Focus Article

Multivariate random forests

Mark Segal* and Yuanyuan Xiao

Random forests have emerged as a versatile and highly accurate classification and regression methodology, requiring little tuning and providing interpretable outputs. Here, we briefly outline the genesis of, and motivation for, the random forest paradigm as an outgrowth from earlier tree-structured techniques. We elaborate on aspects of prediction error and attendant tuning parameter issues. However, our emphasis is on extending the random forest schema to the multiple response setting. We provide a simple illustrative example from ecology that showcases the improved fit and enhanced interpretation afforded by the random forest framework. *WIREs Data Mining Knowl Discov* 2011 1 80–87
DOI: 10.1002/widm.12

INTRODUCTION

Since the mid-1980s, tree-structured (or recursive partitioning) classification and regression methods have enjoyed widespread popularity. This followed the publication of the Classification and Regression Trees (*CART*) monograph[1] that established a rigorous framework for such techniques, and convincingly illustrated one of their greatest virtues: interpretability. Tree-structured methods (TSM) produce interpretable prediction rules by subdividing data into subgroups that are homogenous with respect to both predictors and response. For continuous responses, as considered here, simple (terminal) subgroup summaries (typically means) serve as predictions. The interpretability of the attendant prediction rules derives from (1) the natural, recursive fashion by which predictors are employed in eliciting subgroups, (2) the accessibility of companion tree diagram schematics, and (3) the availability of predictor importance summaries. However, by the mid/late-1990s a serious deficiency of TSM was evident: modest predictive performance, especially in comparison with emerging, flexible competitors such as support vector machines (SVM).[6] In a series of papers, Breiman developed a strategy for remedying this shortcoming: create an ensemble of trees, where each tree in the ensemble is grown in accordance with the realization of a random vector and obtain predictions by aggregating (voting) over the ensemble. Bagging[2] represents an early example whereby each tree is constructed from a bootstrap[10] sample drawn with replacement from the training data. The simple mechanism whereby bagging reduces prediction error for unstable predictors, such as trees, is well understood in terms of variance reduction resulting from averaging.[3] Such variance gains can be enhanced by reducing the correlation between the quantities being averaged. It is this principle that motivates random forests (RF).

Random forests seek to effect such correlation reduction by a further injection of randomness. Instead of determining the optimal subdivision of a given subgroup of a (constituent) tree by evaluating all allowable partitions on all predictors, as is done with single-tree methods or bagging, a subset of the predictors drawn at random, is employed. The size of this subset, designated `mtry`, is the primary tuning parameter for the forest procedure. Breiman[4] argues, on the basis of a comprehensive empiric evaluation employing numerous benchmark datasets excerpted from the University of California at Irvine (UCI) repository, that RF enjoy exceptional prediction accuracy, and that this accuracy is attained for a wide range of settings of the key tuning parameter `mtry`. Here, we provide a very brief overview of RF, there now being excellent accounts in the literature.[11] Rather, we detail (1) the extension of regression trees to multivariate response settings, and the related generalization of RF, (2) an underappreciated aspect of RF pertaining to overfitting and tuning parameters, and (3) an illustrative example of multivariate RF (MRF). Throughout, we concentrate on regression, rather than classification problems.

MULTIVARIATE REGRESSION TREES

The regression tree framework, as developed by Breiman *et al.*[1] involves four components: (1) A set of

*Correspondence to: mark@biostat.ucsf.edu
Department of Epidemiology and Biostatistics, UCSF
DOI: 10.1002/widm.12

binary (yes/no) questions, or splits, phrased in terms of the predictors that serve to partition the predictor space. The subsamples created by assigning cases according to these splits are termed nodes. A node that does not have any descendant nodes is a terminal node or leaf. (2) A node impurity measure, typically relating to response variance in the regression context. (3) A split function, $\phi(s, t)$, that can be evaluated for each allowable split s, of each node t. The best split, which optimizes ϕ, is such that the response distributions in the resultant children nodes are most homogenous among all competing splits, with homogeneity assessed via the impurity measure. (4) A means for determining the appropriate tree size.

In the single (univariate) response setting, let y_i and x_{ij} $(i = 1, \ldots, n;\ \ j = 1, \ldots, p)$ designate response and predictors respectively. Consider a node t containing a sub-sample of cases. We aim to partition t into two child nodes, a 'left' node t_L, and a 'right' node t_R. Let j be the index of a continuous or ordered categorical predictor. Then (default) allowable splits are order-preserving binary cuts of the form $t_L = i \in t : x_{ij} \leq c, t_R = i \in t : x_{ij} > c$ as the cut-point c ranges over all possible values resulting in distinct t_L, t_R. For unordered categorical predictors all splits into disjoint subsets of the categories are allowed. The L_2 node impurity measure is just the sum-of-squares $SS(t) = \sum_{i \in t}(y_i - \mu(t))^2$ where $\mu(t)$ is the mean of y_i in node t. Then the corresponding split function is

$$\phi(s, t) = SS(t) - SS(t_L) - SS(t_R). \tag{1}$$

Now consider multiple response data y_{ik} $(i = 1, \ldots, n;\ \ k = 1, \ldots, m)$. The formulation includes time course and clustered outcomes. In view of anticipated dependencies between the responses, interpretative, and predictive gains can potentially be realized by analyzing all responses simultaneously. Examples illustrating attainment of such gains are provided elsewhere.[20] For simplicity, we assume that each individual has the same number of responses (m) and that the predictors are 'baseline' variables; i.e., do not vary with k. Segal[17] discusses means for removing these restrictions. All that is required to extend regression trees to multiple responses is modification of the split function. A natural formulation is to replace the node impurity measure with a 'covariance' weighted analog:

$$SS(t) = \sum_{i \in t}(y_i - \mu(t))' V^{-1}(t, \eta)(y_i - \mu(t)). \tag{2}$$

Here η represents parameters characterizing prescribed covariance structures (e.g., auto regressive, compound symmetry). Using (2) a multiresponse split function is created as per (1). The prediction for each leaf of a multiresponse regression tree is just the vector response means for cases reaching that leaf.

RANDOM FORESTS

An RF, as defined by Breiman,[5] is a collection of tree predictors $h(\mathbf{x}; \boldsymbol{\theta}_k), k = 1, \ldots, K$ where $\mathbf{x}$ represents the observed input (predictor) vector of length p with associated random vector $\mathbf{X}$ and the $\boldsymbol{\theta}_k$ are independent and identically distributed (iid) random vectors. The observed data is assumed to be independently drawn from the joint distribution of $(\mathbf{X}, Y)$ and comprises n $(p + 1)$-tuples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$. The RF prediction is the unweighted average over the collection: $\bar{h}(\mathbf{x}) = (1/K) \sum_{k=1}^{K} h(\mathbf{x}; \boldsymbol{\theta}_k)$.

As $k \to \infty$ the Law of Large Numbers ensures

$$E_{\mathbf{X},Y}(Y - \bar{h}(\mathbf{X}))^2 \to E_{\mathbf{X},Y}(Y - E_{\boldsymbol{\theta}} h(\mathbf{X}; \boldsymbol{\theta}))^2. \tag{3}$$

The quantity on the right is the prediction (or generalization) error for the RF, designated PE^*_f. The average prediction error for an individual tree $h(\mathbf{X}; \boldsymbol{\theta})$ is $PE^*_t = E_{\boldsymbol{\theta}} E_{\mathbf{X},Y}(Y - h(\mathbf{X}; \boldsymbol{\theta}))^2$. Assume that for all $\boldsymbol{\theta}$ the tree is unbiased, i.e., $EY = E_{\mathbf{X}} h(\mathbf{X}; \boldsymbol{\theta})$. Then

$$PE^*_f \leq \bar{\rho}\, PE^*_t \tag{4}$$

where $\bar{\rho}$ is the weighted correlation between residuals $Y - h(\mathbf{X}; \boldsymbol{\theta})$ and $Y - h(\mathbf{X}; \boldsymbol{\theta}')$ for independent $\boldsymbol{\theta}, \boldsymbol{\theta}'$.

The inequality (4) pinpoints what is required for accurate RF regression: (1) low correlation between residuals of differing tree members of the forest, and (2) low prediction error for the individual trees. Further, the RF will, in expectation, decrease the individual tree error, PE^*_t, by the factor $\bar{\rho}$. The strategy employed to achieve these objectives is: (1) to keep individual error low, grow trees to maximal depth, and (2) to keep residual correlation low, randomize via (a) growing each tree on a bootstrap[10] sample from the training data, and (b) prespecifying $mtry \ll p$ (the number of predictors) and, for each node of every tree, randomly select $mtry$ predictors and pick the best split of that node using only these predictors. Note that limiting step (2) to component (a) or, equivalently, using $mtry = p$ reduces the RF procedure to bagging.

The use of bootstrap resampling in (2)(a) provides for a built-in mechanism for unbiased estimation of PE without requiring recourse to cross validation. Simply, the $\approx 1/3$ of the cases omitted from each bootstrap sample serve as test data for the tree constructed using that sample. These cases, and the associated PE estimate, are termed out-of-bag (OOB).

We revisit issues surrounding the first part of the prescription: growing trees to maximal depth. This represents a notable departure from the strategy advocated under the original[1] (single) tree paradigm. There, while a large tree was initially grown, considerable effort was dedicated to iteratively pruning (upward collapsing) this tree. Subsequent selection of a final right-sized tree from the set of pruned subtrees made recourse to cross validation based assessment of predictive performance. This emphasis on determining tree size derived from the recognition that overly large trees, although unbiased, would incur a prediction variance cost resulting in degraded performance. In the RF context the hope is that by averaging over the (large) ensemble of trees this variance component is reduced. Indeed, the dominant performance of RF in comparative benchmarking studies would support such a notion. However, as we indicate next, the benchmark datasets employed featured idiosyncrasies that may mislead on this issue.

Now consider Figures 1(a) and (b), which display two very distinct cross validated *PE* profiles. The data used in Figure 1(a) comes from a study of RNA splice site identification, as detailed in Ref 19. The minimum *PE* is attained at about 100 splits. The minimum occurs in a plateau region, after which there is an appreciable rise in error. This increase is such that the *PE* at the maximal number of splits is 'significantly' greater than the minimum *PE*; the vertical segments (contained within each plotting symbol) represent ± 1 standard error. Such profiles where, as a function of increasing model size/complexity, *PE* initially decreases, plateaus, and then increases are common. Indeed, prototypic depictions of the relationship between *PE* and model complexity have this form; see, Ref 1 pp. 87 and Ref 11 pp. 38. The spider data, analyzed subsequently via multivariate trees and forests, also exhibits this pattern; see, Figure 2(a). The presence of noise and/or redundant predictors are factors that can contribute to such profiles, the impact of the latter being discussed in Ref 18.

Figure 1(b) differs in that the *PE* at the maximal number of splits is the global minimum. That is to say, no matter how large a (single) tree-structured predictor we fit, we don't overfit the data. This behavior is unusual. The (letter recognition) data used to generate Figure 1(b) were obtained from the UCI Repository of Machine Learning Databases as converted to R,[12] and available from the `mlbench` package. What is remarkable, and seemingly not appreciated, is that almost every dataset in the `mlbench` package exhibits this behavior. And, it was this compendium that was used in establishing the superior predictive performance of RF under the strategy of growing (individual) trees to maximal depth.

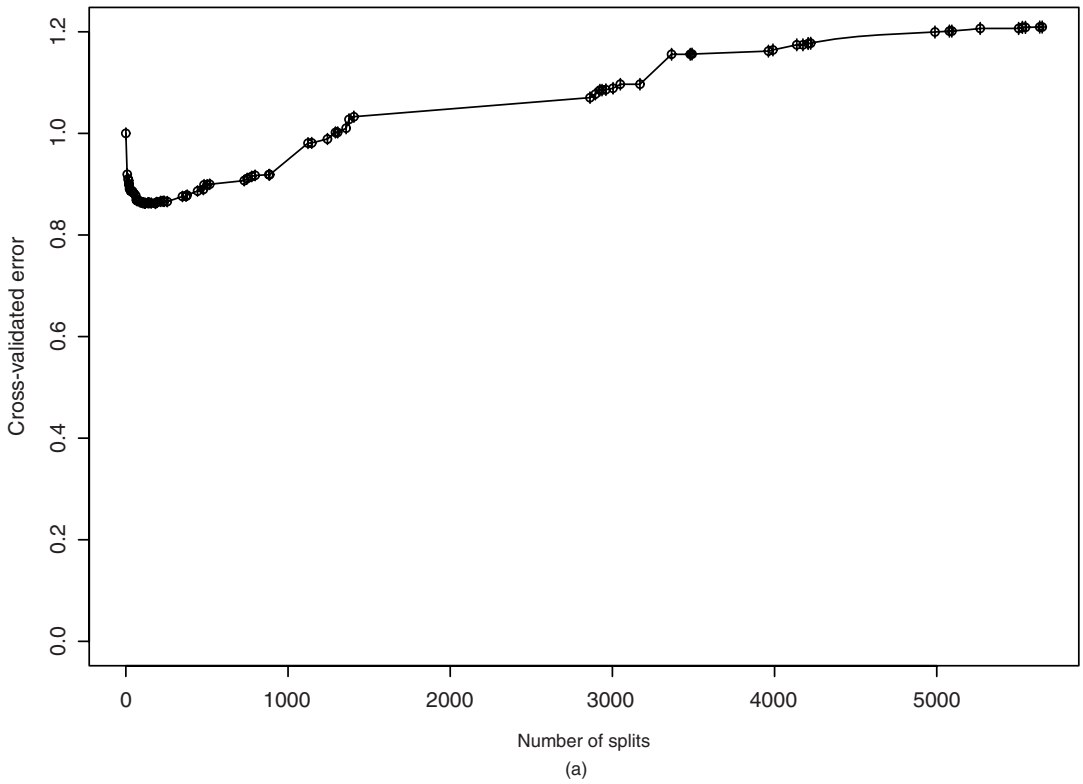
Our central concern, then, is that this strategy yields maximal trees which may be highly unstable. This instability will be reflected in inflated prediction errors, and the variance reduction achieved by averaging over the ensemble may not sufficiently counteract this inflation. Precisely, this behavior was observed for the splice site identification study. That this behavior was not observed in the empirical evaluations of RF using the UCI repository is potentially attributable to the above mentioned property of the repository constituents. The R package RF[15] includes a tuning parameter `maxnodes` that can be used to override growing maximal trees. Judiciously setting this parameter is anticipated to be beneficial in large sample size settings, and/or in situations where maximal trees severely overfit. Such control seems less awkward, especially for large sample sizes, than trying to govern tree size by the (related) parameter `nodesize` which determines the minimum size for terminal nodes: prescribing `maxnodes` more readily allows for penalization of tree complexity whereas not precluding variably sized nodes. Lin and Jeon[16] provide a theoretic perspective that is also contrary to use of maximal trees.

MULTIVARIATE RF

To construct MRF, that accommodate multivariate outcomes, we simply generate an ensemble of MRTs via bootstrap resampling and predictors subsampling as for univariate RF. Here we employ MRTs as described above, but note that other formulations have been advanced.[9,14,21] Under this straightforward extension, we inherit two byproducts of univariate RF useful for enhanced interpretation.

Proximity Matrix

The proximity matrix captures how cases relate to each other, and so provides the underpinnings for (supervised) clustering. For each tree in the ensemble, all data (training and OOB) are run down to their assigned terminal node, as dictated by the split sequence. If cases i and j are assigned to the same terminal node, then the proximity value, $pv_{i,j}$, between i and j is incremented by one. This process is repeated for each tree in the forest, with subsequent normalization by dividing by the number of trees. The proximity matrix is the $n \times n$ matrix of pv's, and is symmetric, positive definite, and bounded by 1. The



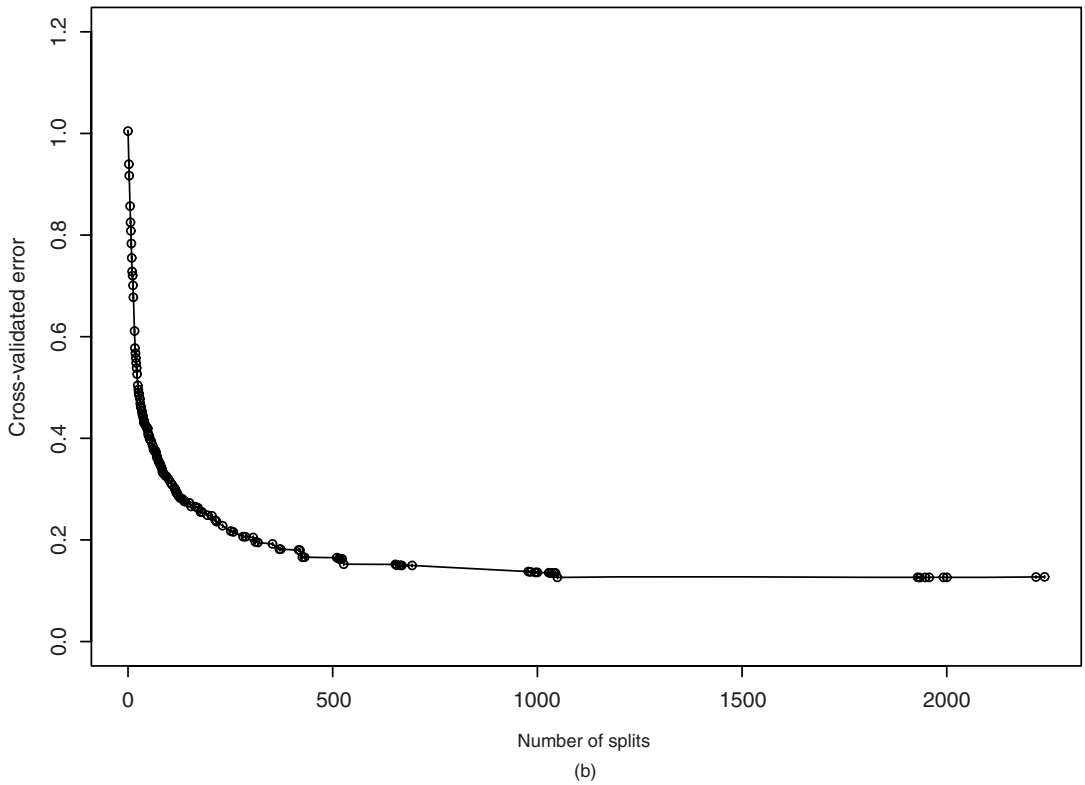


FIGURE 1 | Cross-validated prediction error profiles for trees grown to maximal size on the (a) splice site identification data and (b) letter recognition data from `mlbench`.

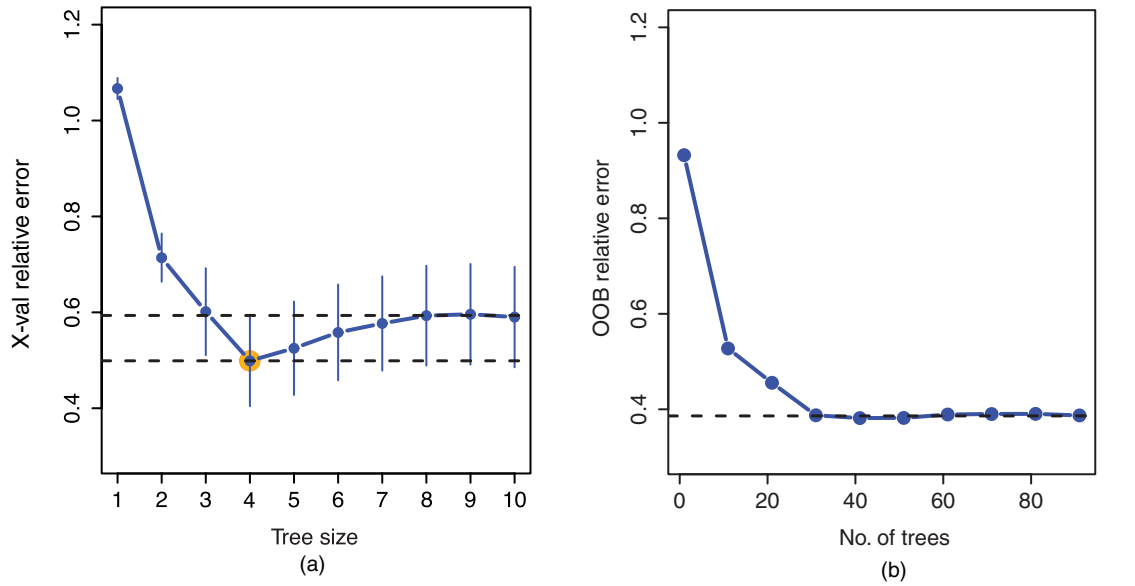


FIGURE 2 | *PE* profiles for (a) MRT and (b) MRF. (a) Cross-validated errors are plotted against tree size. The vertical bars represent *PE* standard error. The orange spot indicates final tree size as selected by the 1-SE rule or overall *PE* minimum. (b) OOB errors are plotted against number of trees in MRF. The dotted line is the minimum error rate.

values 1 minus proximities can be treated as (squared) distances.

Variable Importance

For each tree in the ensemble, we compute *PE* using OOB cases. We then *permute* each variable (one at a time) and again compute *PE*. The difference between the original and permuted *PEs*, averaged over all trees, provides a variable importance summary. This can be used to rank variables and identify those most influencing prediction.

EXAMPLE: CO-OCCURRENCE OF SPIDER SPECIES

We use a simple example pertaining to species abundance and environmental characteristics[7] to illustrate MRF. A much more detailed treatment, pertaining to the cell cycle, is given in Ref 20. Multiple responses are the abundances of 12 spider species, cases are 28 sites, and the predictors are six environmental factors potentially affecting spider habitat. De'ath[7] used MRT, and developed companion software,[8] to predict species co-occurrence based on the predictors. We define the impurity of a node as the sum of squared Euclidean distances of sites within the node to the node centroid; i.e., $V = I$ in Ref 2, as this was utilized by De'ath.[7] Typically, either the tree with the minimum cross validated *PE*, or the smallest tree within one standard error of this minimum (1-SE rule)[1] is selected. Once tree size is determined, the resulting terminal nodes can be interpreted as 'habitats',[7] which are composed of sites that have similar species compositions and environmental attributes.

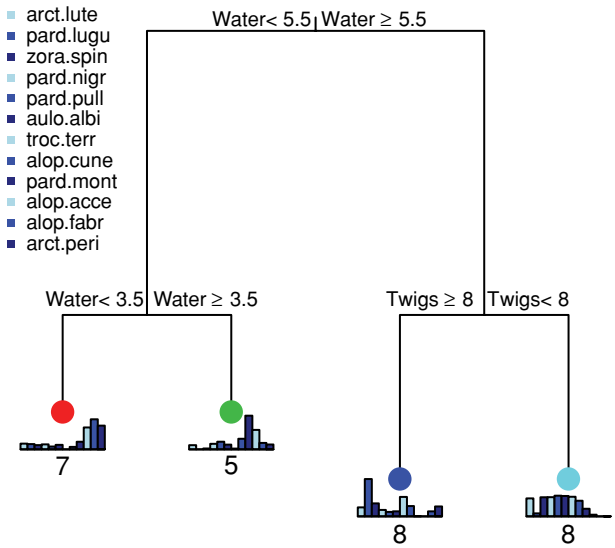


FIGURE 3 | MRT for the spider data. Terminal nodes are indicated by colored dots. Barplots show average abundances of the 12 species at each terminal node. The number of sites in each terminal nodes is given below the barplots.

The application of MRT analysis on the spider data, built using the R package `mvpart`, yields a

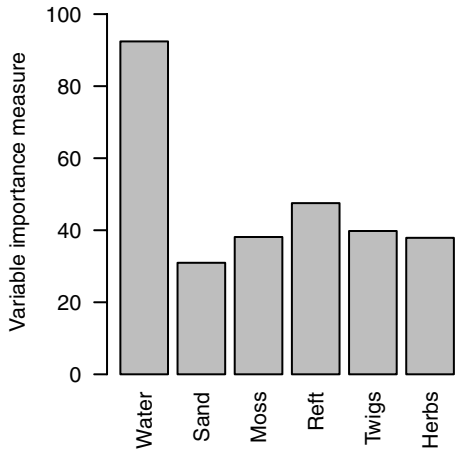


FIGURE 4 | MRF variable importance measures.

tree with four terminal nodes as determined by cross validation using either the 1-SE rule or global PE minimum (of 50%); see, Figure 2(a) where prediction error is plotted against tree size. This tree is defined by two environmental variables, `water` and `twig` (Figure 3). The barplot below each terminal node depicts the species composition.

The tree schematic in Figure 3 enables a ready description of habitats in terms of interplay of spider species composition and environmental features and thereby highlights the interpretability of (single) trees. However, considerable caution is needed in drawing such interpretations. This is on account of the inherent instability of tree-structured predictors, caused in part by the greedy optimization of the split function (1). Consequently, small data changes can produce dramatically different tree topologies and attendant interpretations. Fitting MRTs to bootstrap samples of the original data affirms this by yielding a range of tree sizes and associated split variables (not shown). As described, MRF seek to overcome this instability by aggregating over an ensemble of MRTs. Accordingly, we apply MRF to the spider data using an ensemble of 300 trees with parameters `nodesize` and `mtry` both set to 2. As shown in Figure 2 (B), MRF attain a PE of 38%, a substantial improvement over the 50% achieved by MRT.

Interpretation of MRFs is compromised by the inability to simultaneously view hundreds of tree schematics. However, by making recourse to the byproducts described above interpretability can be regained. Variable importance summaries for the six environmental predictors are presented in Figure 4. Perhaps not surprisingly, `water` has the highest importance value. The proximity matrix can be used for both clustering and identifying outliers. We apply the partition around medioid (PAM) clustering algorithm[13] to group sites into four homogenous groups based on the proximity matrix. To visualize the clustering of the sites, we employ metric multidimensional scaling to project the data onto a two dimensional space (Figure 5) using the first three principal coordinates which together explain 72% of the total variance. The groups are enclosed by convex hulls based on clustering membership by MRF and PAM. Sites 16 and 20, both of which have negative silhouette widths, cannot be clustered with confidence.

The derived habitats by MRF and PAM can be described in terms of their defining predictor

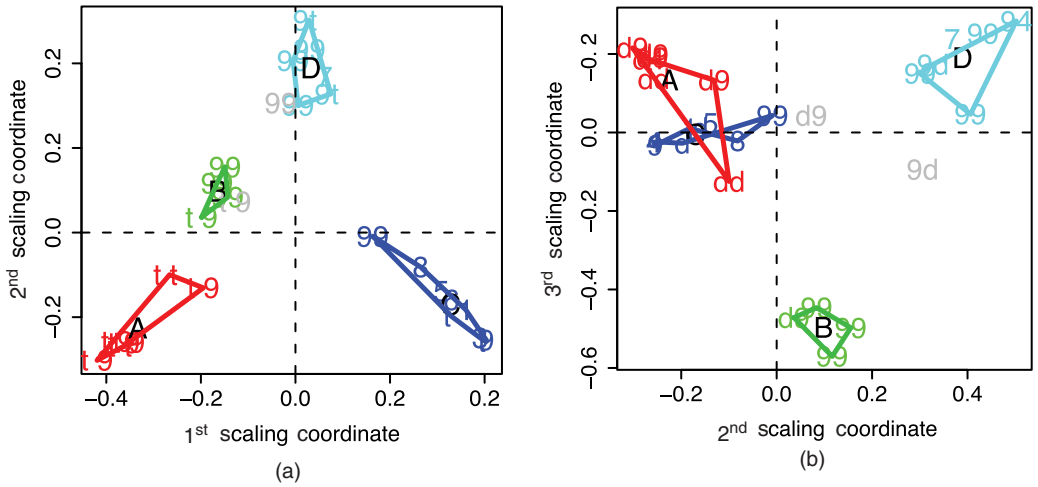


FIGURE 5 | Metric multidimensional scaling of the spider data based on the MRF proximity matrix. Sites are designated 1 - 28. Colors of sites and convex hulls indicate PAM cluster membership; the two sites in gray have negative silhouette widths, suggesting low clustering confidence. Letters, A-D, are located at the cluster means.

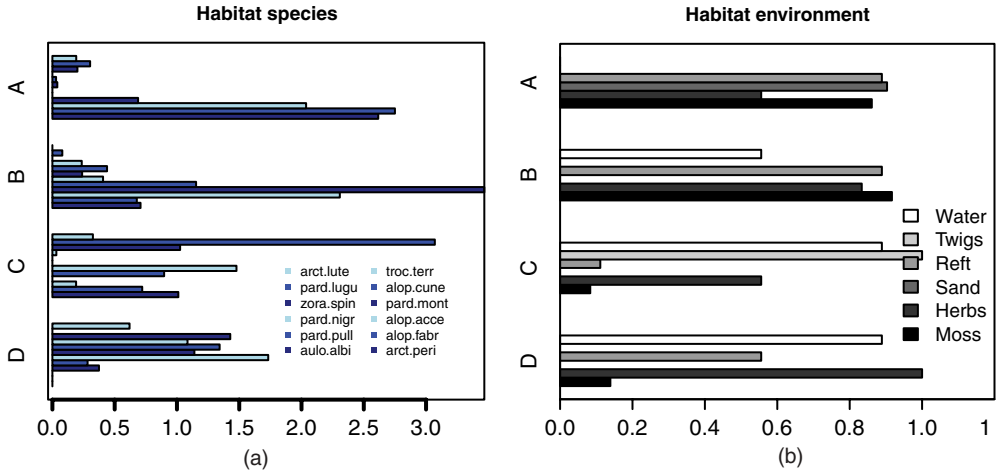


FIGURE 6 | (a) species and (b) environment characteristics of the four habitats.

(environmental factors) and outcome (species compositions) variables. We present corresponding barplots for both sets of variables at each habitat in Figure 6. Integrating information from both panels reveals a dynamic relationship between environment and species. Habitat A is characterized by a complete lack of water and twigs, but with a plentitude of reft, moss, sand and herbs, and is populated by three spider species: *Alopecosa accentuate*, *Alopecosa fabrilis*, and *Arctosa perita*. In contrast, habitat C has abundant water and twigs, but has a lack of moss, reft and sand, and it is dominated by the spider species *Pardosa lugubris*.

CONCLUSION

Random forests have emerged as a forefront classification and regression technique, enjoying exceptional accuracy, and enabling interpretative insight for wide classes of problem, all with minimal tuning. Here, we have illustrated facets of the RF formulation that can benefit from additional tuning: restricting the extent of individual trees in large sample or high noise settings. More importantly, we have demonstrated how readily RF can be generalized. The extension we detailed, to multiple outcomes, can be used to analyze clustered or longitudinal responses.

REFERENCES

1. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*, Belmont, CA: Wadsworth; 1984.
2. Breiman L. Bagging predictors. *Mach Learn* 1996, 24:123–140.
3. Breiman L. Arcing classifiers. *Ann Stat* 1998, 26:801–849.
4. Breiman L. Statistical modeling: the two cultures. *Stat Sci* 2001a, 16:199–215.
5. Breiman L. Random forests. *Mach Learn* 2001b, 45:5–32.
6. Cristianini N, Shawe-Taylor J. *An Introduction to Support Vector Machines*. Cambridge, Cambridge University Press; 2000.
7. Deáth G. Multivariate regression trees: a new technique for modeling species-environment relationships. *Ecology* 2002, 83:1005–1117.
8. Deáth G. mvpart: Multivariate partitioning. R package version 1.3-1. (2010).
9. Džeroski S, Ženko B. Stacking with multi-response model trees. In: *Multiple Classifier Systems, Proceedings of the Third International Workshop*, pp 201–211. Berlin: Springer; 2002.
10. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. New York: Chapman & Hall; 1993.
11. Hastie TJ, Tibshirani RJ, Friedman JH. *The Elements of Statistical Learning*. New York, NY: Springer; 2009.
12. Ihaka R, Gentleman R. R: A language for data analysis and graphics. *J Comput Graph Stat* 1996, 5:299–314.
13. Kaufman L, Rousseeuw PJ. *Finding groups in data: an introduction to cluster analysis*. New York, NY: John Wiley & Sons; 1990.
14. Kim S-J, Lee K. Constructing decision trees with multiple response variables. *International Journal of Management and Decision Making* 2003, 4:337–353.

15. Liaw A, Wiener M. Classification and regression by random Forest. *R News* 2002, 2:18–22.
16. Lin Y, Jeon Y. Random forests and adaptive nearest neighbors. *J Am Stat Assoc* 2006, 97:578–590.
17. Segal MR. Tree-structured methods for longitudinal data. *J Am Stat Assoc* 1992, 87:407–418.
18. Segal MR, Barbour JD, Grant RM. Relating HIV-1 sequence variation to replication capacity via trees and forests. *Stat Appl Genet Mol Biol* 2004, 3.
19. Segal MR. Prediction of RNA Splice Signals. In: Biswas A, Datta S, Fine J, Segal MR, eds. *Statistical Advances in the Biomedical Sciences*. New York: John Wiley & Sons; 2008, 443–463.
20. Xiao Y, Segal MR. Identification of yeast transcriptional regulation networks using multivariate random forests. *PLoS Comput Biol* 2009, 5:e1000414.
21. Zhang H. Classification trees for multiple binary responses. *J Am Stat Assoc* 1998, 93:180–193.

