Reducing the babel in plant volatile communication: using the forest to see the trees

Y. Ranganathan & R. M. Borges
Centre for Ecological Sciences, Indian Institute of Science, Bangalore, India

Abstract

While plants of a single species emit a diversity of volatile organic compounds (VOCs) to attract or repel interacting organisms, these specific messages may be lost in the midst of the hundreds of VOCs produced by sympatric plants of different species, many of which may have no signal content. Receivers must be able to reduce the babel or noise in these VOCs in order to correctly identify the message. For chemical ecologists faced with vast amounts of data on volatile signatures of plants in different ecological contexts, it is imperative to employ accurate methods of classifying messages, so that suitable bioassays may then be designed to understand message content. We demonstrate the utility of 'Random Forests' (RF), a machine-learning algorithm, for the task of classifying volatile signatures and choosing the minimum set of volatiles for accurate discrimination, using data from sympatric Ficus species as a case study. We demonstrate the advantages of RF over conventional classification methods such as principal component analysis (PCA), as well as data-mining algorithms such as support vector machines (SVM), diagonal linear discriminant analysis (DLDA) and k-nearest neighbour (KNN) analysis. We show why a tree-building method such as RF, which is increasingly being used by the bioinformatics, food technology and medical community, is particularly advantageous for the study of plant communication using volatiles, dealing, as it must, with abundant noise.
nant analysis (DA), or MANOVA, e.g. artificial neural network analysis (Cajka et al. 2009). In this paper, we demonstrate the use of Random Forests (RF), a machine-learning algorithm belonging to the class of data-mining techniques, in the analysis of VOC data. This new technique is being increasingly used in data-rich fields such as bioinformatics, chemoinformatics, medical diagnostics, food technology, astronomy and speech analysis, to select the most appropriate candidate variables from the surrounding data babel (Svetnik et al. 2003; Cannon et al. 2006; Díaz-Uriarte & de Andrés 2006; Granitto et al. 2007a,b; Zhang et al. 2008; Gao et al. 2009; Rong et al. 2009). We use data from Borges et al. (2008) on VOCs emitted by sympatric Ficus syconia in seed dispersal phase as an illustrative example. With this dataset, we also compare the performance of RF to other recent data-mining algorithms and tree-building methods, such as support vector machines (SVM), diagonal linear discriminant analysis (DLDA) and k-nearest neighbour analysis (KNN).

METHODS
Dataset

We used data published in Borges et al. (2008) on the volatile emissions produced by the syconia of sympatric Ficus hispida, F. exasperata and F. tsjahela at the seed dispersal stage. The headspace samples were collected from Agumbe Reserve Forest in the Indian Western Ghats (details of study site, volatile collection, composition and analysis are available in Borges et al. 2008). The dataset consisted of 49 samples and 77 VOCs. F. hispida and F. exasperata are dioecious species, in which only syconia on female trees produce seeds while those on male trees breed the pollinating mutualistic wasps. F. tsjahela is a monoecious species, in which all syconia produce seeds and wasps. Since seed dispersers should be attracted only to seed-bearing female syconia in the dioecious species because the male-bearing syconia contain developing and eclosing wasps, we predicted that the VOC signature of male and female ripened syconia should be distinct and different from each other. However, since in F. exasperata male and female ripened syconia are not usually available simultaneously owing to asynchrony between the sexes, we predicted that the VOC signature of male and female ripened syconia in this species could overlap, while they would be distinctive in the synchronous F. hispida. We found these predictions to be true (Borges et al. 2008). Using this dataset as a test case, we now address the following classification problem. We assume that in a forest where ripened syconia of these three Ficus species are available simultaneously, the problem to be solved by a biological entity, such as a seed disperser of Ficus, is to pick out the volatile signature of the ripened syconia of F. tsjahela, female F. hispida or female F. exasperata from other signatures in this data universe using the entire set of 77 VOCs recovered from these samples. Therefore, in this case, we are attempting to differentiate signal from noise within VOC blends produced by the five groups (male/female F. hispida, male/female F. exasperata and F. tsjahela), and also differentiate signal from signal (a unique VOC set to identify a group from the background) for each group. Moreover, since F. hispida is largely bat-dispersed and the other two species are largely bird-dispersed (Borges et al. 2008), we expect that the signatures of the seed-bearing syconia for each group should be distinguishable from the background or ‘the rest’ in this case in order to attract specific dispersal agents. We examine the efficacy of Random Forests (RF) in solving this classification problem. We emphasise that this classification problem was not attempted in Borges et al. (2008), in which clusters of the seed dispersal data were visualised with the help of PCA using only 33 out of the 77 VOCs (i.e. those VOCs with >5% occurrence in an individual sample).

VOC data and Random Forests

VOC data usually consist of samples of volatile emissions collected over several plant individuals or from the same plant over several time points. Whatever the type of sample or its method of collection, the key feature of a VOC dataset is that it is analogous to a microarray gene expression dataset in the sense that there are many more variables than samples. A typical headspace sample would include between 50 and 100 VOCs, just as a typical microarray would yield expression levels for hundreds of genes. This nature of a VOC dataset limited our use of the entire VOC dataset in earlier PCA analyses (Borges et al. 2008) and also limits the usage of classical multivariate analysis methods such as MANOVA or LDA. Conventional multivariate analysis methods require sample sizes to be proportionately increased for each added variable and also assume normality of the dataset, as well as the absence of auto-correlation between variables, besides having other limitations (Stevens 1992). Random Forests (Breiman 2001) is a classification algorithm with the following features that make it best suited for volatile analyses: (i) it allows for more variables than samples; (ii) it has a good classification efficiency, even with a lot of background noise; (iii) it is capable of arriving at a minimal set of variables, which can be used as predictors of that particular group; (iv) it is robust to interactions and correlations among variables; (v) it gives measures of relative variable importance (this objective and identifying the minimum variable set are more efficiently achieved using the varSelRF modifications to RF provided by Díaz-Uriarte & de Andrés 2006); and (vi) it can also be used to analyse time series data that record patterns in volatile emissions over time (achieved by the ‘dyn’ package available at http://cran.r-project.org). RF builds sets of decision trees using bootstrapping from the set of samples, and also selects a variable set of attributes (different VOCs in this case) at each node of the many decision trees thus generated. In this way, RF is also different from other tree-building methods such as PAUP: phylogenetic analysis using parsimony (Pergiguro-Alonso et al. 2008). RF also uses the unselected samples in a given bootstrap iteration to calculate an out-of-bag (OOB) error; i.e. the classification error obtained when the OOB (unselected samples) samples are examined; approximately one-third of the samples are unselected in each iteration. A major advantage of RF is that it does not overfit the data (Breiman 2001; Granitto et al. 2007a,b) so that even if minor fluctuations in variable strength (VOC concentrations or proportions in this case) lead to the building of thousands of classification trees, these fluctuations are not given undue importance in the final model; thus only the minimum set of important predictor variables is obtained.
Classification using RF

Between group classification
In this case, we retained the identity of the groups (e.g. female F. hispida or male F. exasperata) and attempted to find the prediction error of group membership.

One versus the rest classification
In this case, we determined the characteristic set of variables (VOCs) that define a particular group (e.g. female F. hispida) from all other samples, and for this purpose we masked the identity of the other groups (the rest). We attempted to find the smallest number of variables (VOCs) with which each group can be distinguished from the rest. The package varSelRF was used with R software version 2.9.0 (R Development Core Team 2009) for this purpose. The variable selection was allowed to run for 100 iterations. Running the algorithm for 1000 iterations did not yield a different result when compared with 100 iterations (data not shown). We therefore opted to use the results obtained from 100 iterations alone, as these are computationally intensive algorithms. For all analyses presented in this paper, only the proportions of the different VOCs present in the samples were used; this was purely for illustrative purposes; RF can use proportions or actual concentrations of each variable. A coefficient of variation (CV) was calculated for the predictor variables in each of the groups to determine whether RF consistently picks VOC predictors with low variability. An average out-of-bag (OOB) probability of membership in the groups (in the ‘one’ versus ‘the rest’ classification) was also estimated. This is the probability (calculated a` posteriori) of a sample belonging to that group. We also obtained prediction errors for group membership for the bootstrapping procedures as well as the mean decrease in accuracy (MDA) when particular variables (VOCs) are removed from the model. The MDA provides an importance score for that variable. The ‘importance’ function of the package randomForest for R was used to calculate MDA.

Comparative efficiency of Random Forests
To compare the relative performance of RF with other classification methods being currently evaluated in the literature for use in data-rich fields, we used the .632+ bootstrap method with 200 iterations for all methods. This bootstrap method is an improvement on the ‘leave-one-out’ cross-validation method and gives a better estimate of prediction error (Efron & Tibshirani 1997; Díaz-Uriarte & de Andrés 2006). We compared the prediction errors of support vector machines (SVM), diagonal linear discriminant analysis (DLDA), k-nearest neighbour (KNN) to RF. SVM is a data-mining algorithm that uses the concept of data kernels and support vectors that maximise the distance between parallel supporting planes between kernels of data (Bennett & Campbell 2000). DLDA is a type of linear discriminant analysis that uses maximum likelihood and diagonal covariance matrices (Dudoit et al. 2002). KNN is a classification method that classifies each sample based on minimum distance to k nearest neighbours of the sample (Hastie et al. 2001). The ‘errorest’ function of the package ipred for R was used for SVM, the ‘geSignatureBoot’ function of the package geSigna-

RESULTS
The OOB membership probability plots showed that Ficus tsjahela, as well as male and female F. hispida, were clearly separated from the rest, while male and female F. exasperata extractions were not clearly classifiable from the rest (Fig. 1). In the case of female F. exasperata, one sample actually grouped with the rest, while in male F. exasperata many more samples grouped with the rest. The variable selection procedure indicated similar trends (Table 1). The model frequency in this Table indicates the percentage of times the same predictor volatiles appeared in the 100 iterations. Female F. hispida extractions were clearly classifiable from the rest by 2-amyl acetate and iso-amyl acetate. This combination of

Fig. 1. The average out-of-bag (OOB) probability of membership of samples in the different groups. Samples that have a distinct volatile profile have an OOB probability close to one. Samples in the group of interest are indicated by filled circles; samples constituting the rest are indicated by open circles.
predictor compounds appeared in 100% of the iterations. Male *F. hispida* extractions were also uniquely classifiable by indole and α-trans-bergamotene. This combination of predictive compounds also appeared in 100% of the iterations. *F. tsjahela* had α-pinene and camphene that differentiated this group from the rest at 100% model frequency. On the other hand, male *F. exasperata* especially was relatively poorly classifiable and there were many misclassified samples with low model frequency (31%) (Table 1). This result for *F. exasperata* was as expected (see Discussion). Based on classifications achieved by RF, a biological entity (such as a seed disperser) could have less difficulty in finding female figs compared to finding male figs. This is an important and biologically relevant result (see Discussion). Additionally, the coefficient of variation (CV) was also lower for those predictor compounds whose contribution to the volatile signature of each sample was high (Table 1). The RF procedure also indicated the mean decrease in accuracy (MDA) in classification when the principal predictor compounds, as well as others, are removed from the model (Table 2). These results demonstrate the relative importance of certain VOCs in defining the volatile signature of the group.

RF was extremely good at classifying samples with the VOC data. For instance, out of the 77 seed dispersal VOCs found in the five groups examined, only 11 compounds (14% of the complete repertoire of compounds) (Table 1) were sufficient to classify all of the 49 samples with 5.84% error (Table 3). This translates to approximately three samples being misclassified. RF clearly scored over the other tree-building and classification methods examined (Table 3). For example, the prediction error for RF was 5.8% compared to 34.6% for SVM, 22.1% for DLDA and 14.9% for KNN in the between group comparison (Table 3). The same lower prediction error for RF was generally found in the comparison of one versus the rest for all the other groups (Table 3). It must be noted that all classification methods evaluated in this paper performed better than when no information was provided other than representation of the different groups in the total number of samples (no information category in Table 3).

### DISCUSSION

While there are several methods available for multivariate analysis and clustering, RF is ideally suited for the analysis of VOC data, as this case study of dispersal stage volatiles in sympatric *Ficus* has illustrated. RF enables the selection of candidate variables from a large dataset consisting of many more variables (77 VOCs) than samples (49 belonging to five groups) and also provides model statistics for these variables. Such variables can then form the basis of biological assays under controlled conditions. Information such as the mean decrease in accuracy when particular VOCs are removed from the model can also be used to determine candidate VOCs that are important in the particular VOC message that is intended for communication. It must be emphasised that we have merely used our dataset as an illustrative example of RF. Moreover, we attempted this classification exercise in the absence of knowledge of the overall background volatile landscape (i.e. volatiles produced by other plant species at the same site). As we discuss later, this is an important limitation of the dataset and not of the classification method.

Since plants and animals are faced with a complex volatile landscape, the characterisation of the statistical structure of this environment is an important first step towards understanding how volatile signals are encoded (Wright & Thomson 2005). Evaluating the statistics of visual scenes has correspondingly played an important role in understanding visual systems and how they function (Mackay 1986; Price 1987; Wright & Thomson 2005). In the case of VOCs too, understanding and characterising natural olfactory landscapes has been recently advocated, since many studies are now showing the effect of background compounds such as isoprene or various mixtures on olfactory perception and corresponding behavioural or physiological responses (Dicker et al. 2003; Mummi & Hilker 2005; Latthaworkitkul et al. 2008; Loivamäki et al. 2008). Conducting experiments on volatile communication in natural conditions using natural contexts is therefore being increasingly encouraged (Hunter 2002; Hales et al. 2009). However, moving from the controlled olfactory environments of the laboratory to complex natural olfactory landscapes will involve presenting and detecting signals where the signal-to-noise ratio is likely to be very low. Under this scenario, RF can be a powerful exploratory classification tool since it has been demonstrated to work extremely well with noise in a variety of noisy environments ranging from proteomics to astronomy (Gunther et al. 2003; Svetnik et al. 2003; Cannon et al. 2006; Díaz-Uriarte & de Andrés 2006; Granitto et al. 2007a,b; Kwak et al. 2008; Zhang et al. 2008; Gao et al. 2009).
Table 2. Mean decrease in accuracy (MDA) when particular volatiles are removed from the model for each ‘group’ versus ‘the rest’. Only non-zero values of MDA are reported.

<table>
<thead>
<tr>
<th>Male Ficus exasperata versus the rest</th>
<th>Female Ficus exasperata versus the rest</th>
<th>Male Ficus hispida versus the rest</th>
<th>Female Ficus hispida versus the rest</th>
<th>Ficus tsjehela versus the rest</th>
<th>MDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>daucene</td>
<td>0.056</td>
<td>γ-terpinene 0.023</td>
<td>indole 0.044</td>
<td>iso-amyl acetate 0.022</td>
<td>camphene 0.053</td>
</tr>
<tr>
<td>β-copaene</td>
<td>0.032</td>
<td>p-cymene 0.010</td>
<td>α-trans-bergamotene 0.040</td>
<td>2-amyl acetate 0.021</td>
<td>α-pinene 0.048</td>
</tr>
<tr>
<td>allo-aromadendrene</td>
<td>0.029</td>
<td>β-caryophyllene 0.010</td>
<td>isolepidozene 0.019</td>
<td>2-nonyl acetate 0.016</td>
<td>α-terpinene 0.028</td>
</tr>
<tr>
<td>undecane</td>
<td>0.026</td>
<td>terpinolene 0.007</td>
<td>2-heptyl acetate 0.015</td>
<td>n-amyl acetate 0.015</td>
<td>β-pinene 0.022</td>
</tr>
<tr>
<td>γ-terpinene</td>
<td>0.021</td>
<td>γ-muurolole 0.006</td>
<td>(E)-ocimene 0.014</td>
<td>2-nonanone 0.012</td>
<td>α-muurolole 0.019</td>
</tr>
<tr>
<td>γ-muurolole</td>
<td>0.016</td>
<td>tridecane 0.006</td>
<td>longifolene 0.013</td>
<td>3-octenyl acetate 0.012</td>
<td>α-selinene 0.015</td>
</tr>
<tr>
<td>(E)-β-farnesene</td>
<td>0.011</td>
<td>α-trans-bergamotene 0.003</td>
<td>2-heptanone 0.010</td>
<td>γ-terpinene 0.009</td>
<td>2-phenyl ethanol 0.013</td>
</tr>
<tr>
<td>(Z)-ocimene</td>
<td>0.010</td>
<td>(2)-3-hexenyl acetate 0.003</td>
<td>allo-aromadendrene 0.008</td>
<td>2-heptyl acetate 0.008</td>
<td>germacrene D 0.012</td>
</tr>
<tr>
<td>pentadecane</td>
<td>0.009</td>
<td>limonene 0.003</td>
<td>aromadendrene 0.007</td>
<td>α-pinene 0.007</td>
<td>δ-cadinene 0.007</td>
</tr>
<tr>
<td>β-elemene</td>
<td>0.006</td>
<td>tetradecane 0.003</td>
<td>δ-cadinene 0.007</td>
<td>(E)-ocimene 0.006</td>
<td>anisole 0.006</td>
</tr>
<tr>
<td>α-cis-bergamotene</td>
<td>0.006</td>
<td>isolepidozene 0.002</td>
<td>α-pinene 0.007</td>
<td>hexyl acetate 0.004</td>
<td>γ-terpinene 0.005</td>
</tr>
<tr>
<td>δ-cadinene</td>
<td>0.005</td>
<td>benzyl acetate 0.002</td>
<td>p-cymene 0.005</td>
<td>isolepidozene 0.003</td>
<td>alio-aromadendrene 0.005</td>
</tr>
<tr>
<td>sabinene</td>
<td>0.004</td>
<td>α-copaene 0.002</td>
<td>terpinolene 0.004</td>
<td>p-cymene 0.003</td>
<td>terpinolene 0.004</td>
</tr>
<tr>
<td>(E)-ocimene</td>
<td>0.004</td>
<td>(2)-ocimene 0.002</td>
<td>isoleadene 0.003</td>
<td>γ-muurolole 0.002</td>
<td>γ-gurjunene 0.004</td>
</tr>
<tr>
<td>Nl 2</td>
<td>0.004</td>
<td>1,8-cineole 0.002</td>
<td>γ-terpinene 0.003</td>
<td>α-cis-bergamotene 0.001</td>
<td>4-ethyl anisole 0.004</td>
</tr>
<tr>
<td>isoledene</td>
<td>0.004</td>
<td>α-Thuene 0.002</td>
<td>β-pinene 0.003</td>
<td>α-trans-bergamotene 0.001</td>
<td>α-thujene 0.004</td>
</tr>
<tr>
<td>p-cymene</td>
<td>0.004</td>
<td>alio-aromadendrene 0.002</td>
<td>iso-amyl acetate 0.002</td>
<td>limonene 0.001</td>
<td>undecane 0.003</td>
</tr>
<tr>
<td>β-pine</td>
<td>0.003</td>
<td>(E)-ocimene 0.002</td>
<td>daucene 0.002</td>
<td>myrcene 0.001</td>
<td>isolepidozene 0.002</td>
</tr>
<tr>
<td>myrcene</td>
<td>0.003</td>
<td>α-pinene 0.001</td>
<td>2-amyl acetate 0.002</td>
<td>2-heptanone 0.001</td>
<td>1,8-cineole 0.002</td>
</tr>
<tr>
<td>hexadecane</td>
<td>0.003</td>
<td>2-heptyl acetate 0.001</td>
<td>methyl anthranilate 0.002</td>
<td>β-pinene 0.001</td>
<td>β-cedrene 0.002</td>
</tr>
<tr>
<td>indole</td>
<td>0.003</td>
<td>methyl salicylate 0.001</td>
<td>α-thujene 0.002</td>
<td>α-trans-bergamotene 0.001</td>
<td>isoledene 0.002</td>
</tr>
<tr>
<td>β-santalene</td>
<td>0.003</td>
<td>indole 0.001</td>
<td>perillene 0.002</td>
<td>allo-aromadendrene 0.001</td>
<td>sabinene 0.001</td>
</tr>
<tr>
<td>β-caryophyllene</td>
<td>0.003</td>
<td>isoleadene 0.001</td>
<td>germacrene D 0.002</td>
<td>bicyclogermae 0.001</td>
<td>myrcene 0.001</td>
</tr>
<tr>
<td>α-pinene</td>
<td>0.003</td>
<td>α-humulene 0.001</td>
<td>germacrene D 0.002</td>
<td>germacrene D 0.001</td>
<td>sabinene 0.001</td>
</tr>
<tr>
<td>(2)-3-hexenyl acetate</td>
<td>0.002</td>
<td>pentadecane 0.001</td>
<td>undecane 0.001</td>
<td>methyl salicylate 0.001</td>
<td>sabinene 0.001</td>
</tr>
<tr>
<td>perillene</td>
<td>0.002</td>
<td>hexyl acetate 0.001</td>
<td>2 octyl acetate 0.001</td>
<td>sabine 0.001</td>
<td>Nl 1 0.001</td>
</tr>
<tr>
<td>camphene</td>
<td>0.002</td>
<td>2 heptyl acetate 0.001</td>
<td>2-nonanone 0.001</td>
<td>(Z)-3-hexenyl acetate 0.001</td>
<td>(Z)-ocimene 0.001</td>
</tr>
<tr>
<td>tetradecane</td>
<td>0.002</td>
<td>(E)-β-farnesene 0.001</td>
<td>2-nonanone 0.001</td>
<td>(Z)-3-hexenyl acetate 0.001</td>
<td>α-trans-bergamotene 0.001</td>
</tr>
<tr>
<td>α-terpinene</td>
<td>0.001</td>
<td>sabinene 0.001</td>
<td>α-cis-bergamotene 0.001</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NI = not identified.
Random Forest-based volatile selection

Table 3. Error rates estimated for the different classification methods using the .632+ bootstrap method with 200 bootstrap samples. The ‘no information’ column denotes the error rate at random when information from the groups is not used (Díaz-Uriarte & de Andrés 2006).

<table>
<thead>
<tr>
<th></th>
<th>no information</th>
<th>SVM</th>
<th>DLDA</th>
<th>KNN</th>
<th>random forest</th>
</tr>
</thead>
<tbody>
<tr>
<td>between groups</td>
<td>0.6530</td>
<td>0.3462</td>
<td>0.2208</td>
<td>0.1496</td>
<td>0.0584</td>
</tr>
<tr>
<td>female Ficus hispida versus the rest</td>
<td>0.1224</td>
<td>0.0590</td>
<td>0.0011</td>
<td>0.0018</td>
<td>0.0177</td>
</tr>
<tr>
<td>male Ficus hispida versus the rest</td>
<td>0.1837</td>
<td>0.1267</td>
<td>NA</td>
<td>0.0161</td>
<td>0.0054</td>
</tr>
<tr>
<td>Ficus tsjahela versus the rest</td>
<td>0.2245</td>
<td>0.0984</td>
<td>0.0011</td>
<td>0.0010</td>
<td>0.0107</td>
</tr>
<tr>
<td>female Ficus exasperata versus the rest</td>
<td>0.1224</td>
<td>0.0878</td>
<td>NA</td>
<td>0.1341</td>
<td>0.0589</td>
</tr>
<tr>
<td>male Ficus exasperata versus the rest</td>
<td>0.3469</td>
<td>0.2256</td>
<td>0.1339</td>
<td>0.1255</td>
<td>0.0855</td>
</tr>
</tbody>
</table>

NA = not available.

2009; Rong et al. 2009). In the context of noise, it is also significant that the compounds picked as predictors by RF in our case study were mostly those that had very low CVs (Table 1). This is biologically relevant since a VOC that is to be used to constitute a signal with reliable information content should not vary greatly in its representation in the volatile signature. Low CVs of biologically relevant VOCs have been used to constitute a signal with reliable information content. The identification of predictor variables by RF can also be used to make predictions about samples outside the current dataset, a situation analogous to using a set of predictor genes used to make predictions about samples outside the current population (synchrony between sexes), while in F. exasperata there is low overlap between syconia production in the sexes (asynchrony between sexes) (see Borges et al. 2008). Therefore, while there should be selection on female figs to have a VOC profile different from male figs or ‘the rest’, such that only seed figs are consumed so that seeds are dispersed in dioecious species that produce male and female figs simultaneously in the same season, i.e. F. hispida in this case. Thus, there should be less selection pressure on F. exasperata than on F. hispida in the dispersal phase to make the male fig volatile signature different from that of seed figs or ‘the rest’, and this why several male F. exasperata figs are misclassified with ‘the rest’. Furthermore, the ability of RF to separate female signatures from ‘the rest’ is biologically relevant since seed-bearing female figs, rather than wasp-containing male figs, should be consumed.

Another advantage of RF in chemical ecology is that it can deal not only with categorical variables (within a classification framework), such as volatile type, chirality, or any other fea-
nature of stereochemistry that one may want to include, but also with continuous variables, such as concentrations of volatiles or their ratios (in a regression framework) or a combination of both types of variables. Furthermore, as in all good model building, RF allows portions of the data to be used as training sets so that the tree-building algorithm can be refined (Breiman 2001). Such flexibility and feature diversity facilitate comprehensive exploration of volatile landscapes where complex algorithms may be required to determine how olfactory systems find their targets (Bruce et al. 2005; Pareja et al. 2009). Such algorithms can also deal with complex datasets that show great intraspecific variability in VOC emissions (e.g. Degen et al. 2004). From a biologically relevant perspective, using RF to find a limited set of predictor variables from a universe of 77 VOCs in our test dataset can provide a practicable set of compounds to be employed in biological assays with model seed dispersal agents, such as bats or birds, that are the natural dispersers of these Ficus species.

Despite the fact that RF has come into recent use in ecology, including forestry, parasitology and migratory movements (Prasad et al. 2006; Cutler et al. 2007; Iverson et al. 2008; Perdigüero-Alonso et al. 2008; Oppel et al. 2009), it has not yet been employed in understanding the chemical ecology of communication using volatiles in plants. Its versatility can be a great boon in this field and needs further examination. RF is itself in the process of being constantly evaluated (e.g. Amarutunga et al. 2008). We suggest that RF could be a valuable addition to the chemical ecologist’s tool kit, a view that should, however, always be tempered by Wolpert’s ‘no free lunch theorem’; i.e. there is no one algorithm that can be universally suitable for all classification problems (Svetnik et al. 2003).

ACKNOWLEDGEMENTS

This research was funded by the Ministry of Environment and Forests, Government of India. We thank Mahua Ghara and Anusha Krishnan for useful discussions and the reviewers for useful comments.

REFERENCES


