ORIGINAL RESEARCH

Impact of ecological redundancy on the performance of machine learning classifiers in vegetation mapping

Paul D. Macintyre[1] (iD) | Adriaan Van Niekerk[2] (iD) | Mark P. Dobrowolski[1,3] (iD) |
James L. Tsakalos[1] (iD) | Ladislav Mucina[1,2] (iD)

[1]School of Biological Sciences, The University of Western Australia, Perth, Crawley, WA, Australia

[2]Centre for Geographical Analysis, Stellenbosch University, Matieland, Stellenbosch, South Africa

[3]Iluka Resources Limited, Perth, WA, Australia

Correspondence
Paul D. Macintyre, School of Biological Sciences, The University of Western Australia, Perth, Crawley, WA, Australia.
Email: paul.macintyre@research.uwa.edu.au

Funding information
Australian Research Council, Grant/Award Number: LP150100339

Abstract

Vegetation maps are models of the real vegetation patterns and are considered important tools in conservation and management planning. Maps created through traditional methods can be expensive and time-consuming, thus, new more efficient approaches are needed. The prediction of vegetation patterns using machine learning shows promise, but many factors may impact on its performance. One important factor is the nature of the vegetation–environment relationship assessed and ecological redundancy. We used two datasets with known ecological redundancy levels (strength of the vegetation–environment relationship) to evaluate the performance of four machine learning (ML) classifiers (classification trees, random forests, support vector machines, and nearest neighbor). These models used climatic and soil variables as environmental predictors with pretreatment of the datasets (principal component analysis and feature selection) and involved three spatial scales. We show that the ML classifiers produced more reliable results in regions where the vegetation–environment relationship is stronger as opposed to regions characterized by redundant vegetation patterns. The pretreatment of datasets and reduction in prediction scale had a substantial influence on the predictive performance of the classifiers. The use of ML classifiers to create potential vegetation maps shows promise as a more efficient way of vegetation modeling. The difference in performance between areas with poorly versus well-structured vegetation–environment relationships shows that some level of understanding of the ecology of the target region is required prior to their application. Even in areas with poorly structured vegetation–environment relationships, it is possible to improve classifier performance by either pretreating the dataset or reducing the spatial scale of the predictions.

KEYWORDS
functional redundancy, machine learning, predictive modeling, predictive vegetation mapping, vegetation patterns, vegetation–environment relationship

1 | INTRODUCTION

Vegetation maps are simplified models of vegetation complexity carrying important messages about the position of vegetation types along environmental gradients. The utility of such maps extends beyond simple descriptions and audits of vegetation patterns (vegetation types and their complexes) within an area. For instance, they are indispensable tools in land-use and biodiversity conservation planning (Akasaka et al., 2014; Ferrier, 2002; Franklin, Woodcock, & Warbington, 2000) and serve as a major source of predictive modeling in global-change research. Vegetation maps traditionally relied on extensive field surveys (e.g. Beard, 1975; Küchler & Zonneveld, 1988), yet these can be prohibitively costly and time-consuming, especially when the area to be mapped is large and complex (Lee & Lunetta, 1996). Recent technological advances and remotely sensed data collection have changed the way in which vegetation maps are made and enhanced the definition of boundaries between mapped vegetation units at all spatial scales. New technologies also facilitate the production of large and complex spatial (geographical and biological) datasets that can support vegetation mapping (e.g. Farr et al., 2007; Hijmans, Cameron, Parra, Jones, & Jarvis, 2005; Viscarra Rossel et al., 2015). Modern vegetation science is also experiencing a boost through the implementation of novel data-analytical approaches, enhancing our understanding of how the vegetation patterns formed and which environmental (or man-induced) drivers might underpin these patterns (Blois et al., 2013; Lippok et al., 2014; Reynolds, Packer, Bever, & Clay, 2003). The combination of technological advances and improved understanding allowed the development of models to reconstruct past vegetation patterns or predict potential vegetation patterns within a region. Such models can substantially reduce the time and cost of constructing vegetation maps.

Machine learning (ML) algorithms have been shown to produce models that are accurate and robust (de Souza, Boerder, Matwin, & Worm, 2016; Dickson & Perry, 2016; Osis, Hettinga, & Ferber, 2016). In principle, a ML algorithm builds a solution (model) by examining a sample dataset and identifying features or trends. The model is then applied to an unexamined dataset to make predictions. While ML has been applied to examine a diverse range of problems (e.g. Pasolli, Truong, Malik, Waldron, & Segata, 2016; Shipp et al., 2002; Tango & Botta, 2013), there has been a recent increase in its use within the geospatial and ecological sciences. For instance, ML has been successfully applied to predict species distribution (Liu, White, Newell, & Griffioen, 2013), land-use change (Tayyebi & Pijanowski, 2014), and hydrological regimes (Cross et al., 2015) and has also been applied to vegetation mapping across a range of spatial scales using a variety of algorithms (e.g. Bradter, Thom, Altringham, Kunin, & Benton, 2011; Munyati, Ratshibvumo, & Ogola, 2013; Pesch, Schmidt, Schroeder, & Weustermann, 2011; Zhang & Xie, 2013). When applied to vegetation mapping, ML algorithms (hereafter referred to as ML classifiers) aim to create models that depict the relationships between the vegetation types identified within an area and environmental (e.g. climate, geology) or spectral reflectance variables. Although useful maps have been produced using these methods, the performance

(measured in terms of output accuracy) varies with quality of the data and the scale of the classification. The impact of data quality and scale on accuracy has received much attention (Beekhuizen et al., 2014; Ghosh, Fassnacht, Joshi, & Koch, 2014), but the influence of ecological factors, specifically redundancy (see below), on the performance of ML classifiers has to our knowledge not been investigated to date. Given that vegetation classification success is essentially a function of the vegetation–environment relationship, it is critical that we gain a better understanding of how this relationship impacts on the classification results.

The vegetation patterns and their dynamics are a result of interaction of plant species forming the vegetation cover with their environment (Götzenberger et al., 2012). The presence of each species in these complex structures is determined by their life-history: a sum of functional traits that mediate the species response and the abiotic and biotic (interactions with other species) environment (Zobel, 1992). Niche theory predicts that each species would have a unique position along environmental gradients. However, it fails to account that some vegetation types have many species that fulfill the same (or very similar) functional role within the plant communities (Kang et al., 2015) hence the functional role of those species in a plant community is considered redundant (Walker, 1992). This means that floristically distinct communities may occur in similar positions along ecological gradients in the landscape. In such highly redundant systems, many distinct plant communities can be recognized but the relationship of these communities to their environment is unclear. In contrast, low-redundancy systems (where few, if any, plants share responses to the factors of the ecological space) have a stronger ecological link with the environment.

In this study, we investigated the impact of the vegetation–environment relationship (considered as a surrogate for redundancy) on the performance of four ML classifiers. The performance of the ML classification models was tested by predicting (modeling) vegetation in two contrasting vegetation landscapes of Western Australia, namely (a) eucalypt-dominated tropical savannah woodlands (characterized by low redundancy) of the Northern Kimberley and (b) temperate kwongan scrub (characterized by high redundancy) of the Geraldton Sandplains. We also tested whether data pretreatment through feature extraction or feature selection would have any impact on the model performance.

2 | METHODS

2.1 | Datasets

Two datasets representing the vegetation of contrasting regions in Western Australia (Mitchell Plateau and Geraldton Sandplains) were compiled and used as test cases. Each area was analyzed individually. Mitchell Plateau is part of the Northern Kimberley region (approx. 14°44′S, 125°53′E) and covers approximately 480,000 ha. The size of the Mitchell Plateau, combined with the remoteness of the region and poor road infrastructure, prevents traditional field-focused mapping methods from being efficient. A total of 148 vegetation

plots (50 m × 50 m), listing all species and estimating their percentage cover, were sampled in this region. The plots were classified into plant communities by applying the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) clustering on log-transformed data and similarity ratio as resemblance (see Mucina, Stephenson, Daniel, Van Niekerk, & Boonzaaier, 2013 for details). This classification yielded 20 floristically defined plant communities. Three communities were identified as being azonal and were consequently excluded from the modeling process. This dataset is referred to as Woodland.

Geraldton Sandplains are located 270 km north of Perth (approx. 29°49′S, 115°16′E), and the region covers approximately 121,000 ha. The sandplains are part of one of the most diverse floristic regions of Western Australia, with more species diversity of sclerophyllous shrubs than anywhere else in the state (Lamont, Hopkins, & Hnatiuk, 1982). A total of 542 vegetation plots (10 m × 10 m) were sampled in this area (Woodman Environmental Consulting, 2009). This dataset, from now on referred to as the Kwongan dataset, was classified using Beta Flexible Clustering (beta = −0.25) based on Bray–Curtis dissimilarity, with no prior data transformation applied. This classification (Tsakalos, Dršková, Hruban, Mucina, & Dobrowolski, 2014; J. Tsakalos, unpublished data) identified 24 distinct plant communities.

For both regions, rare vegetation classes were merged with their most similar class(es) at the plant community classification (called Level A) scale to ensure a minimum representation of five samples. This resulted in a reduction of the number of classes in the Woodland dataset from 17 to nine, and in the Kwongan dataset from 24 to 14. The classifier comparisons were also conducted at two broader classification scales created by grouping the Level A classes to form a group of plant communities called Level B, and then further by grouping the Level B classes into a high-order group called Level C, to examine whether redundancy changes with classification scale. At the Level B scale, the Woodland dataset contained four classes, while the Kwongan dataset had six classes. Both datasets contained two classes at the Level C scale.

2.2 | Predictor variables

The predictors used as input to the ML classifiers (Supporting Information Table S1) were similar for the two regions and consisted of 67 climatic variables from the BioClim database (Hijmans et al., 2005) and 12 terrain-based layers created using the SAGA-GIS v2.1.2 basic terrain analysis tool. The terrain variables were derived from the 1-ARC second Shuttle Radar Topography Mission (SRTM) digital elevation model (DEM) (Farr et al., 2007). In Kwongan vegetation, 23 in situ sampled soil variables were considered. Because no soil samples were taken in the Woodland vegetation, 10 variables from the 1:150,000 (90 m resolution) National Soil Grid of Australia (Viscarra Rossel et al., 2015) were used instead. None of the vegetation plots occurred within the same cell.

2.3 | Data treatment

Three different predicting datasets were assessed in this study: (a) the full dataset (FD) containing all predicting variables for each region; (b) a dataset containing variables derived from a feature extraction (FE) process, and (c) a dataset containing variables from a feature selection (FS) process. FE involves the construction of a new (smaller) feature set derived from the full dataset (Hira & Gillies, 2015). In this study, this dataset was constructed using the principle component analysis (PCA) tool within ArcMap v10.3. The first five principal components (PCs), which explained more than 95% of the variation in the data, were retained at all scales for both regions. FS is the process of discarding unimportant variables. The FS dataset was created using the random forest (RF) tool within Salford Predictive Modeller v8.0. A proportion (40%) of the sample data was used for (out-of-bag) accuracy assessment. The process started with the full set of variables, after which the importance of each variable was assessed. Subsequent models were then created by excluding less important variables. This process continued until the accuracy of the model could not be improved.

2.4 | Quantification of ecological redundancy

Ecological redundancy depends on the strength of the vegetation–environment relationship. It can be approximated by examining the environmental variables and the quantification of their relationships using Canonical Correlation Analysis (CCA), followed by a Monte Carlo permutation test as implemented in CANOCO v.4.5 (Lepš & Šmilauer, 2003). CCA is a form of constrained ordination involving two matrices: one describing the species co-occurrence in plots, and one comprising environmental variables for the same plots. The latter matrix constrains the ordination of plots sharing species. A Monte Carlo permutation test examines the null hypothesis that the species composition (of communities) is independent of the environmental factors. During the permutations (9,999 runs), the environmental variables were randomly assigned among plots, and then new CCA analyses were performed and assessed whether random environmental data might produce equal or better ordination structure than the ordination of the real data. An F-value was generated after all permutations. If the F-value exceeded 0.05, the null hypothesis was accepted.

2.5 | Machine learning classifiers

Four ML classifiers, namely classification tree (CT), RF, support vector machine (SVM), and nearest neighbor (NN), were tested in this study.

CT was selected owing to its nonparametric nature, ease of interpretability, ability to handle multiple data types, and speed of prediction. CT proceeds through a process of recursive partitioning, which splits the training data into a series of nodes based on increases in homogeneity of the subsequent groups (Breiman, Friedman, Stone, & Olshen, 1984). The result of this process is a tree composed of nodes and terminal leaves that can easily be converted into a simple set of if-then rules. CT outputs are easy to interpret because the resulting tree can be scrutinized to understand why a given output was generated (Chen, Wang, & Zhang, 2011). However, these trees

—WILEY┘

can suffer from overfitting if allowed to grow fully without pruning (Schaffer, 1993).

Random forest is an ensemble CT classifier (Breiman, 2001; Chen et al., 2011). The principle of ensemble classifiers is that a large collection of weaker classifiers (individual CTs in this case) can be used to create a strong classifier. RF involves the construction of large number of individual trees from the training data (Rodriguez-Galiano, Ghimire, Rogan, Chica-Olmo, & Rigol-Sanchez, 2012). How the trees are constructed differs from CT in that a random selection of training data is used for each tree so that each tree is trained on a different set of data. Unlike CT which considers all predictor variables at each node, RF selects a random subset of predictors and these are used to identify the best split pattern. A collection of these trees is the forest where each tree is a unique classification in terms of a random selection of predicting variables and the resulting splitting pattern leading to definition of classes. Once all trees have been constructed, the objects of the studied dataset are presented to each tree in the forest, which then predicts the class to which the object would belong to. The class that was predicted the most frequently is assigned to the unclassified data through a measure of a majority vote. The trees in RF are not pruned as the collection of all trees reduces the likelihood of overfitting. Because of its stochastic nature, RF is also relatively insensitive to noise and outliers (Breiman, 2001). However, the classification accuracy of RF is at a trade-off with interpretability (Zhang & Wang, 2009).

SVM classifiers are widely used in land cover classification studies (Duro, Franklin, & Dubé, 2012; Zhang & Xie, 2013). SVM operates by identifying a hyperplane that separates the samples of two classes in a variable-defined space. Finding the optimal hyperplane is challenging, because there are numerous planes that could separate the training classes (Cracknell & Reading, 2014). SVM addresses this by identifying training samples (support vectors) at the transition between two classes and identifies a plane that aligns with them. The optimal hyperplane is then identified equidistant between these support vectors (Pal & Foody, 2010). Kernels are often used to transform the feature space to improve the fitting of hyperplanes. SVMs have shown capacity for producing reliable classifications even when relatively few training samples are used (Mountrakis, Im, & Ogole, 2011). SVMs are also relatively insensitive to the effects of high dimensionality, which is beneficial when dealing with a large set of predictors (Gualtieri, 2009).

NN sorts training samples according to their similarity (distance in a feature space) to each other (Bhatia, 2010). The distances between the training data can be determined using some metrics, Euclidean distance being the most commonly applied. When data that have not been classified are presented to the classifier, the distance of an unknown sample is calculated to each of the neighboring training samples and the label (class) of the closest training sample is assigned to the unknown data. Unlike many other classifiers, NN retains all training samples during the classification process which can reduce efficiency when the size of the

training set is large or when a high number of predictor variables is considered.

2.6 | Comparison of ML classifiers

The performance of the selected ML classifiers was tested using custom software created by Myburgh and Van Niekerk (2013) based on the GDAL (GDAL Development Team, 2010), OpenCV (Bradski, 2000), and LibSVM (Chang & Lin, 2011) libraries. This software uses a 60/40 split in the observation data to create training and validation datasets, respectively, requiring a minimum of five samples per class (three for training and two for validation). Confusion matrices, which use the validation dataset to determine which samples had their class correctly or incorrectly predicted, were created for each model yielding the overall accuracy (OA) and Kappa (K) values. All datasets were iterated 30 times to reduce the influence of the random selection of training data.

2.7 | Statistical analyses

A range of nonparametric statistical tests are recommended for comparing the performance of ML classifiers (see e.g. Garcia & Herrera, 2008). In this study, we chose the aligned rank transform (ART; Wobbrock, Findlater, Gergle, & Higgins, 2011) to perform a nonparametric factorial analysis (using ANOVA procedures) and multiple comparisons when significance is detected. This analysis was performed using ARTool (Kay & Wobbrock, 2016). The testInteractions function, which is part of the Phia module, was used for carrying out the comparisons (Rosario-Martinez, 2015). The Holm method for p-value adjustment was used as recommended.

3 | RESULTS

The results are summarized by classifier and scale of analysis (Levels) in Figure 1. Examples of the differences in predicted patterns at the finest scale (Level A) for each region are presented in Supporting Information Figures S1 and S2. A summary table of mean accuracy is shown in Supporting Information Table S3, while matrices showing the results of all pairwise comparisons for each region/scale are presented in Supporting Information Appendix S1.

3.1 | Ecological redundancy patterns in the contrasting datasets

On purely statistical grounds, CCA analysis (and associated Monte Carlo permutations) showed that the vegetation patterning of the Woodland vegetation are well explained by the selected environmental variables (F-value = 0.0195), while with the Kwongan vegetation, this was not the case (F-value = 0.0520).
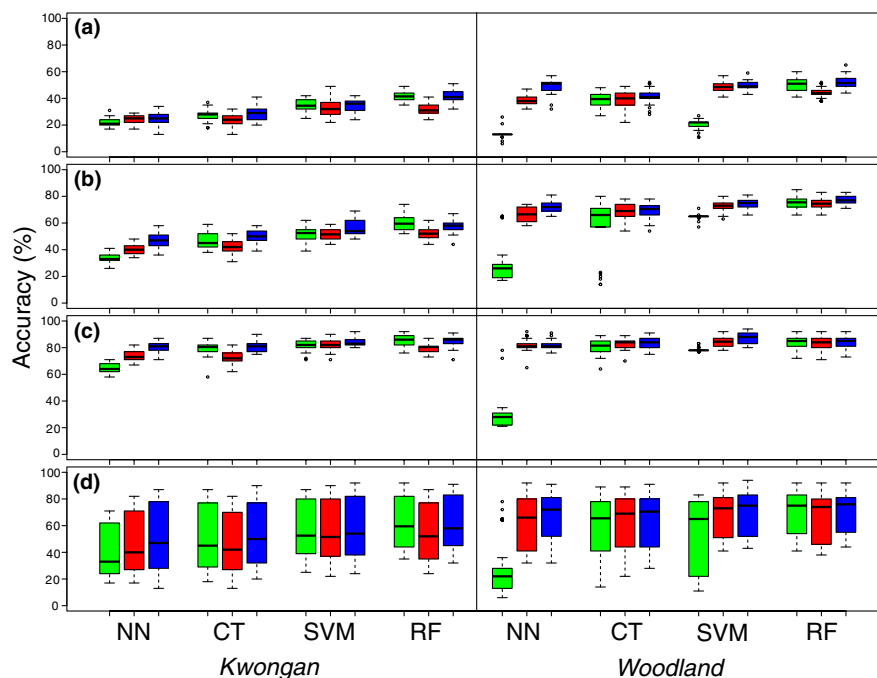
FIGURE 1 Boxplots showing classifier response to data treatment in Kwongan and Woodland at Levels (a, b, and c). The plots in section (d) represent the overall accuracy. Key: Green = FD dataset; Red = FE dataset; Blue = FS dataset; NN = nearest neighbor; CT = classification tree; SVM = support vector machine; RF = random forests

In terms of ecological redundancy, we suggest that the detected vegetation–environment relationship in the Woodland is a sign of low (if any) ecological redundancy, while Kwongan is ecologically redundant.

3.2 | Feature selection

While not all models showed improvements in accuracy (see Supporting Information Table S2), it was possible to reduce the number of predictors in all cases. This dataset contained a reduced number of 15, eight, and seven predictors for Levels A, B, and C, respectively, in Kwongan and 16, 13, and 10 for Levels A, B, and C, respectively, in Woodland. The FS dataset for Kwongan mainly comprised soil and topographic variables, while the Woodland FS dataset included some climatic, in addition to soil and topographic, variables.

3.3 | Classifier performance

The results show that each factor (scale, region, treatment, and classifier) had a significant effect on prediction accuracy with the interactions between these factors also showing significance. The two vegetation datasets used are significantly different overall (and at each scale), with predictions in the Woodland vegetation deemed more accurate. The results, at each scale of analysis, are consequently outlined separately for each region below. Unless otherwise specified, all significant differences were found at the level of $p < 0.001$.

3.4 | Kwongan dataset

3.4.1 | Level A

When the full dataset was used as input to the classifiers at Level A in the Kwongan, the accuracies of all classifiers were significantly different from each other, with RF (41%) and NN (22%) providing the best and worst accuracies, respectively. With the FS dataset, RF (42%) also outperformed the other classifiers with NN (25%) again producing the weakest models. With the FE dataset, two groups of classifiers were found, with SVM (35%) and RF (32%) returning significantly higher accuracies than CT (23%) and NN (24%). Response to pretreatment varied from making no significant difference—when either FS or FE was applied to SVM (35% each time) and NN (22%–25%), and when FS was applied to RF (41% vs. 42%) and CT (27% vs. 29%)—to significantly reducing classifier performance—when FE was applied to RF (32% vs. 41%) and CT (23% vs. 27%). Overall, the RF classifier, combined with the full and FS datasets, provided the best performance (41%, 42%) at Level A in the Kwongan dataset.

3.4.2 | Level B

The experiments carried out at Level B in the Kwongan showed that all classifiers produced significantly different accuracies when the full dataset was used as input. As at Level A, RF generated the strongest models (60%), while the NN classifier performed relatively poorly (34%). With both the FE and FS datasets, the classifiers formed two distinct groups with RF (FE: 52%, FS: 58%) and SVM (FE: 51%, FS:

56%) forming a group with higher accuracy than CT (FE: 42%, FS: 50%) and NN (FE: 40%, FS: 47%). Pretreatment did not make a difference to the performance of the CT or SVM classifiers. The accuracies of the NN classifications were significantly different among datasets, with FE and FS significantly improving (by 6% and 13%, respectively) performance. Overall, three classifier dataset combinations provided the best predictive accuracy, namely RF using the FD (60%) or FS (58%) datasets, and SVM using the FS dataset (56%). Compared to Level A, all classifiers performed better at Level B.

3.4.3 | Level C

Using the full dataset as input at Level C, NN produced significantly lower accuracies (15%–20%) compared to the other classifiers, while the differences between SVM, CT, and RF were marginal. FS had no significant impact on RF and CT, while FE reduced the accuracies of these classifiers substantially (7% and 6%, respectively). NN benefitted (FE: 9%, FS: 15%) the most from the pretreatments, but had no impact on SVM performance. Overall, the highest accuracies were achieved when the FS dataset was used as input to RF (85%) and SVM (84%).

3.5 | Woodland dataset

3.5.1 | Level A

Using the full dataset as the input at Level A, RF returned the highest accuracy (50%) with SVM (20%) and NN (13%) producing the weakest models. When the FE dataset was considered, two groups with notably different accuracy statistics were noted. RF (45%) and SVM (48%) formed a group with significantly higher accuracies than CT (39%) and NN (39%), while no significant differences among classifiers within each group were noted. With the FS dataset, it was found that CT returned the lowest accuracy (42%), while the results of RF (52%), SVM (50%), and NN (49%) were on par with one another.

The treatments resulted in significant improvements to the accuracies of SVM (FE: 48%, FS: 50%) and NN (FE: 39%, FS: 49%), but had no effect on CT. The FS dataset had no effect on the performance of RF, but FE significantly reduced its accuracy (by 5%). Overall, the full dataset using the RF classifier (50%)—or the FS dataset using RF (52%), SVM (50%) or NN (49%)—performed the best at Level A in the Woodland.

3.5.2 | Level B

Using the full Woodland dataset as input at Level B, all classifiers returned statistically different accuracies, with RF (76%) performing the best and NN (31%) generating the weakest models. Pretreating the full dataset made significant improvements to the performance of CT (FE: 69%, FS: 69%), SVM (FE: 73%, FS: 75%), and NN (FE: 66%, FS: 72%), but had no effect on RF. Overall, the RF classifier consistently (using any of the datasets) outperformed

the other classifiers (74%–77%), while the SVM classifier produced similar results when the FE (73%) or FS (75%) datasets were used as input.

3.5.3 | Level C

At Level C, RF (85%) and CT (81%) outperformed the other classifiers when the full dataset was used as input. NN (30%) returned the lowest accuracy. Pretreating the dataset had no effect on the accuracy of CT and RF, but significantly improved the performance of both SVM (FE: 84%, FS: 87%) and NN (FE: 82%, FS: 82%). Overall, the highest performance was obtained using the FE or FS datasets with either SVM (FE: 84%, FS: 87%) or RF (FE: 83%, FS: 84%).

In general, the Kappa values resulting from the experiments agree with the overall accuracies. However, some interesting differences were noted. For example, when the full dataset was used as input in Woodland, the Kappa values of SVM (0.02–0.08) and NN (0.005–0.03) were very low, which indicates that the models produced accuracies that are similar to what one would expect from random class assignment. While accuracy increased at Levels B and C, these classifiers were unable to return Kappa values of above 0.08 (on average) when the full dataset was used as input. However, the application of FE and FS resulted in substantial increases in reliability. For example, at Level A, the average Kappa for the SVM classifier increased by 0.34 (FE) and 0.23 (FS), while for NN, it increased by 0.26 (FE) and 0.21 (FS). These increases became more pronounced as the scale was reduced. For instance, in the case of SVM, accuracies increased by 45% from Levels A to C and 13% from Level B to C.

4 | DISCUSSION

The results of this study show that the vegetation of the Kwongan is ecologically redundant. The hypothesis that such redundancy reduces the power of ML classifiers in predicting vegetation patterns is supported by the observations that the accuracies of the Kwongan classifications—especially at Levels A and B—were consistently lower compared to those of Woodland. It is well known that classifiers behave differently under conditions of data redundancy, however, given that the same number of features was used as input and given that the same dimensionality reduction methods were employed in both study areas, the differences in accuracy are most likely due to inherent redundancies in the environmental variables used. The significant difference in classifier performance between the regions shows that—based on the ecological relationships of the regions—we can predict the vegetation patterns of the nonredundant Woodland vegetation with greater confidence than those of the Kwongan. Although other factors may also have contributed to the differences in accuracy—for example, the classification schemes (the woodland vegetation types are better defined) and the differences in data sources (ground collected soil vs. the interpolated soil data)—the results offer evidence that ecological redundancy must

have been a major driver. While different resemblance and clustering methods can change the final classification, the resulting regional systems used in this study were selected through a comparison and validation process (e.g. Tichý, Chytrý, Hájek, Talbot, & Botta-Dukát, 2010). While different combinations were used in each region, they represent the best systems for each type.

The finding that higher accuracies were obtained when the complexity was reduced (by carrying out the experiments at coarser scales) was expected, but the way in which the classifiers responded to the FE and FS data treatments provided new insights into the impact of ecological redundancy on ML classifications. Both FE and FS have been found in many studies (Babaoğlu, Fındık, & Bayrak, 2010; Chandrashekar & Sahin, 2014; Howley, Madden, O'Connell, & Ryder, 2006) to improve classification accuracy, and it was therefore expected that predictions following these methods would be more accurate than predictions made on the full dataset. However, in the ecologically redundant Kwongan, these methods led to statistically significant classification improvements in only three scenarios, whereas in the nonredundant Woodland, they improved accuracies 12 times. Furthermore, the use of FE in the Kwongan was found to significantly reduce the performance of the classifiers four times. These findings can be explained within the context of the relationship each vegetation type has with the environment. The vegetation–environment relationship has little structure where redundancy is present and data treatments—designed to capture the variability (FE) of the environment or reduce uninformative variables (FS)—thus have little effect. An unexpected result was that the tree-based classifiers (CT and RF) showed little response to both FE and FS in both regions. For FE, this is likely a result of the way that these classifiers create their predictions, with the recursive partitioning possibly not as strong using the transformed data.

RF unresponsiveness to FS is attributed to its use of a randomly selected subset of features during model building. Feature selection is consequently inherent in the algorithm. Although some studies (e.g. Gilbertson & Van Niekerk, 2017) have shown that RF classification accuracies can be improved using FS pretreatments, it was not the case with our data. For both regions and all classifiers, where a difference between the treatments was found in all cases FS was found to be significantly more accurate. Difference in response between the two treatments was more common in the Kwongan than in Woodland, suggesting that in conditions of redundancy, FS is more beneficial (or at least not detrimental) to classification accuracy compared to FE. We suggest that this is due to the creation of the PCs in the FE process. In this process, the input features (environmental variables) which show the greatest variation are considered (reasonably) to be more important in the construction of the new feature set (Shlens, 2014). It is important that, this process is conducted independently of the prediction targets (vegetation classes). This is, in contrast to the FS process in which the feature set, examined against our classes to determine which are the most informative in separating the classes. Given the ecological redundancy of the Kwongan, it is possible that the FE process downplayed the importance of those features which better separate the classes.

In each region, the RF and SVM classifiers consistently outperformed the other classifiers. This is consistent with the results of other studies (Duro et al., 2012; Pal, 2005; Roli & Fumera, 2001), where it was found that RF and SVM were more successful, especially under complex conditions. One can argue that, in this study, ecological redundancy contributed to complexity and that it offers an explanation why these classifiers performed better than NN and DT.

The use of ML classifiers to create robust maps of vegetation patterns is becoming more popular. However, the influence of ecological features on these classifiers is still poorly understood. The results of this study show that, although ecological complexity (i.e. redundancy) has a negative effect on classifier performance, it is not the only factor that contributed to overall performance of classifiers. The unexpectedly poor performance of NN and SVM in the Woodland suggests that the number and quality of training samples may have affected (through the addition of noise) the classifications. The finding that both FE and FS improved the performance of these classifiers in this region provides some support for this notion.

This study focussed on the application of ML to produce potential vegetation maps and as such purposefully omitted additional sources of information (such as satellite reflectance data) as this would have restricted the predictions to extant vegetation. However, it would be beneficial to examine what effect the inclusion of these datasets (either in addition to the environmental data or as sole predictors) may have on improving modeling accuracies. It is clear that more work is needed to find cost-efficient and accurate methods for generating vegetation maps over large and complex areas.

ACKNOWLEDGMENTS

CONFLICT OF INTEREST

None declared.

AUTHORS' CONTRIBUTION

AV, LM, and PM conceived and designed the study. MD managed the collection of field environmental data. PM collated the spatial (GIS) data and performed all GIS-related and statistical analyses. PM and LM led the writing of the manuscript, with AV providing further guidance. JT provided an unpublished vegetation classification scheme. All authors contributed critically to the drafts and gave final approval for publication.

DATA ACCESSIBILITY

The data for this paper are lodged within Dryad. https://doi.org/10.5061/dryad.1m8tg17

ORCID

Paul D. Macintyre http://orcid.org/0000-0003-1792-5774

Adriaan Van Niekerk http://orcid.org/0000-0002-5631-0206

Mark P. Dobrowolski http://orcid.org/0000-0001-5586-4023

James L. Tsakalos http://orcid.org/0000-0001-5067-196X

Ladislav Mucina http://orcid.org/0000-0003-0317-8886

REFERENCES

Akasaka, M., Takenaka, A., Ishihama, F., Kadoya, T., Ogawa, M., Osawa, T., … Yamano, H. (2014). Development of a national land-use/cover dataset to estimate biodiversity and ecosystem services. In S. Nakano, T. Yahara & T. Nakashizuka (Eds.), *Integrative observations and assessments* (pp. 209–229). Dordrecht, NL: Springer.

Babaoğlu, I., Fındık, O., & Bayrak, M. (2010). Effects of principle component analysis on assessment of coronary artery diseases using support vector machine. *Expert Systems with Applications*, *37*, 2182–2185. https://doi.org/10.1016/j.eswa.2009.07.055

Beard, J. S. (1975). The vegetation survey of Western Australia. *Vegetatio*, *30*, 179–187. https://doi.org/10.1007/BF02389706

Beekhuizen, J., Heuvelink, G. B., Huss, A., Bürgi, A., Kromhout, H., & Vermeulen, R. (2014). Impact of input data uncertainty on environmental exposure assessment models: A case study for electromagnetic field modelling from mobile phone base stations. *Environmental Research*, *135*, 148–155. https://doi.org/10.1016/j.envres.2014.05.038

Bhatia, N. (2010). Survey of nearest neighbor techniques. *International Journal of Computer Science and Information Security*, *8*, 302–305.

Blois, J. L., Williams, J. W., Fitzpatrick, M. C., Ferrier, S., Veloz, S. D., He, F., … Otto-Bliesner, B. (2013). Modeling the climatic drivers of spatial patterns in vegetation composition since the Last Glacial Maximum. *Ecography*, *36*, 460–473. https://doi.org/10.1111/j.1600-0587.2012.07852.x

Bradski, G. (2000). The OpenCV library. *Dr. Dobb's Journal of Software Tools*, *120*, 122–125.

Bradter, U., Thom, T. J., Altringham, J. D., Kunin, W. E., & Benton, T. G. (2011). Prediction of national vegetation classification communities in the British uplands using environmental data at multiple spatial scales, aerial images and the classifier random forest. *Journal of Applied Ecology*, *48*, 1057–1065. https://doi.org/10.1111/j.1365-2664.2011.02010.x

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth Books.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers and Electrical Engineering*, *40*, 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024

Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*, 1–27. https://doi.org/10.1145/1961189.1961199

Chen, X., Wang, M., & Zhang, H. (2011). The use of classification trees for bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*, 55–63.

Cracknell, M. J., & Reading, A. M. (2014). Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. *Computers and Geosciences*, *63*, 22–33. https://doi.org/10.1016/j.cageo.2013.10.008

Cross, A. T., Turner, S. R., Merritt, D. J., Van Niekerk, A., Renton, M., Dixon, K. W., & Mucina, L. (2015). Vegetation patterns and hydro-geological drivers of freshwater rock pool communities in the monsoon-tropical Kimberley region, Western Australia. *Journal of Vegetation Science*, *26*, 1184–1197. https://doi.org/10.1111/jvs.12318

de Souza, E. N., Boerder, K., Matwin, S., & Worm, B. (2016). Improving fishing pattern detection from satellite AIS using data mining and machine learning. *PLoS ONE*, *11*, e0158248. https://doi.org/10.1371/journal.pone.0158248

Dickson, M. E., & Perry, G. L. (2016). Identifying the controls on coastal cliff landslides using machine-learning approaches. *Environmental Modelling and Software*, *76*, 117–127. https://doi.org/10.1016/j.envsoft.2015.10.029

Duro, D. C., Franklin, S. E., & Dubé, M. G. (2012). A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery. *Remote Sensing of Environment*, *118*, 259–272. https://doi.org/10.1016/j.rse.2011.11.020

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., … Alsdorf, D. E. (2007). The shuttle radar topography mission. *Reviews of Geophysics*, *45*, 1–33.

Ferrier, S. (2002). Mapping spatial pattern in biodiversity for regional conservation planning: Where to from Here? *Systematic Biology*, *51*, 331–363. https://doi.org/10.1080/10635150252899806

Franklin, J., Woodcock, C. E., & Warbington, R. (2000). Multi-attribute vegetation maps of forest service lands in California supporting resource management decisions. *Photogrammetric Engineering and Remote Sensing*, *66*, 1209–1218.

Garcia, S., & Herrera, F. (2008). An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. *Journal of Machine Learning Research*, *9*, 2677–2694.

GDAL Development Team (2010). *GDAL – Geospatial data abstraction library*. Version 1.70. Open Source Geospatial Foundation. Retrieved from http://www.gdal.org/.

Ghosh, A., Fassnacht, F. E., Joshi, P. K., & Koch, B. (2014). A framework for mapping tree species combining hyperspectral and LiDAR data: Role of selected classifiers and sensor across three spatial scales. *International Journal of Applied Earth Observation and Geoinformation*, *26*, 49–63. https://doi.org/10.1016/j.jag.2013.05.017

Gilbertson, J. K., & Van Niekerk, A. (2017). Value of dimensionality reduction for crop differentiation with multi-temporal imagery and machine learning. *Computers and Electronics in Agriculture*, *142*, 50–58. https://doi.org/10.1016/j.compag.2017.08.024

Götzenberger, L., de Bello, F., Bråthen, K. A., Davison, J., Dubuis, A., Guisan, A., … Pellissier, L. (2012). Ecological assembly rules in plant communities—approaches, patterns and prospects. *Biological Reviews*, *87*, 111–127. https://doi.org/10.1111/j.1469-185X.2011.00187.x

Gualtieri, J. A. (2009). The support vector machine (SVM) algorithm for supervised classification of hyperspectral remote sensing data. In G. Camps-Valls & L. Bruzzone (Eds.), *Kernel methods for remote sensing data analysis* (pp. 49–83). Chichester, UK: John Wiley & Sons. https://doi.org/10.1002/9780470748992

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, *25*, 1965–1978. https://doi.org/10.1002/(ISSN)1097-0088

Hira, Z. M., & Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data. *Advances in Bioinformatics*, *2015*, Article ID 198363. https://doi.org/10.1155/2015/198363

Howley, T., Madden, M. G., O'Connell, M.-L., & Ryder, A. G. (2006). The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data. *Knowledge-Based Systems*, 19, 363–370. https://doi.org/10.1016/j.knosys.2005.11.014

Kang, S., Ma, W., Li, F. Y., Zhang, Q., Niu, J., Ding, Y., … Sun, X. (2015). Functional redundancy instead of species redundancy determines community stability in a typical steppe of Inner Mongolia. *PLoS ONE*, 10, e0145605. https://doi.org/10.1371/journal.pone.0145605

Kay, M., & Wobbrock, J. (2016). *ARTool: Aligned rank transform for non-parametric factorial ANOVAs*. R package version 0.10.4.

Küchler, A. W., & Zonneveld, I. S. (1988). Floristic analysis of vegetation. In A. W. Küchler, & I. S. Zonneveld (Eds.), *Vegetation mapping* (pp. 51–66). Dordrecht, the Netherlands: Dr W Junk. https://doi.org/10.1007/978-94-009-3083-4

Lamont, B. B., Hopkins, A. J. M., & Hnatiuk, R. J. (1982). The flora-composition, diversity and origins. In J. S. Pate, & J. S. Beard (Eds.), *Kwongan: Plant life of the sandplain* (pp. 27–50). Nedlands, WA, Australia: University of Western Australia Press.

Lee, K., & Lunetta, R. (1996). Wetland detection methods. In J. G. Lyons, & J. McCarthy (Eds.), *Wetland and environmental application of GIS* (pp. 249–284). New York, NY: Lewis Publishers.

Lepš, J., & Šmilauer, P. (2003). *Multivariate analysis of ecological data using CANOCO*. Cambridge, UK: Cambridge University Press.

Lippok, D., Beck, S. G., Renison, D., Hensen, I., Apaza, A. E., & Schleuning, M. (2014). Topography and edge effects are more important than elevation as drivers of vegetation patterns in a neotropical montane forest. *Journal of Vegetation Science*, 25, 724–733. https://doi.org/10.1111/jvs.12132

Liu, C., White, M., Newell, G., & Griffioen, P. (2013). Species distribution modelling for conservation planning in Victoria, Australia. *Ecological Modelling*, 249, 68–74. https://doi.org/10.1016/j.ecolmodel.2012.07.003

Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66, 247–259. https://doi.org/10.1016/j.isprsjprs.2010.11.001

Mucina, L., Stephenson, G., Daniel, G., Van Niekerk, A., & Boonzaaier, I. (2013). A floristic-ecological vegetation map of the Mitchel Plateau Region (Northern Kimberley, Western Australia). In L. Mucina, L. & G. Daniel (Eds.), *Vegetation mapping in the Northern Kimberley, Western Australia* (pp. 88–113). Perth, WA, Australia: Curtin University.

Munyati, C., Ratshibvumo, T., & Ogola, J. (2013). Landsat TM image segmentation for delineating geological zone correlated vegetation stratification in the Kruger National Park, South Africa. *Physics and Chemistry of the Earth, Parts A/B/C*, 55–57, 1–10. https://doi.org/10.1016/j.pce.2009.11.014

Myburgh, G., & Van Niekerk, A. (2013). Effect of feature dimensionality on object-based land cover classification: A comparison of three classifiers. *South African Journal of Geomatics*, 2, 13–27.

Osis, S. T., Hettinga, B. A., & Ferber, R. (2016). Predicting ground contact events for a continuum of gait types: An application of targeted machine learning using principal component analysis. *Gait & Posture*, 46, 86–90. https://doi.org/10.1016/j.gaitpost.2016.02.021

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

Pal, M., & Foody, G. M. (2010). Feature selection for classification of hyperspectral data by SVM. *IEEE Transactions on Geoscience and Remote Sensing*, 48, 2297–2307. https://doi.org/10.1109/TGRS.2009.2039484

Pasolli, E., Truong, D. T., Malik, F., Waldron, L., & Segata, N. (2016). Machine learning meta-analysis of large metagenomic datasets: Tools and biological insights. *PLoS Computational Biology*, 12, e1004977. https://doi.org/10.1371/journal.pcbi.1004977

Pesch, R., Schmidt, G., Schroeder, W., & Weustermann, I. (2011). Application of CART in ecological landscape mapping: Two case studies. *Ecological Indicators*, 11, 115–122. https://doi.org/10.1016/j.ecolind.2009.07.003

Reynolds, H. L., Packer, A., Bever, J. D., & Clay, K. (2003). Grassroots ecology: Plant–microbe–soil interactions as drivers of plant community structure and dynamics. *Ecology*, 84, 2281–2291. https://doi.org/10.1890/02-0298

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & 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*, 67, 93–104. https://doi.org/10.1016/j.isprsjprs.2011.11.002

Roli, F., & Fumera, G. (2001). Support vector machines for remote sensing image classification. *Proceedings SPIE*, 4170, 160–166. https://doi.org/10.1117/12.413892

Rosario-Martinez, H. D. (2015). *Phia: Post-hoc interaction analysis*. R package version 0.2-1.

Schaffer, C. (1993). Overfitting avoidance as bias. *Machine Learning*, 10, 153–178.

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C. T., … Golub, T. R. (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*, 8, 68–74. https://doi.org/10.1038/nm0102-68

Shlens, J. (2014). *A tutorial on principal component analysis. arXiv preprint arXiv:1404.1100*.

Tango, F., & Botta, M. (2013). Real-time detection system of driver distraction using machine learning. *IEEE Transactions on Intelligent Transportation Systems*, 14, 894–905. https://doi.org/10.1109/TITS.2013.2247760

Tayyebi, A., & Pijanowski, B. C. (2014). Modeling multiple land use changes using ANN, CART and MARS: Comparing tradeoffs in goodness of fit and explanatory power of data mining tools. *International Journal of Applied Earth Observation and Geoinformation*, 28, 102–116. https://doi.org/10.1016/j.jag.2013.11.008

Tichý, L., Chytrý, M., Hájek, M., Talbot, S. S., & Botta-Dukát, Z. (2010). OptimClass: Using species-to-cluster fidelity to determine the optimal partition in classification of ecological communities. *Journal of Vegetation Science*, 21, 287–299. https://doi.org/10.1111/j.1654-1103.2009.01143.x

Tsakalos, J. L., Dršková, M., Hruban, J., Mucina, L., & Dobrowolski, M. P. (2014). Floristic patterns and drivers of kwongan vegetation patterns in Eneabba region of the Northern Sandplains, Western Australia. In L. Mucina, J. N. Price, & J. M. Kalwij (Eds.), *Biodiversity and vegetation: Patterns, processes, conservation* (p. 196). Perth, WA, Australia: Kwongan Foundation.

Viscarra Rossel, R. A., Chen, C., Grundy, M. J., Searle, R., Clifford, D., & Campbell, P. H. (2015). The Australian three-dimensional soil grid: Australia's contribution to the Global Soil Map project. *Soil Research*, 53, 845–864. https://doi.org/10.1071/SR14366

Walker, B. H. (1992). Biodiversity and ecological redundancy. *Conservation Biology*, 6, 18–23. https://doi.org/10.1046/j.1523-1739.1992.610018.x

Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. (2011). *The aligned rank transform for nonparametric factorial analyses using only ANOVA procedures. Proceedings of the SIGCHI conference on human factors in computing systems*, May 7–12, 2011 (pp. 143–146). Vancouver, BC, Canada: ACM.

Woodman Environmental Consulting (2009). *Flora, vegetation and fauna impact assessment – Iluka ERMP Areas and cumulative impact assessment*. Unpublished report. Perth, WA, Australia: Woodman Environmental Consulting.

Zhang, H., & Wang, M. (2009). Search for the smallest random forest. *Statistics and Its Interface*, 2, 381–388.

Zhang, C., & Xie, Z. (2013). Object-based vegetation mapping in the Kissimmee River watershed using HyMap data and machine

learning techniques. *Wetlands*, *33*, 233–244. https://doi.org/10.1007/s13157-012-0373-x

Zobel, M. (1992). Plant species coexistence: The role of historical, evolutionary and ecological factors. *Oikos*, *65*, 314–320. https://doi.org/10.2307/3545024

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.