Ecological Modelling 227 (2012) 29–33

Contents lists available at SciVerse ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Sample sizes and model comparison metrics for species distribution models

B.B. Hanberry[a,*], H.S. He[a], D.C. Dey[b]

[a] University of Missouri, 203 Natural Resources Building, Columbia, MO 65211, USA
[b] USDA Forest Service, Northern Research Station, University of Missouri, 202 Natural Resources Building, Columbia, MO 65211 USA

ARTICLE INFO

Article history:
Received 22 April 2011
Accepted 5 December 2011

Keywords:
Correlation
Forest inventory and analysis
Intraclass correlation coefficient
Kappa

ABSTRACT

Species distribution models use small samples to produce continuous distribution maps. The question of how small a sample can be to produce an accurate model generally has been answered based on comparisons to maximum sample sizes of 200 observations or fewer. In addition, model comparisons often are made with the kappa statistic, which has become controversial. Therefore, we used sample sizes ranging from 30 to 2500 individuals to model 16 tree species or species groups in Minnesota's Laurentian Mixed Forest. We compared all smaller sample sizes to models for 2500 records and then 1000 records using Cohen's kappa, Pearson's *r*, Cronbach's alpha, and two intraclass correlation coefficients. We then began confirmation of our findings by repeating the process using a smaller extent in a different area, a portion of Missouri's Central Hardwoods. Although there are disadvantages to using the kappa statistic and intraclass correlation coefficients, due to conversion to categories or computation limitations respectively, the model comparison metrics produced similar results. Comparison values depend on the maximum sample size, and at sample sizes roughly around 10–20% of the maximum sample size, values will begin to decrease more rapidly. Models may not be very accurate below a sample size of 200, for our study areas, extents, and grains. Nonetheless, models based on small sample sizes still may provide information for rare species. We recommend using the full sample available for modeling, after using a partial sample for accuracy assessment. Future research is needed to confirm our findings for different areas, extents, grains, and species.

1. Introduction

Species distribution models use small samples from point locations to predict species occurrence probability for a continuous spatial extent. The accuracy of species distribution models may vary by species, statistical method, explanatory variables, and study extent among other factors, although the unique ecological characteristics of species, and consequent diverse distribution patterns, may explain the greatest variance (Guisan et al., 2007; Syphard and Franklin, 2010). Nevertheless, the accuracy of species distribution models also depends on both sample size and the method for comparison of models.

Sample size is an important consideration for modeling accuracy, particularly for rare species where there are few samples. Small sample sizes that produce inaccurate models may provide some information, but uncertainties associated with these models are high. Although there is much research that compares ever-changing statistical methods, establishing an appropriate sample size as a base for appropriate comparisons has not been common. Studies that have focused on sample size also have used small maximum sample sizes for comparison (e.g. Stockwell and Peterson, 2002; Kadmon et al., 2003; Hernandez et al., 2006; Wisz et al., 2008). Even though 100–200 individuals may be more records than available, models for 100 individuals may not be the best standard.

To measure the agreement among species distribution maps, Cohen's kappa commonly is used (Cohen, 1960). Although the kappa statistic is meant to account for chance agreement, the definition of chance is uncertain (Vaughan and Omerod, 2005). The kappa statistic also behaves paradoxically due to prevalence (number of present cases) and location of species distributions (McPherson et al., 2004; Jiménez-Valverde et al., 2008). Therefore, other metrics to measure accuracy may be preferable.

One option for measurement of model agreement is the familiar interclass correlation coefficient. Interclass correlation coefficients, such as the commonly used Pearson's *r* or more rare Cronbach's alpha, are used to correlate different variables (such as height and weight), and consequently, different variance. For Pearson's *r* and Cronbach's alpha, the magnitude of difference between variables does not matter. For example, pairwise values of 0.1 and 0.8, 0.2 and 0.9, and 0.3 and 1.0, would be correlated and yet are very different values for species distribution maps.

Another option is intraclass correlation coefficients, which measure the relationship between the same variable from different

* Corresponding author. Tel.: +1 573 875 5341x230; fax: +573 882 1977.
E-mail address: hanberryb@missouri.edu (B.B. Hanberry).

0304-3800/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2011.12.001

sources (Shrout and Fleiss, 1979). This option may be more suited for model comparison, where the same variable (predicted probability) is compared and variation arises from methodological choices, for example, when comparing predicted probabilities from different statistical methods (i.e. different sources). Two intraclass correlation coefficient values may be calculated: an absolute agreement metric and a consistency metric. The absolute agreement metric incorporates the different sources in the comparison, whereas the consistency metric excludes variance from the source (i.e. different methods), which is desirable when the magnitude of difference is irrelevant. The consistency metric is similar to Pearson's correlation, but based on an additive rather than a linear transformation to relate lower and higher values (McGraw and Wong, 1996).

Due to the importance of sample size and the choice of accuracy assessment metric, we had two objectives. First, we used much larger sample sizes than previous studies to evaluate adequate minimum sample sizes. We examined sample sizes ranging from 30 to 2500 individuals of 16 tree species or species groups in a roughly 5 million hectare area of Minnesota. Secondly, in order to compare species distribution models developed under various sample sizes, we needed to determine if the current option, the kappa statistic, was as reliable as alternative options. We compared all smaller sample sizes to models based on 2500 records and then 1000 records using Pearson's correlation, Cronbach's alpha, Cohen's kappa, and two intraclass correlation coefficients. We then evaluated the comparison metrics for differences to identify the strengths and weaknesses of each metrics. To strengthen our findings, we repeated the process for a smaller extent in the Central Hardwoods of Missouri. Our work will provide guidance in selection of appropriate sample size and metrics for species distribution models.

2. Methods

2.1. Study area

The primary study area covers about half of the 9.3 million hectare Laurentian Mixed Forest province in northeastern Minnesota (Fig. 1; National Hierarchical Framework of Ecological Units; ECOMAP, 1993). In the Laurentian Mixed Forest province, landforms (e.g. moraines and wetlands) were created by glaciers (Albert, 1995). Annual precipitation increases from about 55 cm in the west to 80 cm in the east and long, cold winters prevail (mean annual temperature about 2 °C).

2.2. Tree surveys

The USDA Forest Service Forest Inventory and Analysis (FIA) surveys fixed plots (each composed of four subplots that are a total of 0.065 ha) during a five year cycle. The latest complete cycle was during 2004–2008 for Minnesota's Laurentian Mixed Forest (Fig. 1). The USDA Forest Service joined our predictor variables to plots (in a table but based on accurate spatial locations) for modeling and prediction because the available FIA plot locations are fuzzed (i.e. location moved) and swapped to protect landowner privacy.

We selected tree species that had at least 2500 individuals. The species were American Basswood (*Tilia Americana*), balsam fir (*Abies balsamea*), balsam poplar (*Populus balsamifera*), black ash (*Fraxinus nigra*); black spruce (*Picea mariana*), bur oak (*Quercus macrocarpa*), jack pine (*Pinus banksiana*), northern white cedar (*Thuja occidentalis*), paper birch (*Betula papyrifera*), red maple (*Acer rubrum*), red pine (*Pinus resinosa*), sugar maple (*A. saccharum*), tamarack (*Larix laricina*), and quaking aspen (*Populus tremuloides*). We also created two mixed species groups by genus, aspens (*Populus tremuloides*, *P. balsamifera*) and maples (*Acer rubrum*, *A. saccharum*).

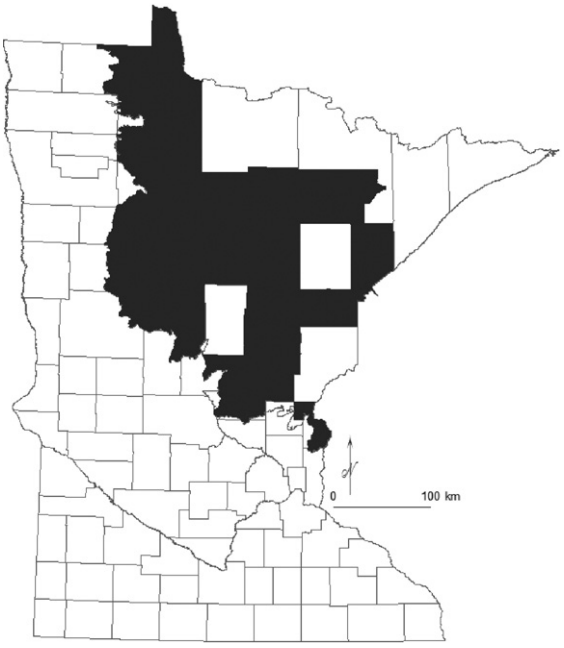

Fig. 1. Primary study area (shaded black), about 5 million ha in the Laurentian Mixed Forest of Minnesota.

2.3. Spatial units and environmental variables

Our spatial units were Soil Survey Geographic (SSURGO) Database (Natural Resources Conservation Service; http://soildatamart.nrcs.usda.gov) polygons. Soil surveys have not been completed in Cook, Crow Wing, Isanti, Koochiching, Lake, Pine, and St. Louis counties, leaving a study extent of about 4,895,238 ha (Fig. 1). After removal of polygons that were water or otherwise miscellaneous areas e.g. mines, pits, dumps), there were 310,000 soil polygons.

We used sixteen predictor variables that are important for tree presence. For soil variables, we determined values based on polygons with similar characteristics by county (map units; 2364 map units total). Soil variables were (1) drainage class (very poorly drained to excessively drained), (2) hydric soil presence class, (3) water holding capacity (cm/cm), (4) pH, (5) organic matter (%), (6) clay (%), and (7) sand (%). We intersected two more categorical variables to each soil polygon: (8) ecological subsection, which is an ecological classification (ECOMAP, 1993), and (9) bedrock geology. From a 30 m DEM (digital elevation model), we determined mean values of terrain variables by a unique unit of map unit, land type association (an ecological classification), and bedrock geology, which contained spatially distinct soil polygons that averaged about 210 ha, and became our unit for predicted probabilities. Terrain variables were (10) elevation (m), (11) slope (%), (12) transformed aspect (1 + sin(aspect/180/3.14 + 0.79; Beers et al., 1966) (13) solar radiation (0700–1900 in 4 h intervals on summer solstice for re-sampled 60 m DEM), (14) topographic roughness (Sappington et al., 2007), (15) wetness convergence, and (16) topographic position index (T. Dilts; http://arcscripts.esri.com).

2.4. Sample sizes and statistical analysis

We randomly selected 2500, 1250, 1000, 5000, 200, 100, 50 and 30 polygons for each tree species or tree species group for modeling. We reserved the rest of the present samples for accuracy

assessment. We selected up to 2500 polygons with plots that did not contain the species to serve as pseudoabsences.

Random Forests classification and regression trees are a non-parametric alternative to traditional statistical methods (Breiman, 2001; Cutler et al., 2007). Random Forests grows multiple trees by drawing different bootstrap samples and only a few randomly selected variables are used to grow individual branches. We used the randomForest package (Liaw and Wiener, 2002) in R statistical software (R Development Core Team, 2010). We fixed the number of classification trees to 1000. We set the number of variables randomly sampled at each split as the square root of the number of predictors. For modeling, we used a modeling prevalence of 80% (1 pseudoabsence: 4 presence) to focus the statistical method on the known presences rather than the uncertain absences. We selected a prevalence that was greater than 50% because we wanted models based on what was known (the present cases) rather than unknown (pseudoabsences).

2.5. Species distribution model comparisons

We compared predicted probabilities for a sample size of 2500 to all sample sizes and a sample size of 1000 to all lesser samples sizes (i.e. 500, 200, 100, 50, 30). We used Pearson's correlation coefficient and Cronbach's alpha (Proc Corr; SAS software, Version 9.2, Cary, NC, USA), weighted kappa (SAS Proc Freq; with weighted kappa, partial credit is given to near groups), and intraclass metrics of (1) an absolute agreement metric (ICC2; includes magnitude of difference) and (2) a consistency metric (ICC3; %INTRACC SAS macro; R.M. Hamer). Due to computer memory requirements for calculating intraclass metrics, we reduced the entire dataset to 2365 records, representing one predicted probability per map unit.

2.6. Accuracy assessment

To compare species distribution models for each species by sample size, we used assessments that did not require known absence cases to evaluate omission and commission error. We calculated the true positive rate (ROCR package in R; Sing et al., 2005) and area (as a surrogate for specificity, the true negative rate) at a 75% threshold for species presence. We also determined the overall mean of predicted probabilities for each sample size by species.

2.7. Confirmation

To determine if our results were applicable to other, smaller areas, we repeated this process in the Missouri Ozarks, part of the Central Hardwoods. Units for terrain variables (and unique prediction probabilities) were about 131 ha, smaller than the units for Minnesota. The FIA sampling density of plots is about 1 per 2500 ha, which is too sparse to have 1000 or more trees of many species in a smaller area. We selected one ecological subsection, the Current River Hills which has an areal extent of about 800,000 ha, that had two tree species with abundance that was greater than 1000 (and enough pseudoabsences to maintain an 80% modeling prevalence). The tree species were black oak (*Quercus velutina*) and scarlet oak (*Q. coccinea*). We compared predictions among sample sizes of 1000, 500, 200, 100, and 50 individuals.

3. Results

For model comparisons of all smaller sample sizes (1000, 500, 200, 100, 50, 30) to a maximum of 2500 records (where the species were present), all the metrics produced similar trends, albeit with different values and slopes (Fig. 2a). Compared to a sample size of 2500, metric values decreased steadily with sample size and then more rapidly when sample sizes reached 500, or 20% of the maximum sample size, and then diminished even more rapidly at 200, or about 10% of the maximum sample size. Pearson's *r* and ICC2 had a similar line slope as did Cronbach's alpha and ICC3. The kappa statistic decreased the most, to a value at sample size of 30 that was 40% of the original value for sample size 2500 (See Appendix A for correlation values among the sample sizes).

When we repeated this process to compare sample sizes of 500 and less to a sample size of 1000, the slopes were similar to the slopes for the comparisons at a sample size of 2500 (given fewer data points), and decreased rapidly at a sample size of 0.2 of the maximum sample size (Fig. 2b). Values were similar at each relative percentage of the comparison sample (i.e. at 20% of the maximum sample size), even though they were re-set to a new comparison.

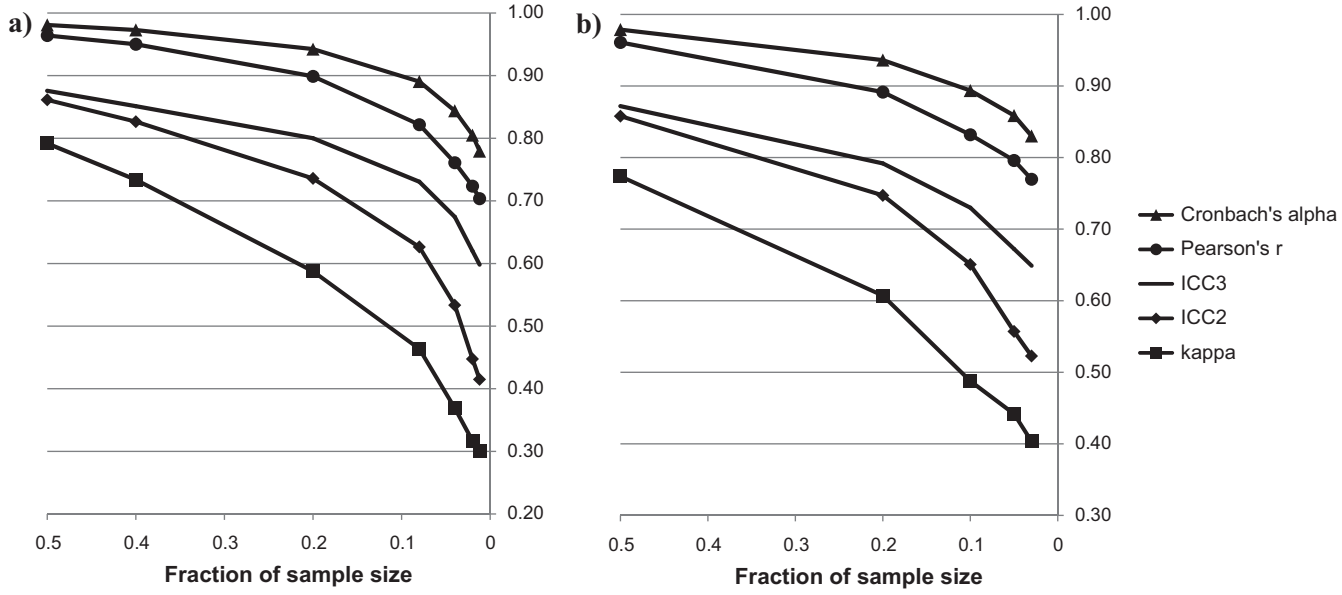

Fig. 2. Accuracy assessment metrics that compare decreasing sample sizes to (a) a maximum sample size of 2500 and (b) a maximum of 1000 for species distribution models covering about 5 million ha in Minnesota.

Table 1
Increase in mean predicted probabilities by species as sample size decreases (Minnesota Laurentian Mixed Forest).

	Sample size							
	2500	1250	1000	500	200	100	50	30
American basswood	0.46	0.51	0.53	0.57	0.59	0.61	0.65	0.66
Aspens	0.66	0.67	0.67	0.69	0.68	0.68	0.71	0.73
Balsam fir	0.42	0.47	0.49	0.54	0.61	0.64	0.68	0.68
Balsam poplar	0.38	0.44	0.45	0.51	0.58	0.58	0.57	0.64
Black ash	0.55	0.60	0.61	0.64	0.67	0.69	0.72	0.69
Black spruce	0.24	0.28	0.29	0.32	0.37	0.44	0.43	0.48
Bur oak	0.57	0.60	0.61	0.65	0.64	0.67	0.72	0.73
Jack pine	0.25	0.29	0.31	0.37	0.40	0.44	0.52	0.56
Maples	0.54	0.58	0.58	0.60	0.64	0.65	0.69	0.68
Northern white cedar	0.18	0.23	0.25	0.32	0.40	0.48	0.52	0.52
Paper birch	0.60	0.63	0.64	0.65	0.69	0.70	0.71	0.65
Quaking aspen	0.69	0.70	0.70	0.69	0.70	0.70	0.70	0.70
Red maple	0.54	0.58	0.59	0.62	0.64	0.62	0.66	0.69
Red pine	0.26	0.30	0.33	0.39	0.47	0.51	0.57	0.58
Sugar maple	0.38	0.45	0.46	0.49	0.51	0.51	0.58	0.62
Tamarack	0.27	0.31	0.33	0.36	0.38	0.44	0.37	0.41

However, values were greater than when compared to greater absolute sample size (i.e. metric values at 500 were greater when compared to 1000 than compared to 2500).

As sample size decreased, at a constant 75% threshold for accepting species presence, true positive rate decreased, to about 75% of the true positive rate value of the 2500 sample size, and area increased rapidly, to about 180% of the area of the 2500 sample size (Fig. 3). In addition, as sample size decreased, mean predicted probabilities increased (Fig. 3), but not by the same amount for all species (Table 1). Indeed, the distance between the two intraclass coefficient metrics (Fig. 2) may represent the difference in mean predicted probability values by sample size.

To substantiate our findings for sample size, we correlated (Pearson's *r*) samples sizes of 1000, 500, 200, 100, and 50 for black oak and scarlet oak in one ecological subsection that was 800,000 ha in the Missouri Ozarks (Fig. 4). The slope for correlation again appeared to decrease at about 10–20% of the best model that had 1000 samples. As a side note, we correlated the sample of 1000 from the subsection to sample of 1000 from the entire Missouri Ozarks (about 9 million ha). For black oak, the correlation was 0.62 and for scarlet oak, the correlation was 0.82, indicating that extent, and the incorporated variation, will influence species distribution models.

4. Discussion

The comparison metrics all showed similar trends and therefore selection of comparison metrics generally is a matter of preference and situation. The kappa statistic uses categories and generally

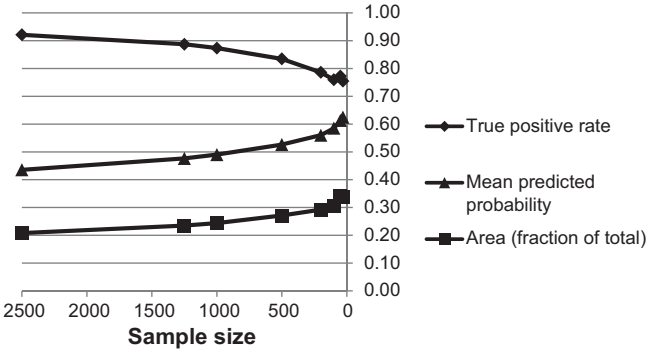

Fig. 3. True positive rate, mean predicted probability, and area (fraction of total 4,895,238 ha) for decreasing sample sizes of species distribution models in Minnesota.

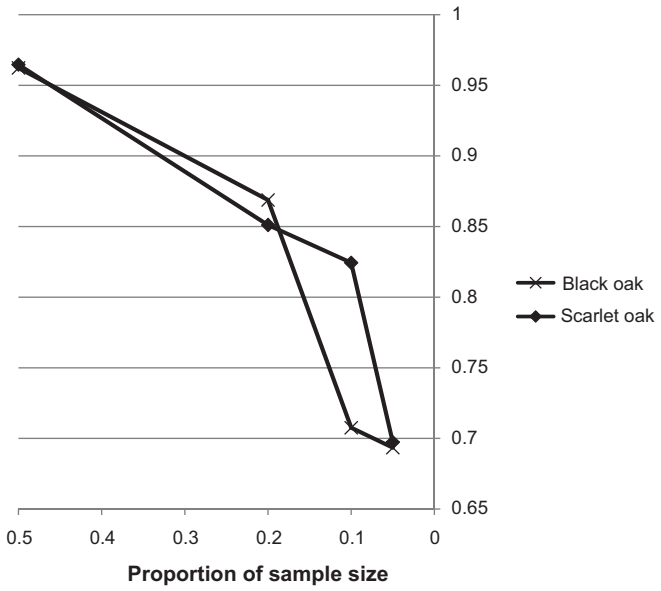

Fig. 4. Correlations that compare decreasing sample sizes (500, 200, 100, 50) to a maximum sample size of 1000 for black oak and scarlet oak species distribution models covering about 800,000 ha in Missouri.

it seems better to directly use predicted probabilities rather than lose information to categories; nevertheless, there may be circumstances when categories are desirable. Because the magnitude of difference in sources is an artifact of sample size (the source), Pearson's *r* and ICC2 should be better metrics for sample size comparisons, but if the magnitude of difference is important, then ICC3 and Cronbach's alpha may be more suitable metrics. Lastly, Pearson's *r* (and Cronbach's alpha) is faster and simpler to calculate than the intraclass metrics.

Comparison of sample sizes is a slippery slope that depends on the maximum sample size. When comparing to 2500 individuals, a sample size of 1000 is very comparable, 500 reasonably so, and there is a steep curve after reaching around 20% (500) to 10% (250) of the maximum sample size. When comparing to 1000 individuals, a sample size of 200 appears reasonably similar, and then again there is a steep curve below 0.2 of the maximum sample size. Therefore, authors who compare to a maximum sample size of 100–200 will find that sample sizes of 10–50 are comparable (e.g. Stockwell and Peterson, 2002; Kadmon et al., 2003; Hernandez et al., 2006).

Although we never reached an asymptote where increased sample size no longer affected predicted probabilities, we believe that sample sizes of 500 individuals and greater will be comparable to a sample size of 2500 based on a correlation of 0.9. At sample sizes between 500 and 200, models will have some uncertainty. Keeping in mind that every species is unique, models may not be very accurate below a sample size of 200, at least for our models and this area, spatial extent, and grain. We generalized the results to an ecological area in the Central Hardwoods of Missouri that was about 5% of the size, but with only two species and one area, the applicability of minimum sample size requirements is tentative. In any event, models based on small sample sizes may still be useful for identifying trends, although prediction uncertainties are high under small sample sizes (Wisz et al., 2008).

Mean predicted probability increased with decreasing sample size perhaps because there was not enough information to fully delineate sites where the species were present, resulting in greater distance between sites where the species were present and not present. To increase sample size, we suggest running the model with a portion reserved for accuracy assessment and then running the model again with the full dataset. Interestingly, decreasing

the modeling prevalence will reduce predicted probabilities (Lobo et al., 2007), which provides an option to reduce predicted probabilities when the values seem inflated compared to probabilities for more common species.

Our research was for a landscape of almost 5 million ha, similar to other studies where the extents were entire countries or states (e.g. Stockwell and Peterson, 2002; Kadmon et al., 2003; Hernandez et al., 2006). We were able to substantiate our sample size minimum of about 200 records for a smaller extent in a limited manner, due to the low density of plots and small number of trees in FIA surveys. However, we expect that even though the minimum sample size requirement may decrease when a smaller study area limits environmental variability, there will not be a large decrease because in order to define a species, a certain amount of variability needs to be captured by the sample. It also appears that species distribution models will change based on the extent and given the choice, the ideal extent for modeling probably depends on the scale of planned applications.

Further research should investigate whether smaller extents also need 200 or more records to produce a model similar to the gold standard. There even may be a sample size rule of thumb based on the relationship between sample size and extent, which may depend heavily on the spatial unit or grain as well. We modeled tree species and therefore minimum sample sizes for other taxa may be variable, due to differing strength of the environmental variables to make predictions.

Acknowledgements

We received funding from the National Fire Plan of the USDA Forest Service, Northern Research Station.

Appendix A. Correlation among predicted probabilities for varying sample sizes (Minnesota Laurentian Mixed Forest)

	2500	1250	1000	500	200	100	50	30
2500	1.00	0.96	0.95	0.90	0.82	0.76	0.72	0.70
1250	0.96	1.00	0.99	0.95	0.87	0.81	0.78	0.75
1000	0.95	0.99	1.00	0.96	0.89	0.83	0.80	0.77
500	0.90	0.95	0.96	1.00	0.94	0.89	0.85	0.82
200	0.82	0.87	0.89	0.94	1.00	0.95	0.91	0.87
100	0.76	0.81	0.83	0.89	0.95	1.00	0.95	0.91
50	0.72	0.78	0.80	0.85	0.91	0.95	1.00	0.96
30	0.70	0.75	0.77	0.82	0.87	0.91	0.96	1.00

References

Albert, D., 1995. Regional Landscape Ecosystems of Michigan, Minnesota, and Wisconsin: A Working Map and Classification. USDA Forest Service, North Central Forest Experiment Station.
Beers, T.W., Dress, P.E., Wensel, L.C., 1966. Aspect transformation in site productivity research. Journal of Forestry 64, 691–692.
Breiman, L., 2001. Random forests. Machine Learning 40, 5–32.
Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20, 37–46.
Cutler, D.R., Edwards Jr., T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2007. Random forests for classification in ecology. Ecology 88, 2783–2792.
ECOMAP, 1993. National Hierarchical Framework of Ecological Units. USDA Forest Service, Washington, DC.
Guisan, A., Zimmermann, N.E., Elith, J., Graham, C.H., Phillips, S., Peterson, A.T., 2007. What matters for predicting the occurrences of trees: techniques, data, or species' characteristics? Ecological Monographs 77, 615–630.
Hernandez, P.A., Graham, C.H., Master, L.L., Albert, D.L., 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29, 773–785.
Jiménez-Valverde, A., Lobo, J.M., Hortal, J., 2008. Not as good as they seem: the importance of concepts in species distribution modelling. Diversity of Distributions 14, 885–890.
Kadmon, R., Farber, O., Danin, A., 2003. A systematic analysis of factors affecting the performance of climatic envelope models. Ecological Applications 13, 853–867.
Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R News 2, 18–22.
Lobo, J.M., Jiménez-Valverde, A., Real, R., 2007. AUC: a misleading measure of the performance of predictive distribution models. Global Ecology and Biogeography 17, 145–151.
McGraw, K.O., Wong, S.P., 1996. Forming inferences about some intraclass correlation coefficients. Psychological Methods 1, 30–46.
McPherson, J.M., Jetz, W., Rogers, D.J., 2004. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artifact? Journal of Applied Ecology 41, 811–823.
R Development Core Team, 2010. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.
Sappington, J.M., Longshore, K.M., Thompson, D.B., 2007. Quantifying landscape ruggedness for animal habitat analysis: a case study using bighorn sheep in the Mojave Desert. The Journal of Wildlife Management 71, 1419–1426.
Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin 86, 420–428.
Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T., 2005. ROCR: visualizing classifier performance in R. Bioinformatics 21, 3940–3941.
Stockwell, D.R.B., Peterson, A.T., 2002. Effects of sample size on accuracy of species distribution models. Ecological Modelling 148, 1–13.
Syphard, A.D., Franklin, J., 2010. Species traits affect the performance of species distribution models for plants in southern California. Journal of Vegetation Science 21, 177–189.
Vaughan, I.P., Omerod, S.J., 2005. The continuing challenges of testing species distribution models. Journal of Applied Ecology 42, 720–730.
Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., Guisan, A., NCEAS Predicting Species Distributions Working Group, 2008. Effects of sample size on the performance of species distribution models. Diversity of Distributions 14, 763–773.