transactions on Knowledge and Data Engineering*, 21(9), pp. 1263–1284. doi: 10.1109/TKDE.2008.239.

Hess, G. (1994) 'Pattern and error in landscape ecology: a commentary', *Landscape Ecology*, 9(1), pp. 3–5.

Houlder, D. (2001) 'ANUCLIM (version 5.1)'. Canberra: Centre for Resource and Environment Studies, Australia National University.

Howell, C. I., Wilson, A. D., Davey, S. M. and Eddington, M. M. (2008) 'Sustainable forest management reporting in Australia', *Ecological Indicators*, 8(2), pp. 123–130. doi: 10.1016/j.ecolind.2006.11.004.

Huang, B. F. F. and Boutros, P. C. (2016) 'The parameter sensitivity of random forests', *BMC Bioinformatics*, 17(1), p. 331. doi: 10.1186/s12859-016-1228-x.

Inglada, J., Vincent, A., Arias, M., Tardy, B., Morin, D. and Rodes, I. (2017) 'Operational High Resolution Land Cover Map Production at the Country Scale Using Satellite Image Time Series', *Remote Sensing*, 9(1), p. 95. doi: 10.3390/rs9010095.

Jacobs, M. R. (1955) *Growth habits of the Eucalypts*. Canberra: Forestry and Timber Bureau.

Japkowicz, N. and Stephen, S. (2002) 'The class imbalance problem: A systematic study', *Intelligent Data Analysis*, 6(5), pp. 429–450.

Jenkins, R. B. and Coops, N. C. (2011) 'Landscape Controls on Structural Variation in Eucalypt Vegetation Communities: Woronora Plateau, Australia', *Australian Geographer*, 42(1), pp. 1–17. doi: 10.1080/00049182.2011.546316.

Jin, H., Stehman, S. V and Mountrakis, G. (2014) 'Assessing the impact of training sample selection on accuracy of an urban classification : a case study in Denver , Colorado', *International Journal of Remote Sensing*. Taylor & Francis, 35(6), pp. 2067–2081. doi: 10.1080/01431161.2014.885152.

Joy, S. M., Reich, R. M. and Reynolds, R. T. (2003) 'A non-parametric supervised classification of vegetation types on the Kaibab National Forest using decision trees', *International Journal of Remote Sensing*, 24(9), p. 1835−1852.

Jupp, D. L. B. and Walker, J. (1997) 'Detecting Structural and Growth Changes in Woodlands and Forests: The Challenge for Remote Sensing and the Role of Geometric-Optical Modelling', in Shimoda, H., Gholz, H. L., and Nakane, K. (eds) *The Use of Remote Sensing in the Modeling of Forest Productivity*. Springer Netherlands, pp. 75–108. doi: 10.1007/978-94-011-5446-8_4.

Kapp, M. N., Sabourin, R. and Maupin, P. (2007) 'An empirical study on diversity measures and margin theory for ensembles of classifiers', in *2007 10th International Conference on Information Fusion*. IEEE, pp. 1–8. doi: 10.1109/ICIF.2007.4408144.

Kayitakire, F., Hamel, C. and Defourny, P. (2006) 'Retrieving forest structure variables based on image texture analysis and IKONOS-2 imagery', *Remote Sensing of Environment*, 102(3–4), pp.

390–401. doi: 10.1016/j.rse.2006.02.022.

Khalyani, A. H., Falkowski, M. J. and Mayer, A. L. (2012) 'Classification of Landsat images based on spectral and topographic variables for land-cover change detection in Zagros forests', *International Journal of Remote Sensing*, 33(21), pp. 6956–6974. Available at: http://www.tandfonline.com/doi/abs/10.1080/01431161.2012.695095 (Accessed: 21 May 2013).

Kleindl, W. J., Powell, S. L. and Hauer, F. R. (2015) 'Effect of thematic map misclassification on landscape multi-metric assessment', *Environmental Monitoring and Assessment*, 187(321). doi: 10.1007/s10661-015-4546-y.

Kohavi, R. and Wolpert, D. (1996) 'Bias plus variance decomposition for zero-one loss functions', in *13th International Conference of Machine Learning, ICML '96*, pp. 275–283. Available at: http://ai.stanford.edu/~ronnyk/biasVar.pdf (Accessed: 30 May 2015).

Kumar, N., Lester, D., Marchetti, A., Hammann, G. and Longmont, A. (2013) *Demystifying cloud computing for remote sensing application*, *Earth Imaging Journal*. Available at: http://eijournal.com/print/column/industry-insights/demystifying-cloud-computing-for-remote-sensing-applications.

Kuncheva, L. and Whitaker, C. (2003) 'Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy', *Machine learning*, 51, pp. 181–207. Available at: http://link.springer.com/article/10.1023/A:1022859003006 (Accessed: 23 January 2014).

Langford, W. T., Gergel, S. E., Dietterich, T. G. and Cohen, W. (2006) 'Map Misclassification Can Cause Large Errors in Landscape Pattern Indices: Examples from Habitat Fragmentation', *Ecosystems*, 9(3), pp. 474–488. doi: 10.1007/s10021-005-0119-1.

Lary, D. J., Alavi, A. H., Gandomi, A. H. and Walker, A. L. (2016) 'Machine learning in geosciences and remote sensing', *Geoscience Frontiers*. Elsevier Ltd, 7(1), pp. 3–10. doi: 10.1016/j.gsf.2015.07.003.

Lawrence, R. L. and Wright, A. (2001) 'Rule-based classification systems using classification and regression tree (CART)', *Photogrammetric Engineering & Remote Sensing*, 67(10), pp. 1137–1142.

Lechner, A. M., Langford, W. T., Bekessy, S. A. and Jones, S. D. (2012) 'Are landscape ecologists addressing uncertainty in their remote sensing data?', *Landscape Ecology*, 27(9), pp. 1249–1261. doi: 10.1007/s10980-012-9791-7.

Lechner, A. M., Reinke, K. J., Wang, Y. and Bastin, L. (2013) 'Interactions between landcover pattern and geospatial processing methods: Effects on landscape metrics and classification accuracy', *Ecological Complexity*. Elsevier B.V., 15, pp. 71–82. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1476945X13000342 (Accessed: 22 April 2013).

Lechner, A. M., Stein, A., Jones, S. D. and Ferwerda, J. G. (2009) 'Remote sensing of small and linear features: Quantifying the effects of patch size and length, grid position and detectability on land cover mapping', *Remote Sensing of Environment*. Elsevier Inc., 113(10), pp. 2194–2204. doi: 10.1016/j.rse.2009.06.002.

Li, J. and Roy, D. P. (2017) 'A Global Analysis of Sentinel-2A, Sentinel-2B and Landsat-8 Data Revisit Intervals and Implications for Terrestrial Monitoring', *Remote Sensing*, 9(9), p. 902. doi: 10.3390/rs9090902.

Liaw, A. and Wiener, M. (2002) 'Classification and Regression by randomForest', *R News*, 2(3), pp. 18–22. Available at: http://cran.r-project.org/doc/Rnews/.

Lillesand, T. M. and Kiefer, R. W. (1994) *Remote Sensing and Image Interpretation*. 3rd edn. John Wiley & Sons Ltd.

Ling, C. X. and Li, C. (1998) 'Data mining for direct marketing: Problems and solutions.', in *Kdd*. New York, pp. 73–79.

Lippitt, C., Rogan, J. and Li, Z. (2008) 'Mapping selective logging in mixed deciduous forest: a comparison of machine learning algorithms', *Photogrammetric Engineering & Remote Sensing*, 74(10), pp. 1201–1211. Available at: http://www.eastwestcenter.org/fileadmin/stored/pics/Lippitt_etal_PE&RS.pdf (Accessed: 22 April 2013).

Liu, W., Song, C., Schroeder, T. A. and Cohen, W. B. (2008) 'Predicting forest successional stages using multitemporal Landsat imagery with forest inventory and analysis data', *International Journal of Remote Sensing*, 29(13), pp. 3855–3872.

Lowell, K., Woodgate, P., Richards, G., Jones, S. D. and Buxton, L. (2005) 'Fuzzy reliability assessment of multi-period land-cover change maps', *Photogrammetric Engineering & Remote Sensing*, 71, pp. 939–945.

Lowry, J., Ramsey, R. D., Thomas, K., Schrupp, D., Sajwaj, T., Kirby, J., Waller, E., Schrader, S., Falzarano, S., Langs, L., Manis, G., Wallace, C., Schulz, K., Comer, P., Pohs, K., Rieth, W., Velasquez, C., Wolk, B., Kepner, W., Boykin, K., O'Brien, L., Bradford, D., Thompson, B. and Prior-Magee, J. (2007) 'Mapping moderate-scale land-cover over very large geographic areas within a collaborative framework: A case study of the Southwest Regional Gap Analysis Project (SWReGAP)', *Remote Sensing of Environment*, 108(1), pp. 59–73. doi: 10.1016/j.rse.2006.11.008.

Lu, D. (2005) 'Aboveground biomass estimation using Landsat TM data in the Brazilian Amazon', *International Journal of Remote Sensing*, 26(12), pp. 2509–2525. doi: 10.1080/01431160500142145.

Lu, D., Mausel, P., Brondízio, E. and Moran, E. (2003) 'Classification of successional forest stages in the Brazilian Amazon basin', *Forest Ecology and Management*, 181, pp. 301–312.

Lu, D. and Weng, Q. (2007) 'A survey of image classification methods and techniques for improving classification performance', *International Journal of Remote Sensing*, 28(5), pp. 823–870. doi: 10.1080/01431160600746456.

Main-Knorn, M., Moisen, G. G., Healey, S. P., Keeton, W. S., Freeman, E. a. and Hostert, P. (2011) 'Evaluating the Remote Sensing and Inventory-Based Estimation of Biomass in the Western Carpathians', *Remote Sensing*, 3(7), pp. 1427–1446. doi: 10.3390/rs3071427.

Marceau, D. J., Gratton, D. J., Fournier, R. A. and Fortin, J.-P. (1994) 'Remote sensing and the

measurement of geographical entities in a forested environment. 2. The optimal spatial resolution', *Remote Sensing of Environment*, 49(2), pp. 105–117. doi: 10.1016/0034-4257(94)90047-7.

Maselli, F. (2011) 'Use of MODIS NDVI data to improve forest-area estimation', *International Journal of Remote Sensing*, 32(21), pp. 6379–6393. Available at: http://www.tandfonline.com/doi/abs/10.1080/01431161.2010.510490 (Accessed: 11 May 2012).

McGarigal, K. and Marks, B. J. (1995) *FRAGSTATS: Spatial Pattern Analysis Program for Quantifying Landscape Structure*. Pacific Northwest Research Station, Portland, OR.

McRoberts, R. E. (2010) 'Probability- and model-based approaches to inference for proportion forest using satellite imagery as ancillary data', *Remote Sensing of Environment*. Elsevier B.V., 114(5), pp. 1017–1025. doi: 10.1016/j.rse.2009.12.013.

McRoberts, R. E. (2011) 'Satellite image-based maps: Scientific inference or pretty pictures?', *Remote Sensing of Environment*. Elsevier B.V., 115(2), pp. 715–724. doi: 10.1016/j.rse.2010.10.013.

McRoberts, R. E., Holden, G. R., Nelson, M. D., Liknes, G. C. and Gormanson, D. D. (2005) 'Using satellite imagery as ancillary data for increasing the precision of estimates for the Forest Inventory and Analysis program of the USDA Forest Service', *Canadian Journal of Forest Research*. NRC Research Press, 35(12), pp. 2968–2980. Available at: http://www.nrcresearchpress.com/doi/abs/10.1139/x05-222 (Accessed: 10 September 2011).

McRoberts, R. E. and Liknes, G. C. (2002) 'Assessing the Effects of Forest Fragmentation Using Satellite Imagery and Forest Inventory Data', *Proceedings of the Fourth Annual Forest Inventory and Analysis Symposium*, pp. 117–120.

McRoberts, R. and Tomppo, E. (2007) 'Remote sensing support for national forest inventories', *Remote Sensing of Environment*, 110(4), pp. 412–419. doi: 10.1016/j.rse.2006.09.034.

Mellor, A. and Boukir, S. (2017) 'ISPRS Journal of Photogrammetry and Remote Sensing Exploring diversity in ensemble classification : Applications in large area land cover mapping', *ISPRS Journal of Photogrammetry and Remote Sensing*. International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS), 129, pp. 151–161. doi: 10.1016/j.isprsjprs.2017.04.017.

Mellor, A., Boukir, S., Haywood, A. and Jones, S. (2014) 'Using ensemble margin to explore issues of training data imbalance and mislabeling on large area land cover classification', in *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, pp. 5067–5071. doi: 10.1109/ICIP.2014.7026026.

Mellor, A., Boukir, S., Haywood, A. and Jones, S. (2015) 'Exploring issues of training data imbalance and mislabelling on random forest performance for large area land cover classification using the ensemble margin', *ISPRS Journal of Photogrammetry and Remote Sensing*, 105, pp. 155–168. doi: 10.1016/j.isprsjprs.2015.03.014.

Mellor, A. and Haywood, A. (2010) 'Remote sensing Victoria's public land forests - a two tiered synoptic approach', in *Proceedings of the 15th Australian Remote Sensing and Photogrammetry*

Conference. Alice Springs.

Mellor, A., Haywood, A., Stone, C. and Jones, S. (2013) 'The Performance of Random Forests in an Operational Setting for Large Area Sclerophyll Forest Classification', *Remote Sensing*, 5(6), pp. 2838–2856. doi: 10.3390/rs5062838.

Melville, P. and Mooney, R. J. (2005) 'Creating diversity in ensembles using artificial data', *Information Fusion*, 6(1), pp. 99–111. doi: 10.1016/j.inffus.2004.04.001.

Millard, K. and Richardson, M. (2015) 'On the importance of training data sample selection in Random Forest image classification: A case study in peatland ecosystem mapping', *Remote Sensing*, 7(7), pp. 8489–8515. doi: 10.3390/rs70708489.

Montreal Process Implementation Group for Australia (2008) *Australia's State of the Forests Report 2008*. Canberra.

Montréal Process Working Group (2015) *The Montréal Process: Criteria and Indicators for the Conservation and Sustainable Management of Temperate and Boreal Forests (5th edition)*.

Morgan, J. L., Gergel, S. E. and Coops, N. C. (2010) 'Aerial Photography: A Rapidly Evolving Tool for Ecological Management', *BioScience*, 60(1), pp. 47–59. doi: 10.1525/bio.2010.60.1.9.

Myers, N. (1996) 'The world's forests: problems and potentials', *Environmental Conservation*, 23(2), pp. 156–168.

Na, X., Zang, S. and Wang, J. (2009) 'Evalution of Random Forest Ensemble Classification for Land Cover Mapping Using TM and Ancillary Geographical Data', *2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery*. Ieee, pp. 89–93. doi: 10.1109/FSKD.2009.165.

NASA (2016) *MODIS*. Available at: https://modis.gsfc.nasa.gov/about/ (Accessed: 25 September 2016).

National Forest Inventory (2003) *National Forest Inventory*. Canberra.

Opitz, D. and Maclin, R. (1999) 'Popular ensemble methods: An empirical study', *Journal of Artificial Intelligence Research*, 11, pp. 169–198. Available at: http://arxiv.org/abs/1106.0257 (Accessed: 28 March 2014).

Paget, M. J. and King, E. A. (2008) *MODIS land data sets for the Australian region*. Canberra. Available at: http://www-data.wron.csiro.au/rs/MODIS/LPDAAC/data/ ModisLand_PagetKing_20081203-final.pdf.

Pal, M. (2005) 'Random forest classifier for remote sensing classification', *International Journal of Remote Sensing*, 26(1), pp. 217–222. doi: 10.1080/01431160412331269698.

Pal, M. and Mather, P. M. (2006) 'Some issues in the classification of DAIS hyperspectral data', *International Journal of Remote Sensing*, 27(14), pp. 2895–2916. doi: 10.1080/01431160500185227.

Parnell, L. D., Lindenbaum, P., Shameer, K., Dall'Olio, G. M., Swan, D. C., Jensen, L. J., Cockell, S.

J., Pedersen, B. S., Mangan, M. E., Miller, C. A. and Albert, I. (2011) 'BioStar: An Online Question & Answer Resource for the Bioinformatics Community', *PLoS Computational Biology*. Edited by P. E. Bourne, 7(10), p. e1002216. doi: 10.1371/journal.pcbi.1002216.

Pearce, J. and Ferrier, S. (2000) 'Evaluating the predictive performance of habitat models developed using logistic regression', *Ecological Modelling*, 133, pp. 225–245.

Pelletier, C., Valero, S., Inglada, J., Champion, N., Sicre, C. M. and Dedieu, G. (2017) 'Effect of training class label noise on classification performances for land cover mapping with satellite image time series', *Remote Sensing*, 9(2). doi: 10.3390/rs9020173.

Peters, J., Verhoest, N. E. C., Samson, R., Van Meirvenne, M., Cockx, L. and De Baets, B. (2009) 'Uncertainty propagation in vegetation distribution models based on ensemble classifiers', *Ecological Modelling*, 220(6), pp. 791–804. doi: 10.1016/j.ecolmodel.2008.12.022.

Pflugmacher, D., Cohen, W. B. and E. Kennedy, R. (2012) 'Using Landsat-derived disturbance history (1972–2010) to predict current forest structure', *Remote Sensing of Environment*. Elsevier Inc., 122, pp. 146–165. doi: 10.1016/j.rse.2011.09.025.

van der Ploeg, T., Austin, P. C. and Steyerberg, E. W. (2014) 'Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints', *BMC Medical Research Methodology*, 14(1), p. 137. doi: 10.1186/1471-2288-14-137.

Polikar, R. (2006) 'Ensemble based systems in decision making', *IEEE Circuits and Systems Magazine*, 6(3), pp. 21–45. doi: 10.1109/MCAS.2006.1688199.

Proisy, C., Couteron, P. and Fromard, F. (2007) 'Predicting and mapping mangrove biomass from canopy grain analysis using Fourier-based textural ordination of IKONOS images', *Remote Sensing of Environment*, 109(3), pp. 379–392. doi: https://doi.org/10.1016/j.rse.2007.01.009.

Provost, F. (2000) 'Machine Learning from Imbalanced Data Sets 101 (Extended Abstract)', in *Proceedings of the AAAI-2000 Workshop on Imbalanced Data Sets*.

Python Software Foundation (2011) *The Python Language Reference*. Available at: http://docs.python.org/release/3.2/reference/index.html.

Qi, Y., Bar-Joseph, Z. and Klein-Seetharaman, J. (2006) 'Evaluation of different biological data and computational classification methods for use in protein interaction prediction', *Proteins: Structure, Function, and Bioinformatics*, 63(3), pp. 490–500. doi: 10.1002/prot.20865.

R Core Team (2013) 'R: A language and environment for statistical computing. R Foundation for Statistical Computing'. Vienna, Austria. Available at: http://www.r-project.org/.

R Development Core Team (2011) 'R: A Language and Environment for Statistical Computing'. Vienna, Austria. Available at: http://www.r-project.org.

Riitters, K. H., Coulston, J. W. and Wickham, J. D. (2012) 'Fragmentation of forest communities in the eastern United States', *Forest Ecology and Management*. Elsevier B.V., 263, pp. 85–93. doi: 10.1016/j.foreco.2011.09.022.

Riitters, K., Wickham, J., O'Neill, R., Jones, B. and Smith, E. (2000) 'Global-Scale Patterns of

Forest Fragmentation', *Conservation Ecology*, 43(2). Available at: http://www.consecol.org/vol4/iss2/art3/.

Rocchini, D., Delucchi, L., Bacaro, G., Cavallini, P., Feilhauer, H., Foody, G. M., He, K. S., Nagendra, H., Porta, C., Ricotta, C., Schmidtlein, S., Spano, L. D., Wegmann, M. and Neteler, M. (2013) 'Calculating landscape diversity with information-theory based indices: A GRASS GIS solution', *Ecological Informatics*. Elsevier B.V., 17, pp. 82–93. doi: 10.1016/j.ecoinf.2012.04.002.

Rodríguez-Galiano, V. F., Abarca-Hernández, F., Ghimire, B., Chica-Olmo, M., Atkinson, P. M. and Jeganathan, C. (2011) 'Incorporating Spatial Variability Measures in Land-cover Classification using Random Forest', *Procedia Environmental Sciences*, 3(0), pp. 44–49. Available at: http://www.sciencedirect.com/science/article/pii/S1878029611000107.

Rodriguez-Galiano, Ghimire, B., Rogan, J., Chica-Olmo, M. and Rigol-Sanchez, J. P. (2012) 'An assessment of the effectiveness of a random forest classifier for land-cover classification', *ISPRS Journal of Photogrammetry and Remote Sensing*. International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS), 67(1), pp. 93–104. doi: 10.1016/j.isprsjprs.2011.11.002.

Rogan, J. (2002) 'A comparison of methods for monitoring multitemporal vegetation change using Thematic Mapper imagery', *Remote Sensing of Environment*, 80(1), pp. 143–156. doi: 10.1016/S0034-4257(01)00296-6.

Rogan, J., Franklin, J., Stow, D., Miller, J., Woodcock, C. and Roberts, D. (2008) 'Mapping land-cover modifications over large areas: A comparison of machine learning algorithms', *Remote Sensing of Environment*, 112(5), pp. 2272–2283. doi: 10.1016/j.rse.2007.10.004.

Schapire, R. E., Freund, Y., Bartlett, P. and Lee, W. S. (1998) 'Boosting the margin: a new explanation for the effectiveness of voting methods.', *The Annals of Statistics*, 26(5), pp. 1651–1686.

Schapire, R. and Freund, Y. (1998) 'Boosting the margin: A new explanation for the effectiveness of voting methods', *The annals of statistics*, 26(5), pp. 1651–1686. doi: 10.1214/aos/1024691352.

Schmidt, G., Jenkerson, C., Masek, J., Vermote, E. and Gao, F. (2013) *Landsat ecosystem disturbance adaptive processing system (LEDAPS) algorithm description*, *Open-File Report*. Reston, VA. Available at: http://pubs.er.usgs.gov/publication/ofr20131057.

scikit-learn developers (2016) *http://scikit-learn.org*. Available at: http://scikit-learn.org.

Selker, H. P., Griffith, J. L., Patil, S., Long, W. J. and D'Agostino, R. B. (1995) 'A comparison of performance of mathematical predictive methods for medical diagnosis: identifying acute cardiac ischemia among emergency department patients', *J Investig Med.*, 43(5), pp. 468–76.

Sesnie, S. E., Gessler, P. E., Finegan, B. and Thessler, S. (2008) 'Integrating Landsat TM and SRTM-DEM derived variables with decision trees for habitat classification and change detection in complex neotropical environments', *Remote Sensing of Environment*, 112(5), pp. 2145–2159. doi: 10.1016/j.rse.2007.08.025.

Shannon, C. E. (1948) 'A Mathematical Theory of Communication', *Bell System Technical Journal*, 27(3), pp. 379–423. Available at: http://www3.alcatel-lucent.com/bstj/vol27-

1948/articles/bstj27-3-379.pdf.

Shao, G., Liu, D. and Zhao, G. (2001) 'Relationships of Image Classification Accuracy and Variation of Landscape Statistics', *Canadian Journal of Remote Sensing*, 27(1), pp. 33–43. doi: 10.1080/07038992.2001.10854917.

Shao, G. and Wu, J. (2008) 'On the accuracy of landscape pattern analysis using remote sensing data', *Landscape Ecology*, 23(5), pp. 505–511. doi: 10.1007/s10980-008-9215-x.

Shen, W., Darrel Jenerette, G., Wu, J. and H. Gardner, R. (2004) 'Evaluating empirical scaling relations of pattern metrics with simulated landscapes', *Ecography*, 27(4), pp. 459–469. doi: 10.1111/j.0906-7590.2004.03799.x.

Sluban, B., Gamberger, D. and Lavrač, N. (2014) 'Ensemble-based noise detection: noise ranking and visual performance evaluation', *Data Mining and Knowledge Discovery*, 28(2), pp. 265–303.

Ståhl, G., Saarela, S., Schnell, S., Holm, S., Breidenbach, J., Healey, S. P., Patterson, P. L., Magnussen, S., Næsset, E., Mcroberts, R. E. and Gregoire, T. G. (2016) 'Use of models in large-area forest surveys : and hybrid estimation', *Forest Ecosystems*. Forest Ecosystems, 3(5), pp. 1–11. doi: 10.1186/s40663-016-0064-9.

Stallman, R. (1985) 'No Title', *The GNUManifesto*. Available at: http://www.gnu.org/gnu/manifesto. html.

Statnikov, A., Wang, L. and Aliferis, C. F. (2008) 'A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification', *BMC Bioinformatics*, 9(1), p. 319. doi: 10.1186/1471-2105-9-319.

Stefanski, J., Mack, B. and Waske, O. (2013) 'Optimization of Object-Based Image Analysis With Random Forests for Land Cover Mapping', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(6), pp. 2492–2504. doi: 10.1109/JSTARS.2013.2253089.

Stehman, S. V (2000) 'Practical Implications of Design-Based Sampling Inference for Thematic Map Accuracy Assessment', *Remote Sensing of Environment*, 72(1), pp. 35–45.

Strobl, C., Boulesteix, A.-L., Zeileis, A. and Hothorn, T. (2007) 'Bias in random forest variable importance measures: illustrations, sources and a solution.', *BMC bioinformatics*, 8, p. 25. doi: 10.1186/1471-2105-8-25.

Stumpf, A. and Kerle, N. (2011) 'Object-oriented mapping of landslides using Random Forests', *Remote Sensing of Environment*, 115(10), pp. 2564–2577. doi: http://doi.org/10.1016/j.rse.2011.05.013.

Sugumaran, R., Hegeman, J., Sardeshmukh, V. and Armstrong, M. (2015) 'Processing Remote-Sensing Data in Cloud Computing Environments', in *Remotely Sensed Data Characterization, Classification, and Accuracies*. CRC Press (Remote Sensing Handbook), pp. 553–562. doi: doi:10.1201/b19294-38.

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P. and Feuston, B. P. (2003) 'Random forest: a classification and regression tool for compound classification and QSAR modeling', *Journal of chemical information and computer sciences*. ACS Publications, 43(6), pp. 1947–1958.

The State of Victoria Department of Environment and Primary Industry (2013) *Victorian Crownland Area Statement 2013*.

The State of Victoria Department of Environment and Primary Industry (2014) *Victoria's State of the Forests Report 2013*. Victoria, Australia.

Tin Kam Ho (1998) 'The random subspace method for constructing decision forests', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8), pp. 832–844. doi: 10.1109/34.709601.

Tomppo, E., Gschwantner, T., Lawrence, M. and McRoberts, R. E. (2010) *National Forest Inventories: Pathways for common reporting*. Berlin: Springer.

Tramontana, G., Ichii, K., Camps-Valls, G., Tomelleri, E. and Papale, D. (2015) 'Uncertainty analysis of gross primary production upscaling using Random Forests, remote sensing and eddy covariance data', *Remote Sensing of Environment*, 168, pp. 360–373. doi: 10.1016/j.rse.2015.07.015.

Tsoumakas, G., Partalas, I. and Vlahavas., I. (2009) 'Applications of Supervised and Unsupervised Ensemble Methods', in *An Ensemble Pruning Primer*. Springer, pp. 1–13.

Tucker, C. J. and Townshend, J. R. (2000) 'Strategies for tropical forest deforestation assessment using satellite data', *International Journal of Remote Sensing*, 21, pp. 1461–1472.

Tumer, K. and Ghosh, J. (1996) 'Error Correlation and Error Reduction in Ensemble Classifiers', *Connection Science*, 8(3–4), pp. 385–404. doi: 10.1080/095400996116839.

Turner, W., Rondinini, C., Pettorelli, N., Mora, B., Leidner, A. K., Szantoi, Z., Buchanan, G., Dech, S., Dwyer, J., Herold, M., Koh, L. P., Leimgruber, P., Taubenboeck, H., Wegmann, M., Wikelski, M. and Woodcock, C. (2015) 'Free and open-access satellite data are key to biodiversity conservation', *Biological Conservation*. Elsevier Ltd, 182, pp. 173–176. doi: 10.1016/j.biocon.2014.11.048.

U.S. Geological Survey (2013) *Earth Explorer*. Available at: http://earthexplorer.usgs.gov.

U.S. Geological Survey (2017) *Landsat project Statistics*. Available at: https://landsat.usgs.gov/landsat-project-statistics.

Uuemaa, E., Antrop, M. and Marja, R. (2009) 'Landscape Metrics and Indices : An Overview of Their Use in Landscape Research Imprint / Terms of Use', *Landscape Research*, 3(1), pp. 1–28.

VanDerWal, J. and Falconi, L. (2014) 'SDMTools: Species Distribution Modelling Tools: Tools for processing data associated with species distribution modelling exercises. R package version 1.1-221'. Available at: http://cran.r-project.org/package=SDMTools.

Vermont Department of Forests Parks and Recreation (2015) *2015 Vermont Forest Fragmentation Report*. Available at: http://vnrc.org/wp-content/uploads/2015/04/FOREST-FRAGMENTATION_FINAL-11.pdf.

Viridans (2000) *Ecosystems and Vegetation*. Available at: http://www.viridans.com/ECOVEG/.

Vogelmann, J. E., Howard, S. M., Yang, L. M., Larson, C. R., Wylie, B. K. and Van Driel, N. (2004) 'Completion of the 1990s National Land Cover Data set for the conterminous United States from Landsat Thematic Mapper data and ancillary data sources', *Photogrammetric Engineering & Remote Sensing*, 67, pp. 650–662.

Weiss, G. M. and Provost, F. (2003) 'Learning When Training Data are Costly : The Effect of Class Distribution on Tree Induction', *Journal of Artificial Intelligence Research*, 19, pp. 315–354.

Wickham, J. D., Neill, R. V. O., Riitters, K. H., Wade, T. G. and Jones, K. B. (1997) 'Sensitivity of Selected Landscape Pattern Metrics to Land-Cover Misclassificat ion and Differences in Land-Cover Composition', *Photogrammetry and Remote Sensing*, 63(4), pp. 397–402.

Wilkes, P., Jones, S., Suarez, L. and Mellor, A. (2015) 'Mapping Forest Canopy Height Across Large Areas by Upscaling ALS Estimates with Freely Available Satellite Data', *Remote Sensing*, 7, pp. 1–25. doi: 10.3390/rs70x000x.

Wilkinson, G. G. (2005) 'Results and implications of a study of fifteen years of satellite image classification experiments', *IEEE Transactions on Geoscience and Remote Sensing*, 43(3), pp. 433–440. doi: 10.1109/TGRS.2004.837325.

Woodgate, P. and Black, P. (1988) *Forest cover changes in Victoria 1869-1987*. Melbourne: Remote Sensing Group, Lands and Forests Division, Dept. of Conservation, Forests and Lands.

Wu, J., Gao, W. and Tueller, P. T. (1997) 'Effects of Changing Spatial Scale on the Results of Statistical Analysis with Landscape Data: A Case Study', *Annals of GIS*, 3(1–2), pp. 30–41. doi: 10.1080/10824009709480491.

Wulder, M. (1998) 'Optical remote-sensing techniques for the assessment of forest inventory and biophysical parameters', *Progress in physical Geography*, 22(4), p. 449.

Wulder, M. A., Masek, J. G., Cohen, W. B., Loveland, T. R. and Woodcock, C. E. (2012) 'Opening the archive : How free data has enabled the science and monitoring promise of Landsat', *Remote Sensing of Environment*. Elsevier B.V., 122, pp. 2–10. doi: 10.1016/j.rse.2012.01.010.

Wulder, M. A., White, J. C., Goward, S. N., Masek, J. G., Irons, J. R., Herold, M., Cohen, W. B., Loveland, T. R. and Woodcock, C. E. (2008) 'Landsat continuity: Issues and opportunities for land cover monitoring', *Remote Sensing of Environment*, 112, pp. 955–969. doi: 10.1016/j.rse.2007.07.004.

Wulder, M. a, White, J. C., Gillis, M. D., Walsworth, N., Hansen, M. C. and Potapov, P. (2010) 'Multiscale satellite and spatial information and analysis framework in support of a large-area forest monitoring and inventory update.', *Environmental monitoring and assessment*, 170(1–4), pp. 417–33. doi: 10.1007/s10661-009-1243-8.

Yu, X., Wu, X., Luo, C. and Ren, P. (2017) 'Deep learning in remote sensing scene classification: a data augmentation enhanced convolutional neural network framework', *GIScience & Remote Sensing*. Taylor & Francis, 54(5), pp. 741–758. doi: 10.1080/15481603.2017.1323377.

Yuan, H., Van Der Wiele, C. F. and Khorram, S. (2009) 'An Automated Artificial Neural Network System for Land Use/Land Cover Classification from Landsat TM Imagery', *Remote Sensing*, 1(3), pp. 243–265. doi: 10.3390/rs1030243.