New machine learning tools for predictive vegetation mapping after climate change: Bagging and Random Forest perform better than Regression Tree Analysis

L.R. Iverson[1], A.M. Prasad[1] and A. Liaw[2]

[1] USDA Forest Service, 359 Main Road, Delaware, OH 43015 USA
E-mail: liverson@fs.fed.us
[2]Merck Research Laboratories, Rahway, New Jersey USA

Abstract

More and better machine learning tools are becoming available for landscape ecologists to aid in understanding species-environment relationships and to map probable species occurrence now and potentially into the future. To that end, we evaluated three statistical models: Regression Tree Analysis (RTA), Bagging Trees (BT) and Random Forest (RF) for their utility in predicting the distributions of four tree species under current and future climate. RTA's single tree was the easiest to interpret but is less accurate compared to BT and RF which use multiple regression trees with resampling and resampling-randomisation respectively. Future estimates of suitable habitat following climate change were also improved with BT and RF, with a slight edge to RF because it better smoothes the outputs in a logical gradient fashion. We recommend widespread use of these tools for GIS-based vegetation mapping.

Introduction

The world's climate has always been undergoing change but now there are new reports almost every month linking recent changes in the climate to some biological trend. It has been estimated that the composition of one-third of the earth's forests could change markedly due to climate changes associated with a doubling of atmospheric CO_2 (e.g., Melillo, 1999). Plant species are expected to shift in range and relative abundance as the climate changes, and this has been the thrust of our research for the eastern half of the United States (e.g., Iverson & Prasad, 1998; 2001). For this paper, we tested three statistical prediction tools for modelling current and potential future suitable habitat under climate change, for four common tree species from eastern North America.

Statistical Models

Regression Tree Analysis (RTA)

RTA differs from classical statistical methods in that it constructs a set of decision rules via recursively partitioning on the predictor variables (Breiman *et al.,* 1984; Iverson and Prasad 1998). These rules allow for the possibility of interactions and non-linearities among variables (Moore *et al.,* 1991), and enable mapping of the predictors with the greatest influence on distributions geographically, which can provide insights into the spatial influence of the predictors (Iverson & Prasad, 1998). RTA predictions can, however, be unstable in that small changes in data can produce largely different models.

Bagging Trees (BT)

Bagging uses a regression tree technique as well but uses many (30-80) training data sets by resampling the data with replacement (on average 67% of cases appear in bootstrap sample), then averages the outputs, so that the variance component of the generalization error is reduced (Breiman 1996). The portion of the data drawn into the sample in a replication is known as the "in-bag" data while the portion not drawn is the "out-of-bag" data. The "out-of-bag" data is not used to build or prune any tree but used to give better estimates of node-error and other generalisation errors for bagged predictors (Breiman, 1996). The main disadvantage of bagging is that the large number of models makes it difficult to interpret the results, especially for species that have relatively unstable models. In contrast, for species with stable models, the interpretation of the original RTA tree may be suitable.

Random Forest (RF)

Random Forest is relatively new, but has been shown to produce very accurate predictions without overfitting models to the data (Breiman 2001). RF is essentially very similar to BT in that bootstrap samples are drawn to construct multiple trees. The differences, however, are that the each tree is grown with a randomised subset of predictors, i.e., the number of predictors (initially fixed) used to find the best split at each node is a randomly chosen subset of the total number of predictors, and that a very large number (500-2000) of trees are grown (hence 'forest' of trees). Like BT, the trees are grown to maximum size without pruning and aggregation takes place by averaging the trees. A main advantage of RF is that the output depends mainly on only one user-selected parameter, i.e., the number of predictors to be chosen randomly at each node, and even this parameter is not highly sensitive. Obviously, it is not possible to interpret each of the trees in RF, but the procedure does provide tables on relative importance among predictor variables.

To our knowledge, we are the first to use BT and RF in the field of ecology (Prasad *et al.*, submitted). Furthermore, it appears that only one biologically based study has used RF (Furlanello *et al.*, 2003).

Methods

We used the three statistical modelling techniques on a data set consisting of: (1) tree importance values, based on tree density and basal area from over 100,000 inventory plots in the eastern U.S. (Iverson and Prasad 1998), for four tree species (red spruce (*Picea rubens*), jack pine (*Pinus banksiana*), white ash (*Fraxinus americana*) and chestnut oak (*Quercus montana*)); (2) 36 environmental (predictor) variables describing climate, soil, land-use, landscape, and topography of each grid cell; and (3) potential future climate based on the Canadian Climate Centre (CCC) global circulation model (Boer *et al.*, 2000). Each cell was 20 x 20 km, for a total of 9,782 cells in the study area. We first ran the models using the dataset with current climate variables and then re-ran the models using CCC variables to get the, CCC-derived, future predictions of suitable habitat. We used the statistical software R (R Development Core Team, 2003), based on the S language (e.g., Chambers & Hastie, 1993), to run RTA ("rpart", Therneau and Atkinson, 1997), BT ("ipred", Peters *et al.*, 2002), and RF ("randomForest", Liaw & Wiener, 2002).

We used three map similarity measures to conduct a pixel-by-pixel comparison (actual vs. predicted) among the three models and four species; correlation, Kappa, and fuzzy Kappa (Hagen 2003). With the Kappa statistic, the level of agreement between maps is based on a contingency table, which details how the distribution of categories in map A differs from map B. Fuzzy Kappa recognizes that categories are often not crisp, i.e., there are grades of similarity between pairs of cells in two maps. The fuzziness of location is set with a function that defines the level to which the neighbouring cells influence the target cell (Hagen 2003).

Results and discussion

Although space does not allow displaying of maps for this paper, the RTA, BT, and RF maps all replicated current distributions well. However, the correlation, Kappa, and fuzzy Kappa tests verified that the BT and RF models were clearly superior to RTA in predicting current distributions of the four species studied (Table 1). The "multiple-perturbed" trees in BT and "multiple-perturbed-randomised" trees in RF allow for better prediction capabilities. While the outputs of BT and RF are fairly similar, we prefer the output of RF for two reasons: 1) RF slightly outperformed BT in most statistics; and 2) RF smoothes the response more than BT. We believe this smoothing provides an advantage because the IVs grade smoothly from lower to higher and there are no abrupt changes or skips in IV classes. These abrupt changes happened more in BT.

Table 1. Correlation and Kappa scores for RTA, BT, and RF among four tree species.

	Correlation			Kappa			Fuzzy Kappa		
	RTA	BT	RF	RTA	BT	RF	RTA	BT	RF
Picea rubens	0.864	0.945	0.953	0.576	0.586	0.589	0.660	0.659	0.660
Pinus banksiana	0.734	0.896	0.919	0.430	0.447	0.477	0.497	0.517	0.539
Fraxinus americana	0.693	0.907	0.923	0.357	0.417	0.441	0.375	0.443	0.455
Quercus montana	0.795	0.940	0.947	0.506	0.513	0.532	0.567	0.579	0.590

Predictions of potential future suitable habitat were biogeographically reasonable and logical, especially for BT and RF. The RF models were slightly more biogeographically realistic because of the smoothed output, though we realise that this kind of reasoning about potential future habitat is subjective and fraught with uncertainty because many factors, including many not considered here, will influence the final distribution. Additional support for the superiority of the RF model compared to BT has also been shown by others (Hawkins and Musser 1999; Meyer *et al.*, 2003; Svetnik *et al.*, 2003).

We propose to use RTA, BT, and RF as a toolbox for species modelling. The superior prediction capability of RF is best used to map future scenarios, while RTA and to some extent BT can be used for their interpretive abilities. If the individual trees (among BT) are similar, a single RTA tree can be used to map what predictors are driving the distribution of the species spatially; a very unique aspect of RTA that offers additional insights into the

species distribution (Iverson and Prasad, 1998; Iverson *et al.*, 1999). We are currently using this procedure to model the future climate distributions of 135 eastern US tree species. Our website (http://www.fs.fed.us/ne/delaware/4153/4153.html) has an online atlas of an earlier version of this work (RTA only) which will be updated using these new tools. We advocate this package of tools for widespread use in predictive biological mapping.

References

Boer, G.J.; Flato, G.M. & Ramsden, D. (2000). A transient climate change simulation with historical and projected greenhouse gas and aerosol forcing: projected climate for the 21st century. *Climate Dynamics* **16**, 427-451.

Breiman, L. (1996). Bagging predictors. *Machine Learning* **24**, 123-140.

Breiman, L. (2001). Random forests. *Machine Learning* **45**, 5-32.

Breiman, L.; Freidman, J.; Olshen, R.; & Stone, C. (1984). Classification and regression trees. Wadsworth, Belmont, CA. 358 pp.

Furlanello, C.; Neteler, M.; Merler, S.; Menegon, S.; Fontanari, S.; Donini, A.; Rizzoli, A. & Chemini, C. (2003). GIS and the random forest predictor: integration in R for tick-borne disease risk assessment. K. Hornik, F. Leisch, & A. Zeileis, (Eds.), *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, 1-11, Vienna.

Hagen, A. (2003). Fuzzy set approach to assessing similarity of categorical maps. *Geographic Information Science* **17**, 235-249.

Hawkins, D.M. & Musser, B.J. (1999). One tree or a forest? Alternative dendrographic models. *Computing Science and Statistics* **30**, 534-542.

Iverson, L.R. & Prasad, A.M. (1998). Predicting abundance of 80 tree species following climate change in the eastern United States. *Ecological Monographs* **68**, 465-485.

Iverson, L.R. & Prasad, A.M. (2001). Potential changes in tree species richness and forest community types following climate change. *Ecosystems* **4**, 186-199.

Iverson, L.R.; Prasad, A.M.; Hale, B.J. & Sutherland, E.K. (1999). *An atlas of current and potential future distributions of common trees of the eastern United States*. General Technical Report NE-265. Northeastern Research Station, USDA Forest Service.

Liaw, A. & Wiener, M. (2002). Classification and regression by Random Forest. *R News* **2/3**, 18-22, URL: http://CRAN.R-project.org/doc/Rnews/.

Melillo, J.M. (1999). Climate change: Warm, warm on the range. *Science* **283**, 183-184.

Meyer, D.; Leisch, F. & Hornik, K. (2003). The support vector machine under test. *Neurocomputing* **55**, 169-186.

Moore, D.E.; Lees, B.G. & Davey, S.M. (1991). A new method for predicting vegetation distributions using decision tree analysis in a geographic information system. *Journal of Environmental Management 15*, 59-71.

Peters, A.; Hothorn, T. & Lausen, B. (2002). Ipred: improved predictions. *R News* **2**, 33-36, URL :http://CRAN.R-project.org/doc/Rnews/.

Prasad, A.; Iverson, L.; Liaw, A. (submitted). A comparison of four computer intensive modelling techniques: RTA, MARS, Bagging Trees and Random Forest, for predicting current and future distribution of trees. *Ecosystems*.

Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J.C.; Sheridan, R.P. & Feuston, B.P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Science* **43**, 1947-1958.

Therneau, T.M. & Atkinson, E.J. (1997). *An introduction to recursive partitioning using the RPART routines*. Technical Report #61. Mayo Clinic. Rochester, MN.