Transactions in GIS

Transactions in GIS, 2010, 14(5): 709–726

Research Article

Modeling the Potential Distribution of Pine Forests Susceptible to *Sirex Noctilio* Infestations in Mpumalanga, South Africa

Riyad Ismail
Department of Geography and Environmental Studies
University of KwaZulu-Natal

Onisimo Mutanga
Department of Geography and Environmental Studies
University of KwaZulu-Natal

Lalit Kumar
Department of Ecosystem Management
University of New England

Abstract
Reducing the impact of the siricid wasp, *Sirex noctilio* is crucial for the future productivity and sustainability of commercial pine resources in South Africa. In this study we present a machine learning model that serves as a spatial guide and allows forest managers to focus their existing detection and monitoring efforts on key areas and proactively adopt the most appropriate course of intervention. We implemented the random forest model within a spatial framework to determine which pine forests in Mpumalanga are highly susceptible to *S. noctilio* infestations. Results indicate that a majority (63%) of pine forest plantations located in Mpumalanga have a high susceptibility (>70%) to *S. noctilio* infestation. A KHAT value of 0.84 and F measures above 0.87 indicate that the random forest model is a robust classifier that produces accurate results. Additionally, the use of the backward variable selection method enabled us to simplify the random forest modeling process and identify the minimum number of explanatory variables that offer the best discriminatory power and help in the empirical interpretation of the final random forest model. Overall, the results show that pine forests that experience stress caused by evapotranspiration and evaporation followed by rainfalls, especially during the summer months are more susceptible to *S. noctilio* infestations.

Address for correspondence: Riyad Ismail, Department of Geography and Environmental Studies, University of KwaZulu-Natal, Private Bag X01, Scottsville 3209, South Africa. E-mail: riyad.ismail@sappi.com

doi: 10.1111/j.1467-9671.2010.01229.x

1 Introduction

Reducing the impact of the siricid wasp, *Sirex noctilio* (Hymenoptera: Siricidae) is crucial for the future productivity and sustainability of commercial pine resources in South Africa. *S. noctilio* has caused extensive damage to pine forests located in KwaZulu-Natal and the Eastern Cape (Hurley et al. 2007, Ismail et al. 2007, Slippers 2006). It is now a major concern that the wasp will spread further north, to the province of Mpumalanga, where the majority of the country's pine forests are located (DWAF 2005). Detection and monitoring methods have been identified as important tools that provide forest managers with valuable information on the current location and extent of *S. noctilio* infestations (Carnegie 2005, Haugen et al. 1990, Hurley et al. 2007, Ismail et al. 2007).

Researchers have recommended the combined use of aerial and field surveys (Carnegie 2005, Haugen 1990) or the use of multispectral remote sensing (Ismail et al. 2007, 2008) to spatially quantify the location and extent of *S. noctilio* infestations. Due to operational limitations (namely, cost and labor), and the initial scattered pattern of infestations (Ciesla 2003), it is not feasible to implement any of the suggested detection and monitoring methods consistently at national or provincial levels. Therefore, the strength of current detection and monitoring methods would be greatly enhanced if we could proactively identify pine forests that are highly susceptible to *S. noctilio* infestations before any concerted monitoring and detection methods are implemented. Maps showing the distribution of susceptible forests will then serve as a spatial guide and allow forest managers to focus their existing detection and monitoring efforts to these key areas ("hotspots"). Additionally, forest managers will have the ability to adopt the most appropriate remediation measures (Carnegie 2005, Haugen 1990, Haugen and Underdown 1990, Neumann and Minko 1981, Spradberry and Kirk 1978, Taylor 1981, Tribe and Cillie 2004) before the wasp can colonize these uninfected pine forests.

Statistical modeling approaches have been increasingly recognized as important tools that improve our understanding of forest pests and pathogens. When used within a spatial framework, these models have the ability to identify areas that are highly susceptible to infestations (Candau and Fleming 2005, Carnegie et al. 2006, Guo et al. 2005, Kelly and Meentemeyer 2002, Negron 1998, Rosso and Hansen 2003, van Staden et al. 2004). For example, Carnegie et al. (2006) developed a model based on climate matching in order to understand the potential global distribution of *S. noctilio*. However, the explanatory variables used in the CLIMEX model (http://www.hearne.com.au/products/climex) were based on the wasp's endemic habitat conditions in Eurasia and northern Africa. These areas experience dry warm summers and cool moist winters (Carnegie 2005), whereas, in contrast, *S. noctilio* has successfully established itself in the summer rainfall areas of South Africa (Hurley et al. 2007, 2008). With the exception of the CLIMEX model, spatially-based studies that empirically relate the potential distribution of *S. noctilio* infestations to a set of explanatory variables (for example, environmental data) are non-existent. Therefore, it would be beneficial from a pest management perspective, to model pine forests that are highly susceptible to *S. noctilio* infestations at a more regional scale in an effort to understand localized variations of environmental conditions in relation to the distribution of the wasp.

Machine learning techniques such as classification and regression trees or C&RT (Breiman et al. 1984) have been frequently used to model the damage associated with

forest pests and pathogens (Candau and Fleming 2005, Kelly and Meentemeyer 2002, Kelly et al. 2007, Rosso and Hansen 2003). C&RT are non-parametric models that construct a set of decision rules by recursively splitting the response variable (for example, species data) into smaller homogenous groups, where each split is based on a single explanatory variable. The final output is a tree diagram with the terminal nodes of the tree indicating the final response (De'ath and Fabricius 2000, Prasad et al. 2006, Vayssieres et al. 2000). C&RT are popular amongst researchers because the model has the following benefits: no advanced variable selection is required; explanatory variables do not need to have a Gaussian distribution; due to the graphical nature of the tree; the results are easy to interpret; the model can use a combination of categorical and continuous explanatory variables; the model has the ability to capture hierarchical and non-linear relationships and finally; the model provides insight into the spatial influence of the explanatory variables (De'ath and Fabricius 2000, Kelly and Meentemeyer 2002, Kelly et al. 2007, Prasad et al. 2006, Vayssieres et al. 2000). However, C&RT are very sensitive to small changes in the training dataset and have been identified as unstable classifiers that are prone to overfitting (Breiman 1996). Researchers have suggested that by bootstrapping (Efron and Tibshirani 1993) the original training dataset and then averaging the class predictions, C&RT can be stabilized (Archer and Kimes 2008).

The Breiman-Cutler, random forest (RF) model is an improvement of C&RT that includes bootstrap aggregation (bagging) and randomly selects a subset of explanatory variables to create an ensemble classifier that avoids overfitting and is successful in combining unstable learners like C&RT (Breiman 2001). The RF model has been notably exploited for the analysis of microarray data (Archer and Kimes 2008, Diaz-Uriarte and Alvarez de Andres 2006, Jiang et al. 2004, Strobl et al. 2007). However, in recent years, researchers have successfully applied the RF model to a variety of spatial datasets. Within a spatial framework, RF has been used to map invasive plants (Lawrence et al. 2006), land cover (Gislason et al. 2006, Pal 2005), tick-borne disease (Furlanello et al. 2003), climate change (Leng et al. 2007, Prasad et al. 2006) and habitat suitability (Garzon et al. 2006).

Results from the above studies indicated that the RF model is competitive with commonly used modeling approaches and provides an effective method for estimating the importance of explanatory variables. More specifically, Garzon et al. (2006) compared the predictive ability of RF, neural networks and C&RT to map the distribution of *Pinus sylvestris*. Results from the study, concluded that RF was the most accurate classifier followed by neural networks and then by C&RT. Prasad et al. (2006) modeled the distribution of loblolly pine, sugar maple, American beech and white oak under current and future climate scenarios using regression trees, RF, bagging trees and multivariate adaptive regression splines. Results from the study indicated that random forest was superior in reproducing the current and future distribution of the four tree species. Using the RF internal measure of variable importance, Furlanello et al. (2003) determined that climatic variables are vital for predicting the occurrence of ticks. The experimental results were considered novel for the study area, since previous published models indicated that the geological substratum was important for identifying tick habitats (Furlanello et al. 2003).

In this study, we expanded the application of RF to susceptibility mapping. We implemented the RF model within a spatial framework to determine which pine forests in an unaffected area (i.e. Mpumalanga), are susceptible to *S. noctilio* infestations. We

assumed that if pine forests in Mpumalanga share similar environmental conditions with those areas with confirmed *S. noctilio* infestations in KwaZulu-Natal, they are more likely to be susceptible to infestation. More specifically, we examined the robustness of RF, firstly in terms of its classification accuracy and secondly for the empirical selection of explanatory variables. This study ultimately focused on developing a Geographic Information Systems (GIS) susceptibility model that could eventually be applied to all pine forests located in South Africa and for the first time introduced RF for mapping pine forests that are susceptible to *S. noctilio* infestations. Although RF is capable of carrying out regression as well as classification (Liaw and Wiener 2002), this study focused on classification, since the response variable in this study was binary and denoted the absence or presence of *S. noctilio* infestations.

2 Methods and Materials

2.1 Response and Explanatory Variables

We assessed the robustness and accuracy of the RF model by applying the algorithm to 1301, Sappi and Mondi, *Pinus patula* compartments located in the southern region of KwaZulu-Natal (Figure 1). These compartments were field checked and visually inspected for the presence or absence of *S. noctilio* infestations by experienced forestry personnel over a period of three years (i.e. from 2004 to 2006). Additionally, in compartments that were classified as infested, a subset of infested trees was destructively sampled by foresters to verify the presence or absence of *S. noctilio* larvae. Of the 1301 response variables, 458 (35.20%) had *S. noctilio* infestations and 843 (64.80%) had no *S. noctilio* infestations. The response variables were further divided into a training dataset for model development and a test dataset for independent accuracy assessments (Table 1). Additionally, class frequencies (absence and presence) were approximately balanced in both the test and training datasets.

The explanatory variables used in this study consisted of one minute by one minute, historical climatic as well as topographic layers projected to Transverse Mercator (Hartebeesthoek; central meridian 31). The climatic variables used for developing the model were obtained from the South African agrohydrology atlas (Schulze et al. 1997) and included: mean annual precipitation, mean annual temperature, monthly median rainfall, monthly minimum temperature, monthly maximum temperature, monthly solar radiation, monthly evapotranspiration and monthly potential evaporation. These historical climatic datasets (1990–1997) were derived from 1,000 meteorological stations located across South Africa (van Staden et al. 2004) and a detailed methodological description of the datasets is provided by Schulze et al. (1997).

The topographic variables used in the study consisted of a digital elevation model (DEM), slope and aspect. The DEM (90 m spatial resolution) was derived from shuttle radar topographic mission (STRM) data and was obtained from the global land cover facility (GLCF) at the University of Maryland (http://glcf.umiacs.umd.edu/index.shtml). Slope (percentage) and aspect (degrees) were then calculated from the DEM using Spatial Analyst (ESRI 2006). Data from climatic and topographic datasets (n = 77) were then extracted for the test and training datasets using the zonal statistics functionality in ArcGIS 9.1 (ESRI 2006). The complete list of explanatory variables used in this study is shown in Table 2.

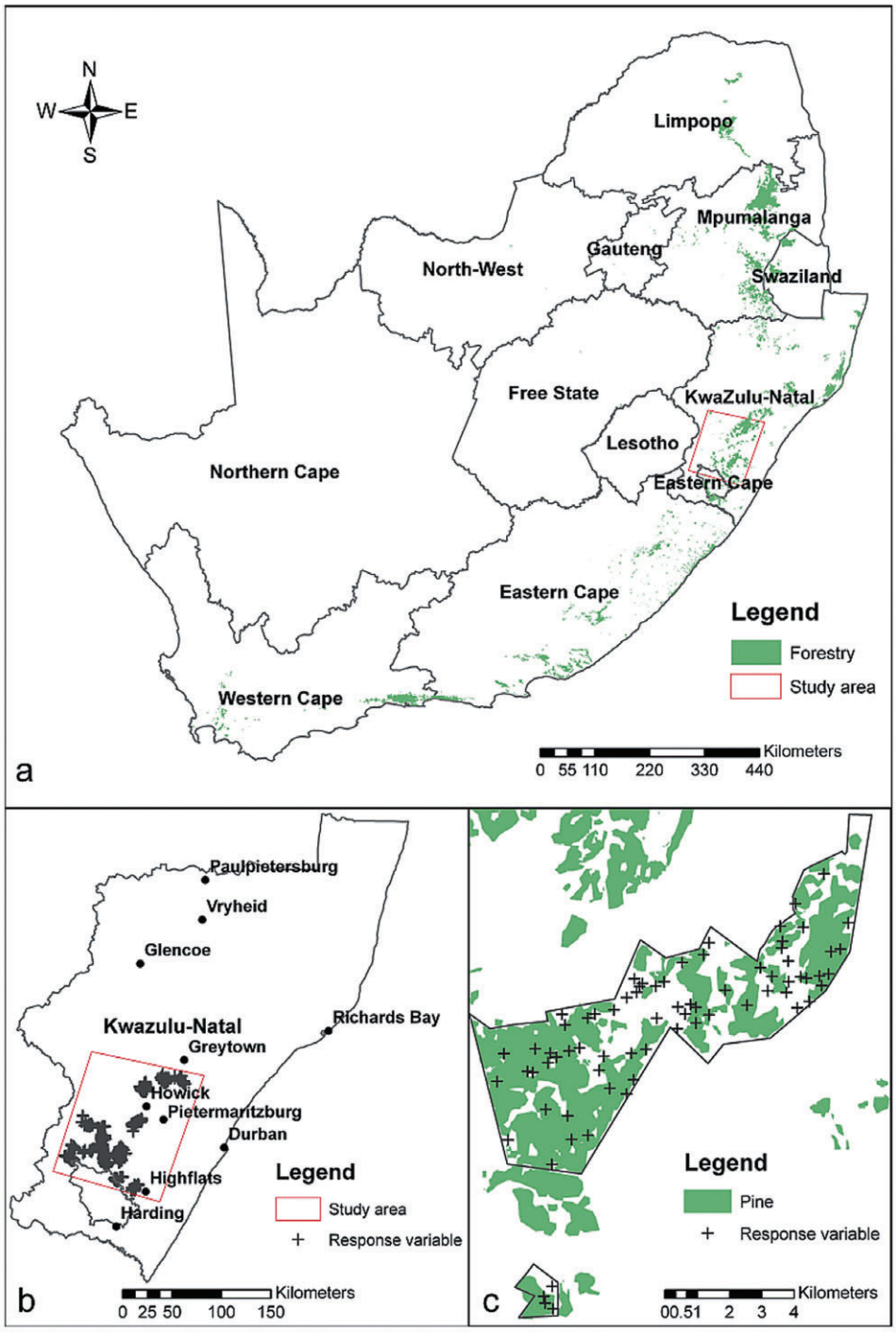

Figure 1 Maps (inserts (a) and (b)) showing the study area in relation to the spatial distribution of commercial forestry plantations in South Africa. Insert (c) provides a detailed view of the samples that were collected at the Sappi Pinewoods plantation

Table 1 Training and test datasets used in the study

	Training dataset	Test dataset
Presence (Y)	321 (35.24%)	137 (35.13%)
Absence (N)	590 (64.76%)	253 (64.87%)
Total	911 (70.02%)	390 (29.98%)

Table 2 Climatic and topographic datasets used in the study

Variable	Abbreviation	Description	Coverage
Solar radiation	SR	Monthly solar radiation	January to December
Precipitation	MAP	Mean annual precipitation	
	MR	Median rainfall	January to December
Temperature	MAXT	Daily maximum temperature	January to December
	MINT	Daily minimum temperature	January to December
	MAT	Mean annual temperature	
Evaporation	APAN	Potential evaporation	January to December
	PEMO	Potential evapotranspiration	January to December
Digital elevation model	DEM	Elevation	
	SLOPE	Slope	
	ASPECT	Aspect	

2.2 Model Description

2.2.1 Random forest

The RF methodology was developed by Breiman (2001) and uses a classification tree as the base classifier. Firstly, RF generates an ensemble of classification trees with each tree in the ensemble grown to maximum size without any pruning. The classification trees in the ensemble then vote by plurality on the correct classification. Secondly, RF searches only across randomly selected subsets of explanatory variables to determine the split at each node. By limiting the number of variables used for each split, the computational complexity of the RF algorithm is reduced and the correlation between trees in the ensemble decreases. Each tree in the ensemble is then constructed using bootstrapped sampling with replacement and contains randomly drawn samples from approximately two-thirds of the samples from the original training dataset. The excluded one-third of the random samples that are left out from each bootstrapped sample is known as the out of bag (OOB) samples. Finally, the OOB samples are then used to determine misclassification error and variable importance. The misclassification error (or OOB error) is calculated by putting each OOB sample down the corresponding classification tree from which it was excluded. The error estimate is then calculated as the misclassified proportion of that OOB sample (Breiman 2001, Garzon et al. 2006, Gislason et al. 2006, Liaw and Wiener 2002, Pal 2005, Peters et al. 2007, Prasad et al. 2006). The

calculation of the variable importance using the OOB sample is described in section 2.2.2.

There are only two tuning parameters required for RF, the number of trees in the ensemble (*ntree*), and the number of possible splitting variables (*mtry*) which are sampled at each node (Peters et al. 2007). Researchers have shown that sensitivity of the *ntree* and *mtry* parameters are minimal and that the default values are often a good choice (Lawrence et al. 2006, Liaw and Wiener 2002). However, in this study we optimised both parameters using the OOB error. A more detailed statistical description of the algorithm is provided by Breiman (2001). We used the *randomForest* library (Liaw and Wiener 2002) for the R statistical software (R Development Core Team 2008) to implement the algorithm.

2.2.2 Using random forest for variable selection

The importance of variable selection is not only to reduce the amount of explanatory variables used in the analysis, but also to improve our understanding of which explanatory variables are most suitable for modeling the distribution of pine forest that are susceptible to *S. noctilio* infestations. RF calculates the importance of each explanatory variable by random permutation of all values of the explanatory variables in the OOB sample. The number of votes for the correct class in the permutated data is subtracted from the number of correct votes in the original data which is then averaged over all trees in the forest. This represents the importance value for each explanatory variable and is the percentage increase in the misclassification rate as compared to the OOB rate of the non-permutated data (Prinzie and Van den Poel 2008). As opposed to other methods of calculating variable importance (for example, the Gini index), the permutation method is regarded as the most reliable measure for determining variable importance (Breiman 2001).

However, it is often difficult to set a cut off value when there are many explanatory variables and when most of them have very similar importance measures (Jiang et al. 2004). Also, in order to simplify the modeling process we would like to identify the smallest number of explanatory variables that offer the best discriminatory power and help in the empirical interpretation of the final *S. noctilio* susceptibility model. To address these issues we examined two variable selection methods which iteratively measure the importance of each explanatory variable (as determined by RF) and then remove the least relevant explanatory variables. The backward variable selection method builds multiple RF's and after building each RF iteratively discards those explanatory variables with the smallest variable importance as determined by the OOB error rate (Diaz-Uriarte and Alvarez de Andres 2006). The recursive variable selection method is very similar to the backwards approach except that variable importance is recalculated for each RF that is built, thus producing a new ranking of variables before the variables with the smallest importance are discarded (Jiang et al. 2004, Svetnik et al. 2003). We used the *varSelRF* library (Diaz-Uriarte and Alvarez de Andres 2006) for the R statistical software to implement the recursive and the backward variable selection methods.

2.2.3 Accuracy assessments

It has been suggested that when using RF there may be no need for cross validation or a separate test dataset to determine the misclassification error because the OOB error

provides an unbiased estimate of error (Lawrence et al. 2006, Prasad et al. 2006, Prinzie and Van den Poel 2008). However, according to Diaz-Uriarte and Alvarez de Andres (2006) and Granitto et al. (2006) using the OOB error to determine the misclassification error could result in a biased estimation of the error because the samples used to calculate the error are not independent of the model being evaluated. In this study, the OOB error was used to fine tune the user-defined *mtry* parameter and for the empirical selection of explanatory variables and not to calculate the final misclassification error (Diaz-Uriarte and Alvarez de Andres 2006).

To avoid bias in the accuracy assessments we use an independent test dataset (n = 390) to calculate the final misclassification error (Reunanen 2003) and the results were tabulated using a confusion matrix. Several measures can be calculated from the confusion matrix (Fielding and Bell 1997), however, we calculated the precision and recall measures because our main interest was in correctly modeling "presence" rather than "absence" (Peters et al. 2007). From the confusion matrix, precision (p) is calculated as the proportion of predicted presences that are observed to be present rather than absent and is defined as: $p = TP/TP + FP$. Recall (r) is calculated as the proportion of observed presences that were predicted correctly and is defined as: $r = TP/TP + FN$ (Peters et al. 2007). The weighted F measure (van Rijisbergen 1979) that combines precision and recall is stated as:

$$F_\beta(p, r) = (\beta^2 + 1)pr/\beta^2 p + r \tag{1}$$

where β is the weighting factor that controls the relative importance of precision versus recall. If $\beta = 1$, precision and recall have equal importance if $\beta = 0.5$, precision is twice as important as recall and if $\beta = 2$ then recall is twice as important as precision. According to Peters et al. (2007) the magnitude of F varies from no predictive power (0) to perfect prediction (1). Furthermore, we carried out a discrete multivariate technique called Kappa analysis to determine if the overall classification as determined by RF was better than if it was classified by a random classifier. The result of performing a Kappa analysis is the k (KHAT) statistic which measures agreement or accuracy (Cohen 1960). KHAT values range from –1 to +1 and if the values are one or close to one then there is perfect agreement between test and training datasets (Congalton and Green 1999, Lillesand et al. 2004, Skidmore 1999).

3 Results

3.1 Fine Tuning Random Forest

Before using RF to model the potential distribution of pine forests that are susceptible to *S. noctilio* infestations, we examined the effect of the number of randomly selected variables (*mtry*) on the classification error. According to Peters et al. (2007) reducing the *mtry* value decreases the strength of individual trees (resulting in an increase in classification error) and the correlation between any two trees in the forest (resulting in a decrease in classification error). Therefore, the user defined *mtry* value has to be optimized in order to achieve a minimal classification error. Four different sized RF models (*ntree*) were constructed for all possible unique values (n = 77) of *mtry*. We then used the lowest OOB error to determine the optimal *mtry* value (Diaz-Uriarte and Alvarez de Andres 2006, Granitto et al. 2006, Peters et al. 2007, Svetnik et al. 2003).

Table 3 Maximum and minimum OOB errors obtained using four different *ntree* values and all possible *mtry* values

ntree value	100	200	500	1000
Minimum OOB error	9.22%	8.89%	9.00%	9.22%
Optimal *mtry* value	2	3	6	11
OOB error (*mtry* = 3)	9.55%	8.89%	9.33%	9.44%

Table 3 shows that the lowest OOB error (8.89%) is obtained when 200 trees are built using an *mtry* value of three. Furthermore, using an *mtry* value of three for the other models (100, 500 and 1,000) produces a negligible increase in OOB error (<1%). Similarly, when classifying microarray data Diaz-Uriarte and Alvarez de Andres (2006) show that the OOB error rate is largely independent of *ntree* sizes even for *ntree* values ranging from 1,000 to 40,000 trees. Based on the results obtained we opted to use the following parameters: *mtry* = 3 with *ntree* = 200 because of the low OOB error produced.

3.2 Variable Selection Using Backward and Recursive Approaches

As mentioned earlier, the RF model estimates the importance of an explanatory variable by looking at how much the OOB error increases when the OOB data for that particular explanatory variable is permutated while the other variables are not permutated. Figure 2 shows the mean decrease in accuracy of explanatory variables (n = 35) as determined by the OOB error. For visualization purposes, only variables that have a greater than 30% decrease in accuracy are shown.

Results show that the highest ranked variables in terms of their importance include: evapotranspiration (April and August) followed by the median rainfall during the summer months (February, April, January, November and December). Additional high ranked variables include: solar radiation, evaporation and slope. Temperature (minimum and maximum), aspect and the digital elevation model are low ranked and have very low importance scores (not shown in Figure 2). To determine the minimum number of explanatory variables required to accurately model the potential distribution of *S. noctilio* infestations, we implemented the backward and recursive variable selection methods. For both variable selection methods, 20% of the least important explanatory variables (as determined by the RF model) were discarded from the previous RF iteration. This allowed for faster computations and is based on an aggressive variable selection approach (Diaz-Uriarte and Alvarez de Andres 2006). Figure 3 shows the results for both variable selection methods.

Results show that by using 14 variables the minimum OOB error obtained for the recursive variable selection method was 8.34% and by using 21 variables the minimum OOB error was 8.23% for the backward variable selection method. However, the best solution is based on selecting the least amount of variables with the proviso that the final solution has an OOB error rate that is within one standard error of the minimum error rate of all forest created (Diaz-Uriarte and Alvarez de Andres 2006). Under these conditions the recursive method then selects the best solution based on seven variables

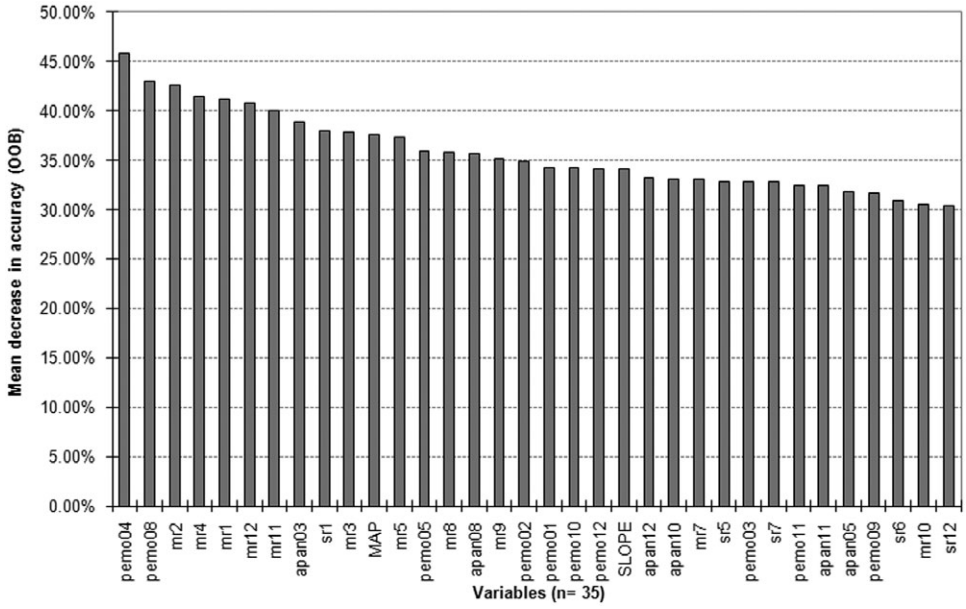

Figure 2 Variable importance as determined by random forest (*mtry* = 3 and *ntree* = 200). The full names for the variables are shown in Table 2 and the numbers refer to the month of the year

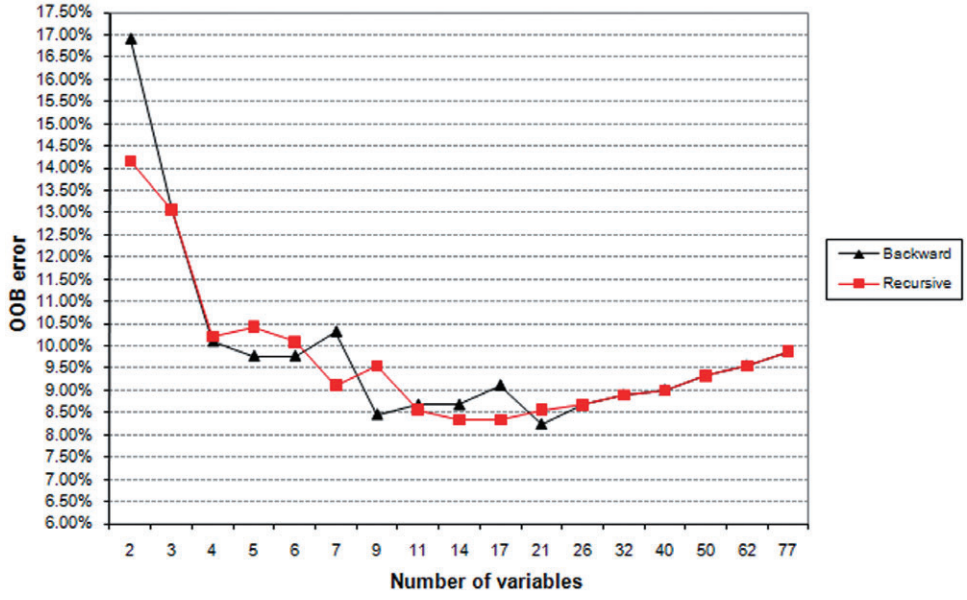

Figure 3 The OOB error obtained during the backward and recursive variable selection process. The arrows indicate the number of explanatory variable that produce an OOB error within one standard deviation of the lowest OOB error

with an OOB error of 9.11%, while the backward method selects nine variables with an OOB error of 8.45%. The variables selected by the recursive method were as follows: median rainfall for January (MR1), median rainfall for February (MR2), median rainfall for April (MR4), median rainfall for November (MR11), evapotranspiration for April (PEMO4), evapotranspiration for August (PEMO8) and potential evaporation for August (APAN08). The variables selected by the backward method included: the mean annual precipitation (MAP), median rainfall for January (MR1), median rainfall for February (MR2), median rainfall for April (MR4), median rainfall for December (MR12), evapotranspiration for April (PEMO4), evapotranspiration for August (PEMO8), evapotranspiration for October (PEMO10) and potential evaporation for August (APAN08). The backward variable selection method provides the better solution with a lower OOB error than the recursive approach.

3.3 Stability of the Backward Variable Selection Method

According to Granitto et al. (2006) the selection of explanatory variables is an unstable process and could lead to the selection of very different subsets of explanatory variables for each replicate of the study. To examine the stability of the RF model with the backward variable selection method, we determined the number of times an explanatory variable (MAP, MR1, MR2, MR4, MR12, PEMO4, PEMO8, PEMO10 and APAN08) is selected when the backward variable selection method is bootstrapped (n = 100) (Efron and Thibshirani 1997). Results indicate that all the variables selected using the backward method have a very high selection probability (90% and greater). According to Figure 4 the explanatory variables with the highest probability of selection are: MAP (99%), MR4 (97%), PEMO08 (97%), APAN08 (96%) followed by PEMO04 (96%). The variables selected using the backward method (n = 9) were then used as the input explanatory variables for the final RF model.

3.4 Classification Accuracy

To evaluate the accuracy and robustness of RF for mapping the potential spatial distribution of *S. noctilio* infestations, we compared the RF accuracy assessments against the widely used C&RT. Table 4 shows the accuracy assessments for both machine learning models. The KHAT value obtained by the RF model (0.84) is much higher than the KHAT value obtained by the C&RT (0.74), indicating that there is a strong agreement between the observations (n = 911) and the RF model predictions (n = 390). For both models, precision and recall are high, implying that there was more correctly predicted presence rather than absence of *S. noctilio* infestations. However, the weighted F measures ranges from 0.78 to 0.87 for C&RT while the weighted F measures for the RF model were all above 0.87. Overall, the RF model produces better results than C&RT as determined by the weighted F measures as well as by the kappa analysis.

3.5 Modeling Sirex Noctilio Susceptibility

Finally, we extrapolated the RF model developed for KwaZulu-Natal to all pine forest plantations located in Mpumalanga (Figure 5a). Each pixel (one minute by one minute) that contained pine forests was classified 200 times and the proportion of votes over all 200 trees indicated the susceptibility to *S. noctilio* infestations. Figure 5b shows

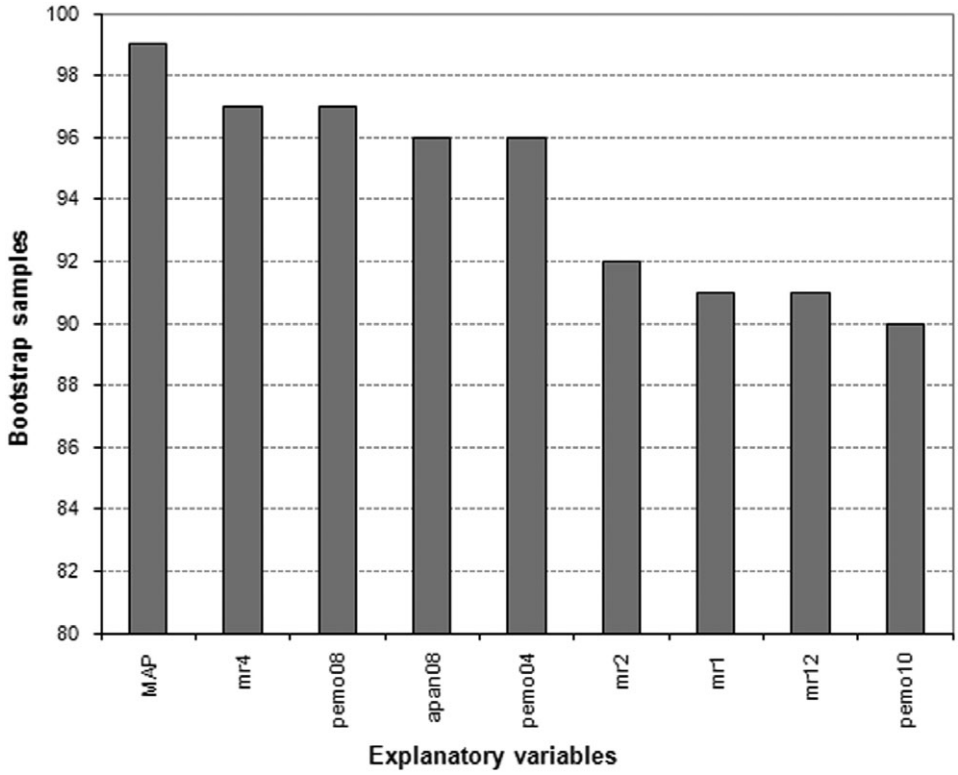

Figure 4 The number of times each explanatory variable is selected during the bootstrap process (n = 100). The full names for the variables are shown in Table 2 and the numbers refer to the month of the year

Table 4 Accuracy assessments using the test dataset (n = 390)

	Classification and regression trees (C&RT)	Random forest (RF)
Precision	0.90	0.91
Recall	0.76	0.88
F_2	0.78	0.89
F_1	0.82	0.90
$F_{0.5}$	0.87	0.90
KHAT	0.74	0.84

the potential distribution of pine forest plantations that are susceptible to *S. noctilio* infestations in Mpumalanga as determined by RF model. Of the 1,909 pixels that were classified, 1,204 pixels have a high susceptibility (>70%) to *S. noctilio* infestation with the remaining pixels (n = 705) having a moderate (50%–70%) to low (<50%)

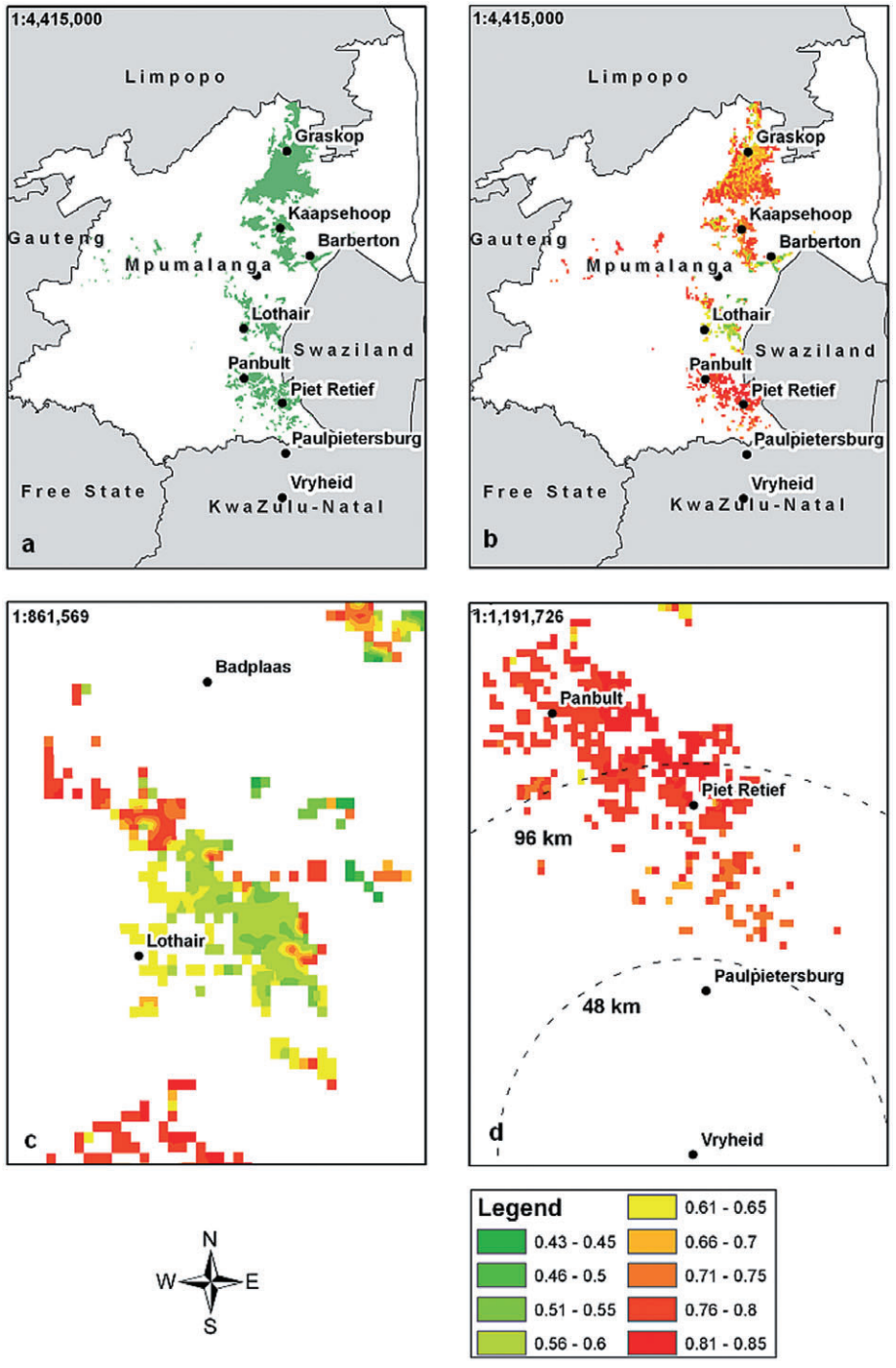

Figure 5 Insert (a) shows the current distribution of pine forest located in Mpumalanga. Insert (b) shows the potential distribution of pine forests that are susceptible to *Sirex noctilio* infestations in Mpumalanga. Inserts (c) and (d) provide a detailed view of pine forests that are susceptible to *Sirex noctilio* infestations

susceptibility to *S. noctilio* infestation. Overall, the majority (63%) of pine forest plantations located in Mpumalanga have a high susceptibility to *S. noctilio* infestation with the exception of pine plantations located in the vicinity of Lothair (Figure 5c), which have a moderate to low susceptibility to *S. noctilio* infestation.

4 Discussion

4.1 Classification Accuracy

The results obtained from this study are very encouraging and show that RF is a robust and accurate machine learning technique that can be implemented within a spatial framework to determine which pine forests are susceptible to *S. noctilio* infestations. Comparisons between RF and C&RT confirm that RF produces better accuracies than C&RT (Breiman 2001, Garzon et al. 2006, Prasad et al. 2006). More specifically, results from this study indicate that RF obtains a KHAT value of 0.84 while C&RT obtained a KHAT value of 0.74. Furthermore, Gislason et al. (2006) and Hamza and Larocque (2005) showed that RF obtains the best overall classification results even when compared to other ensemble methods that use tree classifiers as the base model. Overall, the RF model is relatively easy to implement and only requires the user to specify the (1) number of trees to be grown (*ntree*) and (2) number of variables to split the nodes of individual trees (*mtry*). While results of this study indicated that these parameters have to be optimized to produce the best results, researchers have noted that the default *mtry* (square root of the total amount of variables) and *ntree* (500) values often produce acceptable results (Liaw and Wiener 2002). This is an important property of RF because the model can be run with minimal human guidance (Gislason et al. 2006).

4.2 Variable Importance

In addition to providing accurate classification results, the RF model also provided insight into which variables are most important with respect to the modeling process. The backward variable selection method enabled us to simplify the RF modeling process successfully and identify the smallest number of explanatory variables that offer the best discriminatory power and help in the empirical interpretation of the final model. These findings were also reported by Diaz-Uriarte and Alvarez de Andres (2006) and Jiang et al. (2004), when they applied variable selection and RF to microarray datasets. More specifically, results from this study showed that by implementing the backward variable selection method, we reduced the total number of explanatory variables (n = 77) to an optimal number of variables (n = 9) that explained the presence or absence of *S. noctilio* infestations.

Variables that were selected by the backward variable selection method included: mean annual precipitation (MAP), median rainfall for January (MR1), median rainfall for February (MR2), median rainfall for April (MR4), median rainfall for December (MR12), evapotranspiration for April (PEMO4), evapotranspiration for August (PEMO8), evapotranspiration for October (PEMO10) and potential evaporation for August (APAN08). Although the backward variable selection method is not expected to describe the causal relationship between the explanatory variables and presence or

absence of S. *noctilio* infestations, results empirically show that pine forests that experience stress caused by evapotranspiration and evaporation (PEMO4, PEMO8, PEMO10 and APAN08) followed by rainfalls especially during the summer months (MR1, MR2, MR4, MR12) are more susceptible to *S. noctilio* infestations. These results are consistent with the view of Madden (1988) who hypothesized that intermittent stress (for example drought) contributes significantly to *S. noctilio* outbreaks by increasing tree attractiveness and susceptibility through rapid physiological changes following rains of short duration. Additionally, it is well documented that trees that are experiencing stress are more likely to be attacked by *S. noctilio*. For example, researchers have suggested that pine forests that experience drought, fire or have a high density of tree plantings are more likely to be attacked by *S. noctilio* (Ciesla 2003, Haugen and Underdown 1990, Tribe and Cillie 2004).

4.3 Modeling Susceptibility

Developing a model that spatially defines the potential distribution of pine forests that are susceptible to *S. noctilio* infestations is an important step in understanding the nature of the epidemic in South Africa. *S. noctilio* is currently the most important pest of pines in South Africa (Hurley et al. 2008). Knowledge of the potential distribution of pine forests that are susceptible to *S. noctilio* infestations is important because it serves as a spatial guide and allows forest managers to focus their existing detection and monitoring efforts on key areas and to proactively adopt the most appropriate course of intervention before the wasp actually colonizes these unaffected pine forests located in Mpumalanga. For example, results show that a potential hotspot exists around the town of Piet Retief (Figure 5d). Pine forests located in the vicinity are highly susceptible to *S. noctilio* infestations and are within a close proximity to the current *S. noctilio* infestation in KwaZulu-Natal (Vryheid). With an annual flight radius of 48 km (Tribe and Cillie 2004), the wasp will most probably colonize pine forest in the area within the next two years. It is recommended that pine forests located in the area should be continuously monitored for the early symptoms of *S. noctilio* infestation and prioritized for remediation efforts.

Remediation of established *S. noctilio* populations is achieved by biological means using the nematode *Beddingia siricidicola* and by using various parasitic wasps (Carnegie 2005, Ciesla 2003, Hurley et al. 2007, Tribe and Cillie 2004). However, in unaffected pine forests in Mpumalanga, silvicultural practices, especially thinnings, have been recommended to improve tree vigor and to increase resistance to future *S. noctilio* infestations (Hurley et al. 2007). It is important, however, that thinnings are not carried out during the flight season as the practice could increase stress and favor a build up *S. noctilio* infestation (Carnegie 2005).

5 Conclusions

The RF model when used in conjunction with GIS provides a useful and robust tool that can assist with current forest pest management initiatives. The added benefit of using the RF model is that it only requires the fine tuning of two user-defined parameters in order to achieve good classification. Overall, there is a high probability of *S. noctilio* infestation for the majority (63%) of pine forest plantations located in Mpumalanga. Compared to

previous studies the RF model identified highly susceptible pine forests at a more regional scale and provided an understanding of localized variations of environmental conditions in relation to the distribution of the wasp. Knowledge of the potential distribution of pine forests that are susceptible to *S. noctilio* infestations is important because it serves as a guide and allows forest managers to focus their existing detection and monitoring efforts to key areas and proactively adopt the most appropriate course of intervention before the wasp actually colonizes these unaffected pine forests.

Acknowledgements

We thank Mondi and Sappi for allowing us access to field data. This project was partially funded by the National Research Foundation (South Africa). The early contributions of Marcel Verleur, Philip Croft and Mark Norris-Rogers are gratefully acknowledged.

References

Archer K J and Kimes R V 2008 Empirical characterization of random forest variable importance measures. *Computational Statistics and Data Analysis* 52: 2249–60

Breiman L 1996 Bagging predictors. *Machine Learning* 26: 123–40

Breiman L 2001 Random forests. *Machine Learning* 45: 5–32

Breiman L, Friedman J, Olshen R, and Stone C 1984 *Classification and Regression Trees*. Monterey, CA, Wadsworth and Brooks

Candau J and Fleming R A 2005 Landscape-scale spatial distribution of spruce defoliation in relation to bioclimatic conditions. *Canadian Journal of Forest Resources* 35: 2218–32

Carnegie A J 2005 History and management of sirex wood wasp in pine plantations in New South Wales, Australia. *New Zealand Journal of Forestry* 35: 3–24

Carnegie A, Matsuki M, Haugen D A, Hurley B, Ahumada R, Klasmer P, Sun J, and Iede E 2006 Predicting the potential distribution of *Sirex noctilio* Fabricius (Hymenoptera: Siricidae), a significant exotic pest of Pinus plantations. *Annals of Forest Science* 63: 119–28

Ciesla W M 2003 European woodwasp: A potential threat to North America's conifer forests. *Journal of Forestry* 101: 18–23

Cohen J 1960 A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20: 37–46

Congalton R G and Green K 1999 *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*. Boca Raton, FL, Lewis Publishers

De'ath G and Fabricius K E 2000 Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology* 81: 3178–92

Diaz-Uriarte R and Alvarez de Andres S 2006 Gene selection and classification of microarray data using random forest BMC. *Bioinformatics* 7: 1–13

DWAF 2005 *Commercial Timber Resources and Primary Roundwood Processing in South Africa 2003/2004*. Pretoria, Department of Water Affairs and Forestry

Efron B and Thibshirani R J 1997 Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association* 92: 548–60

Efron B and Tibshirani R 1993 *An Introduction to Bootstrapping*. Boca Raton, FL, Chapman and Hall

ESRI 2006 *ArcGIS 9.1*. Redlands, CA, ESRI Press

Fielding A H and Bell J F 1997 A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation* 24: 38–49

Furlanello C, Neteler M, Merler S, Menegon S, Fontanari S, Donini A, Rizzoli A, and Chemini C 2003 GIS and random forest predictor: Integration in R for tick-borne disease risk assessment. In *Proceedings of the Third International Workshop on Distributed Statistical Computing (DSC 2003)*, Vienna, Austria: 1–10

Garzon M B, Blazek R, Neteler M, de Dios R S, Ollero H S, and Furlanello C 2006 Predicting habitat suitability with machine learning models: The potential area of *Pinus sylvestris* L. in the Iberian Peninsula. *Ecological Modelling* 197: 383–93

Gislason P O, Benediktsson J A, and Sveinsson J R 2006 Random Forests for land cover classification. *Pattern Recognition Letters* 27: 294–300

Granitto P M, Furlanello C, Biasioli F, and Gasperi F 2006 Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems* 83: 83–90

Guo Q, Kelly M, and Graham C 2005 Support vector machines for predicting distribution of Sudden Oak Death in California. *Ecological Modelling* 182: 75–90

Hamza M and Larocque D 2005 An empirical comparison of ensemble methods based on classification trees. *Journal of Computation and Simulation* 75: 629–43

Haugen D A 1990 Control procedures for *Sirex noctilio* in the Green Triangle: Review from detection to severe outbreak (1977–1987). *Australian Forestry* 53: 24–32

Haugen D A and Underdown M G 1990 *Sirex noctilio* control program in response to the 1987 Green Triangle outbreak. *Australian Forestry* 53: 33–40

Haugen D A, Bedding R A, Underdown M G, and Neumann F G 1990 National strategy for control of *Sirex noctilio* in Australia. *Australian Forest Grower* 13: 8

Hurley B, Slippers B, and Wingfield J 2007 A comparison of the control results for the alien invasive woodwasp, *Sirex noctilio*, in the southern hemisphere. *Agricultural and Forest Entomology* 9: 159–71

Hurley B P, Slippers B, Croft P K, Hatting H J, van der Linde M, Morris A R, Dyer C, and Wingfield M J 2008 Factors influencing parasitism of *Sirex noctilio* (Hymenoptera : Siricidae) by the nematode *Deladenus siricidicola* (Nematoda : Neotylenchidae) in summer rainfall areas of South Africa. *Biological Control* 45: 450–9

Ismail R, Mutanga O, and Bob U 2007 Forest health and vitality: The detection and monitoring of *Pinus patula* trees infected by *Sirex noctilio* using digital multispectral imagery (DMSI). *Southern Hemisphere Forestry Journal* 69: 39–47

Ismail R, Mutanga O, Kumar L, and Bob U 2008 Determining the optimal resolution of remotely sensed data for the detection of *Sirex noctilio* infestations in Pinus patula plantations in KwaZulu-Natal, South Africa. *South African Geographical Journal* 90: 196–204

Jiang H, Deng Y, Chen H, Tao L, Sha Q, Chen J, Tsai C, and Zhang S 2004 Joint analysis of two microarray gene-expression datasets to select lung adenocarcinoma marker genes. *BMC Bioinformatics* 5: 81

Kelly M and Meentemeyer R K 2002 Landscape dynamics of the spread of sudden Oak death. *Photogrammetric Engineering and Remote Sensing* 68: 1001–9

Kelly M, Guo Q, Liu D, and Shaari D 2007 Modeling the risk for a new invasive forest disease in the United States: An evaluation of five environmental niche models. *Computers, Environment and Urban Systems* 31: 689–710

Lawrence R L, Wood S D, and Sheley R L 2006 Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (RandomForests). *Remote Sensing of Environment* 100: 356–62

Leng W, He S H, Bu R, Dai L, Hu Y, and Wang X 2007 Predicting the distribution of suitable habitat for three larch species under climate warming in Northern China. *Forest Ecology and Management* 254: 420–8

Liaw A and Wiener M 2002 Classification and regression by randomForest. *R News* 2/3: 18–22

Lillesand T, Kiefer R, and Chipman J 2004 *Remote Sensing and Image Interpretation*. New York, John Wiley and Sons

Madden J L 1988 Sirex in Australasia. In Berryman A A (ed) *Dynamics of Forest Insect Populations*. New York, Plenum Press: 407–29

Negron J F 1998 Probability of infestation and extent of mortality associated with Douglas-fir beetle in the Colorado Front Range. *Forest Ecology and Management* 107: 71–85

Neumann F G and Minko G 1981 The Sirex woodwasp in Australian radiata pine plantations. *Australian Forestry* 44: 46–63

Pal M 2005 Random forest classifier for remote sensing classification. *International Journal of Remote Sensing* 26: 217–22

Peters J, De Baets B, Verhoest N, Samson R, Degroeve S, De Becker P, and Huybrechts W 2007 Random forest as a tool for ecohydrological distribution modelling. *Ecological Modelling* 27: 304–18

Prasad A, Iverson L, and Liaw A 2006 Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems* 9: 181–99

Prinzie A and Van den Poel D 2008 Random Forests for multiclass classification: Random MultiNomial Logit. *Expert Systems with Applications* 34: 1721–32

R Development Core Team 2008 *R: A Language and Environment for Statistical Computing*. Vienna, R Foundation for Statistical Computing

Reunanen J 2003 Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research* 3: 1371–82

van Rijisbergen C J 1979 *Information Retrieval*. London, Butterworths

Rosso P H and Hansen M E 2003 Predicting Swiss needle cast disease distribution and severity in young Douglas-Fir plantations in coastal Oregon. *Epidemiology* 93: 790–98

Schulze R E, Maharaj M, Lynch S D, Howe B J, and Melvil-Thomson B 1997 *South African Atlas of Agrohydrology and Climatology*. Pretoria, Water Research Commission Report No. TT82/96

Skidmore A 1999 Accuracy assessment of spatial information. In Stein A, van der Meer F, and Gorte B (eds) *Spatial Statistics for Remote Sensing*. Amsterdam, Kluwer: 197–209

Slippers B 2006 The sirex epidemic in KwaZulu-Natal and its control. *Wood and Timber Times Southern Africa* 31: 24–5

Spradberry J P and Kirk A 1978 Aspects of the ecology of siricid woodwasps (Hymenoptera : Siricidae) in Europe, North Africa and Turkey with special reference to the biological control of *Sirex noctilio* F. *Australia. Bulletin of Entomological Resources* 68: 341–59

van Staden V, Erasmus B F N, Roux J, Wingfield M J, and van Jaarsveld A S 2004 Modelling the spatial distribution of two important South African plantation forestry pathogens. *Forest Ecology and Management* 187: 61–73

Strobl C, Boulesteix A, Zeileis A, and Hothorn T 2007 Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* 8: 25

Svetnik V, Liaw A, Tong C, Culberson J, Sheridan R, and Feuston B 2003 Random Forest: A classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Science* 43: 1947–58

Taylor K L 1981 The sirex woodwasp: Ecology and control of an introduced forest insect. In Kitching R L and Jones R E (eds) *The Ecology of Pests: Some Australian Case Histories*. Melbourne, Vic, CSIRO: 231–48

Tribe G D and Cillie J J 2004 The spread of *Sirex noctilio* Fabricius (Hymenoptera: Siricidae) in South African pine plantations and the introduction and establishment of its biological control agents. *African Entomology* 12: 9–17

Vayssieres M P, Plant R E, and Allen-Diaz B H 2000 Classification trees: An alternative nonparametric approach for predicting species distributions. *Journal of Vegetation Science* 11: 679–94