Am. Midl. Nat. (2012) 168:443–455

Comparing Predicted Historical Distributions of Tree Species Using Two Tree-based Ensemble Classification Methods

BRICE B. HANBERRY[1] AND HONG S. HE

Department of Forestry, University of Missouri, 203 Natural Resources Building, Columbia 65211

AND

BRIAN J. PALIK

USDA Forest Service, Northern Research Station, Grand Rapids, Minnesota 55744

ABSTRACT.—Fine scale spatial mapping of historical tree records over large extents is important for determining historical species distributions. We compared performance of two ensemble methods based on classification trees, random forests, and boosted classification, for mapping continuous historical distributions of tree species. We used a combination of soil and terrain predictor variables to predict species distributions for 21 tree species, or species groups, from historical tree surveys in the Missouri Ozarks. Mean true positive rates and AUC values of all species combined for random forests and boosted classification, at a modeling prevalence and threshold of 0.5, were similar and ranged from 0.80 to 0.84. Although prediction probabilities were correlated (mean r = 0.93), predicted probabilities from random forests generated maps with more variation within subsections, whereas boosted classification was better able to differentiate the restricted range of shortleaf pine. Both random forests and boosted classification performed well at predicting species distributions over large extents. Comparison of species distributions from two or more statistical methods permits selection of the most appropriate models. Because ensemble classification trees incorporate environmental predictors, they should improve current methods used for mapping historical trees species distributions and increase the understanding of historical distributions of species.

INTRODUCTION

The United States General Land Office (GLO) records represent vegetation preceding and during Euro-American settlement, and thus the GLO records provide a source for reconstructing historical forests, including historical species distributions. Comparison with current species distributions allows assessment of effects of various processes, such as climate change and fire suppression, which influence species persistence. The GLO instituted Public Land Surveys based on a rectangular survey system of townships and ranges (White, 1983). Each township, at the intersection of north and south-running ranges and east and west-running townships, is composed of 36 1.6 km^2 sections. Surveyors recorded species at the corners and middle of each section line (*i.e.*, every 0.8 km) for two to four trees.

Statistical methods are necessary to predict continuous species distributions from surveys. Competing methods for mapping species distributions include generalized linear models, generalized additive models, linear and mixed discriminant analysis, multivariate adaptive regression splines, artificial neural networks, hierarchical Bayesian models, and classification methods (*e.g.*, Guisan and Thuiller, 2005; Elith *et al.*, 2006). Two relatively novel classification methods are random forests and boosted classification. Random forests classification (Breiman, 2001; Cutler *et al.*, 2007), is a method based on bootstrap aggregation (bagging) by the majority vote of many trees grown using random samples of

[1] Corresponding author: Telephone: (573) 875-5341; e-mail: hanberryb@missouri.edu

both predictor variables and training data. Random forests classification is accurate, non-parametric, non-linear, user-friendly, and generally avoids predictor collinearity issues (but *see* Strobl *et al.*, 2008). Boosted classification (Ridgeway, 1999; Friedman *et al.*, 2000; Elith *et al.*, 2008) is similar to random forests but is sequential and assigns greater weights (boosting) to misclassified cases, to create a better fitting model with the addition of each tree. Ensemble classification methods, and many contemporary modeling methods, have not been applied previously to historical tree surveys. Indeed, most modeling methods used to generate continuous surfaces for historical species distributions have been based on geographical information systems (GIS), such as kriging (*see* He *et al.*, 2000; Manies and Mladenoff, 2000; Cogbill, 2002; Wang, 2009) or triangular irregular networks (Batek, 1999), or even simple means (Friedman and Reich, 2005). These methods do not use environmental information (*i.e.*, predictor variables) to inform the models, and consequently may have lower mapping accuracy than other methods.

Because General Land Office surveys were incomplete, a recorded presence of a tree species does not mean that unrecorded species were absent; rather, these are methodological absences due to scarceness in the surveys (Lobo *et al.*, 2010). For presence-only data, pseudo-absences from surveyed areas produce more consistent models than pseudo-absences that represent the study background (Mateo *et al.*, 2010). Additionally, imbalanced data, where there are few surveyed species compared to the number of methodological absences, are an issue. Classification trees will minimize the overall error rate of both absent and present species rather than more correctly model the present species; however, it is possible to change the focus to model present species through differential sampling and weighting to set the modeling prevalence. Although there was bias in selection of trees by surveyors (Hanberry *et al.*, 2012), if selected trees for the most part were identified and located correctly, bias should not affect species distribution models.

Our objective was to map the continuous historical distributions of 21 tree species or groups of species across the Ozarks ecological section of Missouri at fine spatial resolution (*i.e.*, the hectare scale) using predictive models to improve current methods for modeling historical vegetation. We used random forests and boosted classification combined with environmental predictor variables to model and predict probability of species presence over a large spatial extent. We compared the performance of random forests and boosted classification using true positive rates, species prediction area, area under the curve (AUC) of the receiver operating characteristic (ROC), correlations, and visual assessment by an expert.

METHODS

GENERAL LAND OFFICE SURVEY DATA

We selected all trees, about 285,000, located in the Ozarks section (the southern portion of Missouri; *see* Figures; Nigh and Schroeder, 2002) of the Missouri GLO dataset (J. Harlan, Geographic Resources Center, University of Missouri, Columbia, MO, USA, http://msdis.missouri.edu). The majority (95%) of survey dates occurred between 1815 and 1850. We eliminated duplicates and line trees, which were descriptive of the line but not referenced to a point on the line, and trees that did not have a distance and bearing from the survey point. We grouped the most common tree species into the following categories (Table 1): black gum (*Nyssa sylvatica* var *sylvatica*), eastern redcedar (*Juniperus virginiana*), hackberry (*Celtis occidentalis*), sassafras (*Sassafras albidum*), sycamore (*Platanus occidentalis*), shortleaf pine (*Pinus echinata*), blackjack oak (*Quercus marilandica*), bur oak (*Q. macrocarpa*), chinkapin oak (*Q. muehlenbergii*), pin oak (*Q. palustris*), post oak (*Q. stellata*), white oak (*Q. alba*), black oaks

TABLE 1.—Tree species/group and number of individuals that intersect with spatial modeling units (soil polygons) for GLO (1815–1850) surveys in the Missouri Ozarks. Trees are bearing trees selected by surveyors

Species/group		Count
Ashes	*Fraxinus americana, F. pennsylvanica*	1898
Black gum	*Nyssa sylvatica*	1528
Black oaks	*Quercus velutina,* also *Q. falcata, Q. coccinea, Q. rubra*	46209
Blackjack oak	*Quercus marilandica*	18171
Bottomland (Cottonwood, willow)	*Populus* spp., *Salix* spp.	279
Bur oak	*Quercus macrocarpa*	857
Cherries	*Prunus* spp.	190
Eastern redcedar	*Juniperus virginiana*	249
Chinkapin oak	*Quercus muehlenbergii*	558
Elms	*Ulmus alata, U. americana, U. rubra*	3952
Hackberry	*Celtis occidentalis*	884
Hickories	*Carya cordiformis, C. glabra, C. laciniosa, C. ovata, C. texana, C. tomentosa*	11199
Maples	*Acer saccharum, A. negundo* also *A. rubrum, A. saccharinum*	1943
Mesic (American basswood, American hornbeam, black locust, birch, honeylocust, red and white mulberry, sweetgum)	*Tilia americana, Carpinus caroliniana, Robinia pseudoacacia, Betula* spp., *Gleditsia triacanthos, Morus alba* and *M. rubra, Liquidambar styraciflua*	1551
Walnuts	*Juglans nigra, J. cinerea*	1998
Shortleaf pine	*Pinus echinata*	12289
Pin oak	*Quercus palustris*	1383
Post oak	*Quercus stellata*	52676
Sassafras	*Sassafras albidum*	90
Sycamore	*Platanus occidentalis*	1343
White oak	*Quercus alba*	53614

(primarily black oak, *Quercus velutina*; but including red oak, *Q. rubra*; southern red oak, *Q. falcata*; scarlet oak, *Q. coccinea;* due to patchy distribution of these species), ashes (white ash, *Fraxinus americana*; green ash, *F. pennsylvanica*), cherries (*Prunus* spp.), elms (winged elm, *Ulmus alata*; American elm, *U. americana*; slippery elm, *U. rubra*), hickories (bitternut hickory, *Carya cordiformis*; pignut hickory, *C. glabra*; shellbark hickory, *C. laciniosa*; shagbark hickory, *C. ovata*; black hickory, *C. texana*; mockernut hickory, *C. tomentosa*), maples (primarily sugar maple, *Acer saccharum*; boxelder, *A. negundo*; also red maple, *A. rubrum;* silver maple, *A. saccharinum*), walnuts (black walnut, *Juglans nigra*; butternut, *J. cinerea*), a bottomland, early successional group (cottonwood, *Populus* spp.; willow, *Salix* spp.), and a mesic (moist, low fire frequency) group (American basswood, *Tilia americana*; American hornbeam, *Carpinus caroliniana*; black locust, *Robinia pseudoacacia*; birch, *Betula* spp.; honeylocust, *Gleditsia triacanthos*; sweetgum, *Liquidambar styraciflua;* red and white mulberry, *Morus alba* and *M. rubra*). Post oak had some obvious square survey gaps in areas where white oak had more intense survey presence, suggesting that surveyors recorded post oak as white oak, and for one area in southern Missouri along the Arkansas border, we removed the white oaks recorded by four surveyors.

ENVIRONMENTAL VARIABLES

We used Soil Survey Geographic (SSURGO) Database (Natural Resources Conservation Service, http://soildatamart.nrcs.usda.gov) polygons as our spatial unit for analyses. Due to rugged topography, we split SSURGO polygons that were backslopes (*i.e.*, slopes ≥15%) into backslopes of either southwestern or northeastern aspect (based on DEM aspects; please contact authors for ESRI ModelBuilder model). We removed soil polygons of water and miscellaneous areas disturbed by human developments (*e.g.*, mines, pits, dumps) and we removed seven very large polygons that had areas ≥5000 ha to reduce variance in polygon size. After processing, there were about 835,000 polygons, with a mean area of about 10 ha, and a total extent of 9,073,300 ha.

Soil information from the SSURGO database is based on map unit by county, where map units represent groups of discontinuous polygons with similar soil characteristics in a county (3828 total map units, mean area of 2465 ha). However, because map units characterize a collection of smaller soil polygons, mean values for topographic variables can represent a smaller area than map unit and therefore reduce the areal extent for each unique prediction. We developed a unique unit (mean area of 131 ha) based on map unit, landform (protected southwestern backslope, exposed northeastern backslope, and other), land type association (an ecological classification; Cleland *et al.*, 1997), and bedrock geology; these units thus became our unit for predicted probability.

We used 16 soil variables from the SSURGO database and we calculated seven topographic variables from a 30 m digital elevation model (DEM; Table 2). We also joined subsection, an ecological classification based on geomorphology, soil, climate, and potential natural communities, and bedrock geology designations to each individual soil polygon (Nigh and Schroeder, 2002; Table 2). In addition to ecological characteristics, the larger scale ecological subsection and geology provided information about spatial locations; ecological subsection also was informed by expert opinion, because experts classified subsections based on similar ecological characteristics (Cleland *et al.*, 1997).

MODELING AND PREDICTION

We selected all polygons (about 88,000) that contained trees of any species to use for modeling samples. That is, surveyed polygons were our source for both polygons with a specific tree species and pseudo-absences, or polygons without a specific tree species. We then oversampled the minority class (polygons with a specific tree species) to reduce the inequality between the number of spatial sampling units with and without the tree species for the modeling dataset. For each species or species group, we oversampled by randomly selecting 67% of polygons with the species, up to 2500 polygons, for modeling, and held back the rest for prediction and validation. For pseudo-absences of each species or species group, we randomly selected 2500 polygons that contained survey points without a recorded species presence for modeling, rather than points randomly drawn from (the background of) the entire study area (Mateo *et al.*, 2010). We made predictions for polygons not used in modeling and polygons without trees.

We used the randomForest package (Liaw and Wiener, 2002) in R statistical software (R Development Core team, 2010) to model and predict each species or species group. With random forests, we can down-sample the majority class (if there were not 2500 present cases) and thus, we used an equal number of pseudo-absent and present cases to fix the modeling prevalence to 0.5, giving an equal weight to polygons with and without trees in the modeling. We set the bag fraction, or sub-sampling rate (*i.e.*, the sampsize option, which is sampled without replacement), at 67% of the selected polygons with the species. We then

TABLE 2.—Predictor variables for modeling

Variable	Type	Notes
Landform	Categorical soil	Bottomlands, northeastern backslope, southwestern backslope, and uplands
Parent material	Categorical soil	*e.g.*, alluvium, colluvium, residuum
Limestone class	Categorical soil	No limestone, limestone in combination, limestone
Drainage class	Categorical soil	Very poorly drained to excessively drained
Soil taxonomic order	Categorical soil	
Flooding frequency	Categorical soil	
Soil layer restriction type	Categorical soil	None, fragipan or claypan, bedrock
Soil depth (cm)	Continuous soil	Depth to either the bottom of the soil profile or soil restriction
Water holding capacity (cm/cm)	Continuous soil	
PH	Continuous soil	
Base saturation	Continuous soil	ECEC/sum of bases
Fragments (%)	Continuous soil	
Organic matter (%)	Continuous soil	
Clay (%)	Continuous soil	
Sand (%)	Continuous soil	
Elevation (m)	Continuous topographic	
Slope (%)	Continuous topographic	
Transformed aspect	Continuous topographic	1 + sin(aspect/180*3.14 + 0.79); Beers *et al.*, 1966
Solar radiation	Continuous topographic	0700 to 1900 in 4 hour intervals on summer solstice for re-sampled 60 m DEM
Roughness	Continuous topographic	Sappington *et al.*, 2007
Wetness (convergence) index	Continuous topographic	T. Dilts, University of Nevada, Reno, NV, USA; http://arcscripts.esri.com
Topographic position index	Continuous topographic	
Ecological subsection	Categorical	
Bedrock geology	Categorical	

used standard or recommended options, which appeared to have low sensitivity to change. We set the number of classification trees at 1000 (increasing the number of trees did not improve true positive rate for predictions) and the number of variables randomly sampled at each split as the square root of the number of predictors.

For boosted classification, we used code by J. Leathwick and J. Elith (Elith *et al.*, 2008) based on the R gbm package (G. Ridgeway, http://www.i-pensieri.com/gregr/gbm.shtml) with the Bernoulli distribution to model and predict each species or species group. For boosted classification, we could increase the weight of polygons with the given tree species. We set the weight so that the present case sample size was equal to the modeling sample size of 2500 for polygons without present trees (*e.g.*, if there were 1000 present trees for modeling, then the weight was 2.5), an adjustment corresponding to setting the modeling prevalence at 0.5 in random forests. We selected standard options that were similar to values to random forests. We set the bag fraction or sub-sampling rate at 0.67, the shrinkage or learning rate (*i.e.*, the influence of each tree to the developing model) at 0.005 (decreased learning rate did not improve performance), the interaction depth or tree complexity at 8, and then allowed the program to determine the optimal number of trees based on AUC values.

VALIDATION AND COMPARISONS

We used the ROCR package (Sing *et al.*, 2005) in R to calculate the true positive rate (*i.e.*, present cases correctly identified as present) and AUC (*i.e.*, true positive rate plotted against false positive rate or 1- true negative rate) values for predictions at a threshold of 0.5, equal to the modeling prevalence; predicted probabilities ≥ the threshold of 0.5 represented species presence (Liu *et al.*, 2005). We also calculated the area predicted as present at a threshold of 0.5 as a surrogate for the true negative rate (*i.e.*, absent cases correctly identified as absent), because although the absent cases are unknown, a smaller area allocated to presence relative to the true positive rate indicates less error in prediction of presence. We statistically compared predicted probabilities using Pearson's correlation coefficient (Proc Corr; SAS software, version 9.1, Cary, North Carolina). We grouped the predictions into four cut-off bins (0–25%, 25–50%, 50–75%, 75–100%) and mapped the distributions. We consulted a regional expert (Tim Nigh, Missouri Department of Conservation) to visually assess the maps for correspondence with expected (*i.e.*, by field and ecological) distributions. Due to potential problems with random forests classification (Strobl *et al.*, 2007, 2008), we examined (1) correlation among the most important variables and (2) number of classes in categorical variables.

RESULTS

Predicted probabilities from random forests and boosted classification generally were similar and accurate for predicting presence based on true positive rate, AUC, and species area values. Mixed and/or wide-ranging species had lower performance in accuracy metrics. Mean true positive rates for all species were 0.83 and 0.80 for random forests and boosted classification, respectively (Table 3). True positive rates ranged from 0.67 to 0.98 for random forests and 0.53 to 0.94 for boosted classification. Mean values (all species combined) for species prediction area (area predicted as present at a threshold of 0.5) and AUC were the same (0.27 vs. 0.26 for species prediction area and 0.84 vs. 0.83 for AUC, for random forests and boosted classification, respectively). Differences among metrics were minor among the species, except for sassafras, for which random forests performed better than boosted classification (Table 3). With fine tuning of the modeling parameters for boosted classification, we could increase the true positive rate of sassafras at the expense of the AUC value.

Overall, predicted probabilities for species distribution models from random forests and boosted classification were correlated strongly (mean $r = 0.93$; Table 3). By species, black oak and hickories had the lowest correlation ($r = 0.89$) between predicted probabilities from the two methods. Variation was due, in part, to the weight assigned to predictor variables by the different methods. Subsection was one of the top five predictor variables for all models produced by boosted classification (Table 4). Mean rank of subsection for all species was 1.7 for boosted classification and 3.8 for random forests. Elevation, aspect, and geology were weighted more heavily and wetness index and slope were weighted lower by boosted classification than by random forests.

TABLE 3.—True positive rates, predicted area (fraction of total area; proxy for true negative rate), AUC, and correlation values (r) for random forests (RF) and boosted classification (BC)

	True positive rate		Area		AUC		
Species	RF	BC	RF	BC	RF	BC	r
Ashes	0.80	0.78	0.27	0.29	0.83	0.81	0.94
Black gum	0.86	0.86	0.23	0.24	0.86	0.84	0.97
Black oak	0.67	0.69	0.57	0.57	0.54	0.51	0.89
Blackjack oak	0.85	0.86	0.39	0.40	0.84	0.83	0.95
Bottomland	0.91	0.89	0.07	0.06	0.98	0.98	0.90
Bur oak	0.86	0.81	0.13	0.13	0.93	0.92	0.93
Cherries	0.73	0.65	0.18	0.16	0.90	0.88	0.95
Chinkapin oak	0.88	0.88	0.18	0.17	0.90	0.90	0.95
Eastern redcedar	0.98	0.92	0.19	0.14	0.95	0.94	0.91
Elms	0.80	0.79	0.34	0.33	0.81	0.80	0.92
Hackberry	0.85	0.81	0.12	0.11	0.96	0.95	0.95
Hickories	0.69	0.70	0.53	0.51	0.63	0.62	0.89
Maples	0.87	0.85	0.26	0.26	0.87	0.85	0.95
Mesic	0.80	0.79	0.21	0.22	0.88	0.87	0.95
Pin oak	0.83	0.80	0.24	0.26	0.88	0.85	0.93
Post oak	0.81	0.82	0.49	0.48	0.73	0.72	0.94
Sassafras	0.80	0.53	0.16	0.09	0.89	0.84	0.91
Shortleaf pine	0.94	0.94	0.20	0.21	0.90	0.89	0.97
Sycamore	0.88	0.87	0.11	0.11	0.95	0.95	0.93
Walnuts	0.78	0.77	0.28	0.27	0.83	0.82	0.94
White oak	0.73	0.74	0.50	0.49	0.59	0.58	0.93

Species distribution maps from the two modeling techniques had subtle differences. Maps based on probability bins from random forests contained more variation (measured visually by color differences) within subsections, whereas boosted classification produced cleaner species distribution maps that conformed to the subsection units (*e.g.,* Fig. 1 for hickories, which have the least correlation). In general, we believed that greater variation generated a better map than correspondence to subsection, with the one exception of shortleaf pine. Shortleaf pine has a restricted range in the southern portion of the Ozarks, and boosted classification was better able to predict pine on its known contemporary range, although the

TABLE 4.—Most frequent variables for all species from the top five variables for each species

Random forests		Boosted classification	
Variable	Number of models	Variable	Number of models
Subsection	16	Subsection	21
Wetness index	15	Elevation	16
Slope	12	Aspect	15
Elevation	11	Wetness index	8
Parent material	8	Solar radiation	7
Solar radiation	8	Geology	6
Roughness	7	Parent material	6
Aspect	4	Roughness	6
Flooding	4	Slope	5
Geology	3	Flooding	4

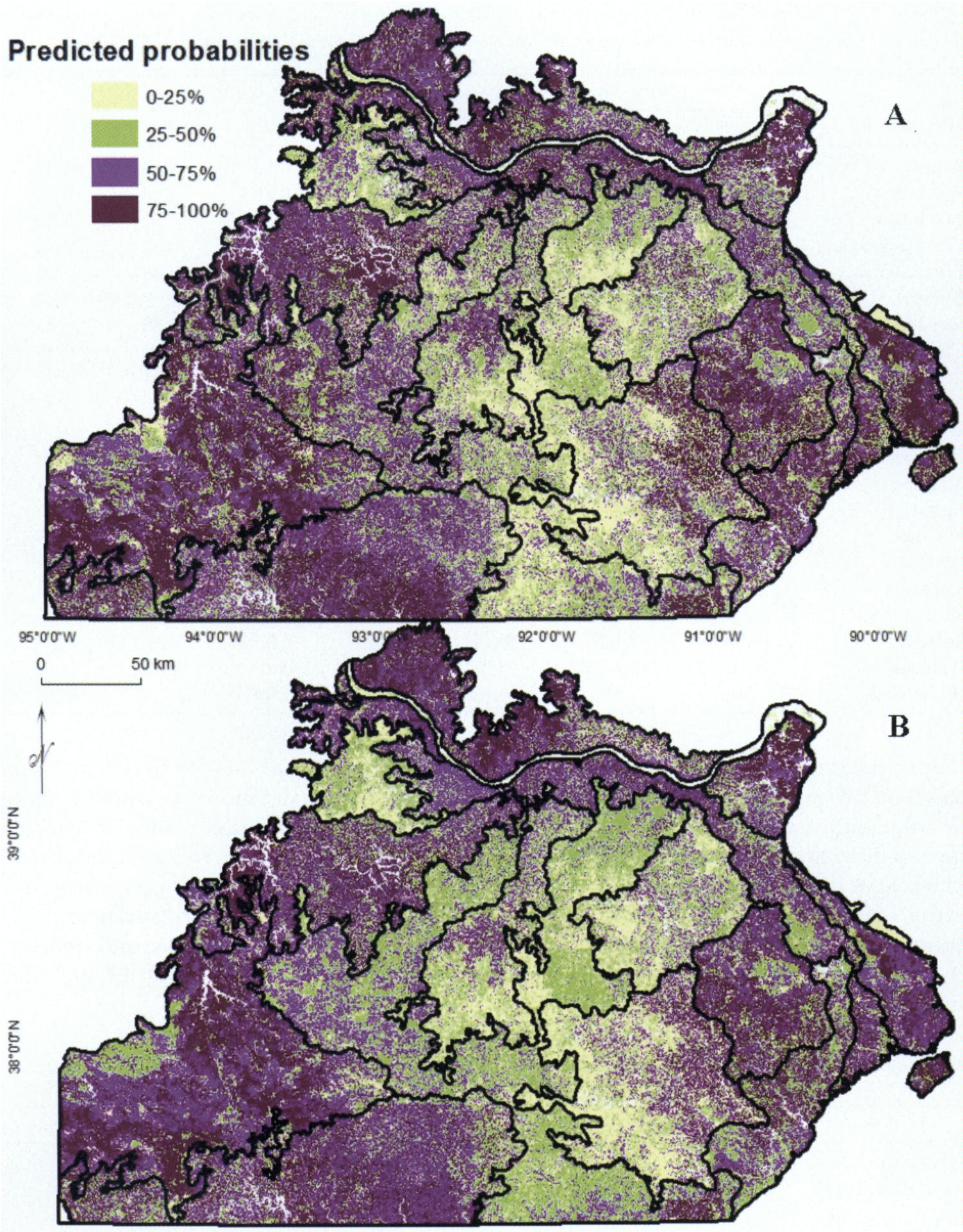

FIG. 1.—Predicted probabilities for hickories by (**A**) random forests and (**B**) boosted classification. Random forests produced slightly more variation within subsections (outlined)

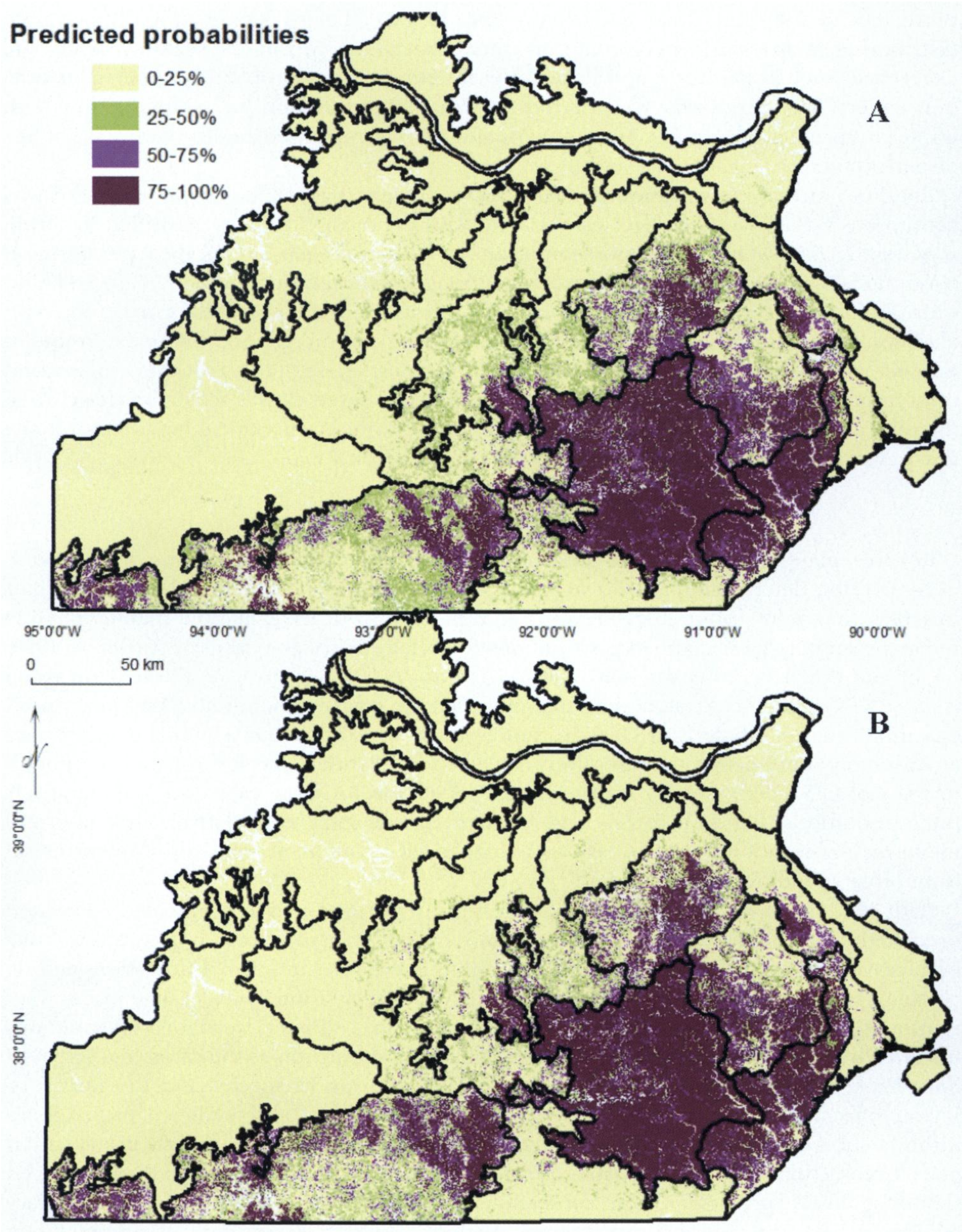

FIG. 2.—Predicted probabilities for shortleaf pine by (**A**) random forests and (**B**) boosted classification. Random forests produced greater predicted probabilities beyond current shortleaf pine distribution than boosted classification

predicted distribution still did extend the known species range (Fig. 2). Overall, distribution maps did not contain as much variation within subsections from site differences such as landform and slope differentiating species probability of occurrence as our expert had expected. We tried to increase the variation by reducing the least important predictor variables, but the maps and metrics remained similar, with a slight loss of performance.

Random forests classification did not exhibit clear bias in selection of correlated continuous variables and categorical variables with the most classes, as identified by Strobl *et al.* (2007, 2008). Of the five most important variables for each model, the most frequent continuous variable was wetness index, and although wetness index had a −0.69 correlation with slope, the second most frequent variable, wetness index had a 0.008 correlation with elevation, the third most frequent variable. Ecological subsection, the most frequent variable, also was the categorical variable with the most classes (13). However, the second most frequent categorical variable, parent material, had seven classes, which was fewer than the nine classes of bedrock geology, the fourth most frequent categorical variable, while the third most frequent categorical variable, flooding, had only two classes.

DISCUSSION

Reliable mapping of historical tree records over large extents with fine resolution is necessary for determining historical species distributions and the stand-scale predictor variables that were influential. Previous studies of historical vegetation that used GLO records generally have mapped distributions for relatively small extents (*e.g.*, Batek, 1999; Leahy and Pregitzer, 2003; He *et al.*, 2007). He *et al.* (2000), Schulte *et al.* (2002), and Wang *et al.* (2009) conducted statewide mapping but at coarse resolutions. We introduced ensemble methods based on classification trees, a statistical approach that incorporates environmental predictors that are often neglected in historical species distribution models in favor of GIS operations, to map historical vegetation for large extents at comparatively finer resolutions. Random forests and boosted classification are relatively new improvements on classification trees that take advantage of averaging many trees and randomization to improve model fit.

Both random forests and boosted classification performed well and provided correlated predicted probabilities. Ecological subsection overall was the most important variable because subsections represent the effects of multiple variables in space. Interactions among climate, topography at different scales, soils, and vegetation differentiate areas that reinforce or suppress disturbance processes such as fire, which determined distributions historically. Generalized boosting identified ecological subsection as the variable of greatest importance; ecological subsection was the top variable for 12 species and the fourth or higher most important variable for two other species. Therefore, predicted probabilities within a subsection were comparatively constant. In contrast, random forests classification placed subsection importance as first for eight species and as low as 14th and thus produced slightly greater variation within subsections. Visual assessment, rather than accuracy assessment, of the species distribution models allowed recognition of these subtle differences.

We were not able to increase the amount of variation in spatial distribution. Reasons for lack of variation include trees outside of their typical distribution, due to error in tree location or tree presence in an atypical ecological site. Additionally, we chose to use vector shapes as mapping units instead of raster pixels to better correspond with soil and ecological variables; however, the individual polygons that formed the map units do not conform to landform shape well. Users of these models should moderate predicted probabilities based on local physiography.

To reconstruct historical forests, species distribution modeling and prediction are important steps. Ensemble classification trees provide distribution maps and important environmental variables, which allow comparisons to current and future predicted distributions for restoration and management goals. Changes in species distributions, whether directional, expansions to new ecological areas, or contractions, provide patterns to assess land use changes. Although not yet demonstrated, historical tree distributions that were more southerly than current distributions may indicate responses to climate change. Maps of historical oak, hickory, and pine distributions versus mesic species such as maples, elms, and ashes may show distribution patterns under a constant fire regime, where fire-dependent species were present in open and exposed areas and fire-intolerant species were restricted to fire shadows along streams and steep slopes (*e.g.,* Nowacki and Abrams, 2008). Strong regional patterns along ecological lines may be replaced by uniformity, indicating current homogenous land use across a landscape (*e.g.,* Schulte *et al.,* 2007). Historical disturbance regions may be shown by expanded prediction areas for late successional species and contracted prediction areas of early successional species, or potentially vice versa (*e.g.,* Friedman and Reich, 2005; Wang *et al.,* 2009). From this information and combined with historical forest density estimates, we can develop a greater understanding of historical forest conditions and the changes that have occurred in contemporary forests.

CONCLUSIONS

All statistical methods will create different models and multiple methods permit flexibility of choice. For our study area, random forests classification provided more variation spatially, thus better representing environmental gradients, but boosted classification was better able to predict the restricted range of shortleaf pine. Lawrence *et al.* (2004) found that gradient boosting produced overall lower accuracy but greater accuracy for some classes of spectral imagery than classification tree analysis. If there is no apparent difference in performance metrics and maps appear equally reasonable, then averaging of models is an option. Regardless, the use of a statistical method that requires predictor variables is an improvement on models that use GIS operations alone.

Acknowledgments.—We received help from Rencang Bu, Bill Dijak, Jane Elith, Jennifer Goyne, Erica Serna, Kyle Steele, Tim Nigh, and funding from the National Fire Plan and the USDA Forest Service, Northern Research Station.

LITERATURE CITED

BATEK, M. J., A. J. REBERTUS, W. A. SCHROEDER, T. L. HAITHCOAT, E. COMPAS, AND R. P. GUYETTE. 1999. Reconstruction of early nineteenth-century vegetation and fire regimes in the Missouri Ozarks. *J. Biogeogr.,* **26**:397–412.

BEERS, T. W., P. E. DRESS, AND L. C. WENSEL. 1966. Aspect transformation in site productivity research. *J. Forest.,* **64**:691–692.

BREIMAN, L. 2001. Random Forests. *Mach. Learn.,* **40**:5–32.

CLELAND, D. T., P. E. AVERS, W. H. MCNAB, M. E. JENSEN, R. G. BAILEY, T. KING, AND W. E. RUSSELL. 1997. National Hierarchical Framework of Ecological Units, p. 181–200. *In:* M. S. Boyce and A. Haney (eds.). Ecosystem Management Applications for Sustainable Forest and Wildlife Resources. Yale University Press, New Haven, Connecticut, USA.

COGBILL, C. V., J. BURK, AND G. MOTZKIN. 2002. The forests of presettlement New England, USA: spatial and compositional patterns based on town proprietor surveys. *J. Biogeogr.*, **29**:1279–1304.

CUTLER, D. R., T. C. EDWARDS, Jr., K. H. BEARD, A. CUTLER, K. T. HESS, J. GIBSON, AND J. J. LAWLER. 2007. Random forests for classification in ecology. *Ecology*, **88**:2783–2792.

ELITH, J., C. H. GRAHAM, R. P. ANDERSON, M. DUDÍK, S. FERRIER, A. GUISAN, R. J. HIJMANS, F. HUETTMANN, J. R. LEATHWICK, A. LEHMANN, J. LI, L. G. LOHMANN, B. A. LOISELLE, G. MANION, C. MORITZ, M. NAKAMURA, Y. NAKAZAWA, J. McC. OVERTON, A. T. PETERSON, S. J. PHILLIPS, K. S. RICHARDSON, R. SCACHETTI-PEREIRA, R. E. SCHAPIRE, J. SOBERÓN, S. WILLIAMS, M. S. WISZ, AND N. E. ZIMMERMANN. 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**:129–151.

———, J. R. LEATHWICK, AND T. HASTIE. 2008. A working guide to boosted regression trees. *J. Anim. Ecol.*, **77**:802–813.

FRIEDMAN, J., T. HASTIE, AND R. TIBSHIRANI. 2000. Additive logistic regression: a statistical view of boosting. *Ann. Statist.*, **28**:337–407.

FRIEDMAN, S. K. AND P. B. REICH. 2005. Regional legacies of logging: departure from pre-settlement forest conditions in northern Minnesota. *Ecol. Applicat.*, **15**:726–744.

GUISAN, A. AND W. THUILLER. 2005. Predicting species distribution: offering more than simple habitat models. *Ecol. Lett.*, **8**:993–1009.

HE, H. S., D. J. MLADENOFF, T. A. SICKLEY, AND G. G. GUNTENSPERGEN. 2000. GIS interpolations of witness tree records (1839–1866) for northern Wisconsin at multiple scales. *J. Biogeogr.*, **27**:1031–1042.

———, D. C. DEY, X. FAN, M. B. HOOTEN, J. M. KABRICK, C. K. WIKLE, AND Z. FAN. 2007. Mapping pre-European settlement vegetation at fine resolutions using a hierarchical Bayesian model and GIS. *Plant Ecol.*, **191**:85–94.

LAWRENCE, R., A. BUNN, S. POWELL, AND MICHAEL. ZAMBON. 2004. Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. *Remote Sens. Env.*, **90**:331–336.

LEAHY, M. J. AND K. S. PREGITZER. 2003. A comparison of pre-settlement and present-day forests in northeastern lower Michigan. *Am. Midl. Nat.*, **149**:71–89.

LIAW, A. AND M. WIENER. 2002. Classification and Regression by randomForest. *R News*, **2**:18–22.

LIU, C., P. M. BERRY, T. P. DAWSON AND R. G. PEARSON. 2005. Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, **28**:385–393.

LOBO, J. M., A. JIMÉNEZ-VALVERDE, AND J. HORTAL. 2010. The uncertain nature of absences and their importance in species distribution modeling. *Ecography*, **33**:103–114.

MANIES, K. L. AND D. J. MLADENOFF. 2000. Testing methods to produce landscape-scale presettlement vegetation maps from the U.S. public land survey records. *Landsc. Ecol.*, **15**:741–754.

MATEO, R. G., T. B. CROAT, A. M. FELICÍSIMO, AND J. MUÑOZ. 2010. Profile or group discriminative techniques? Generating reliable species distribution models using pseudo-absences and target-group absences from natural history collections. *Divers. and Distribut.*, **16**:84–94.

NIGH, T. A. AND W. A. SCHROEDER. 2002. *Atlas of Missouri ecoregions*. Missouri Department of Conservation Publication, Missouri, USA. 212 p.

NOWACKI, G. J. AND M. D. ABRAMS. 2008. The demise of fire and 'mesophication' of forests in the eastern United States. *Bioscience*, **58**:123–138.

RIDGEWAY, G. 1999. The state of boosting. *Comput. Science and Stat.*, **31**:172–181.

SAPPINGTON, J. M., K. M. LONGSHORE, AND D. B. THOMPSON. 2007. Quantifying landscape ruggedness for animal habitat analysis: a case study using bighorn sheep in the Mojave Desert. *J. Wildl. Manage.*, **71**:1419–1426.

SCHULTE, L. A., D. J. MLADENOFF, AND E. V. NORDHEIM. 2002. Quantitative classification of historical northern Wisconsin (U.S.A.) landscape: mapping forests at regional scales. *Can. J. For. Res.*, **32**:1616–1638.

———, ———, T. R. CROW, L. C. MERRICK, AND D. T. CLELAND. 2007. Homogenization of northern U.S. Great Lakes forests due to land use. *Landsc. Ecol.*, **22**:1089–1103.

SING, T., O. SANDER, N. BEERENWINKEL, AND T. LENGAUER. 2005. ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**:3940–3941.

STROBL, C., A-L. BOULESTEIX, A. ZEILEIS, AND T. HOTHORN. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinform.*, **8**:25. doi:10.1186/1471-2105-8-25

———, ———, T. KNEIB, T. AUGUSTIN, AND A. ZEILEIS. 2008. Conditional variable importance for random forests. *BMC Bioinform.*, **9**:307. doi:10.1186/1471-2105-9-307

WANG, Y-C., B. J. KRONENFELD, AND C. P. S. LARSEN. 2009. Spatial distribution of forest landscape change in western New York from presettlement to the present. *Can. J. For. Res.*, **39**:76–88.

WHITE, C. A. 1983. A History of the Rectangular Survey System. Bureau of Land Management, Government Printing Office, Washington D.C., U.S. 774 p.

SUBMITTED 11 NOVEMBER 2011 ACCEPTED 2 MARCH 2012