Assessing Effective Sampling Method and Sample Size for Species Distribution Modeling of Korean Red Pine (Pinus densiflora)

SUNG, Sun-Yong
Interdisciplinary Program in Landscape Architecture, Seoul National University, Seoul | Laboratory of Forest Resources Management, Division of Forest Environmental Sciences, Department of Agro-Environmental Sciences, Faculty of Agriculture, Kyushu University

LEE, Dong-Kun
Department of Landscape Architecture and Rural Systems, Engineering, Seoul National University | Laboratory of Forest Resources Management, Division of Forest Environmental Sciences, Department of Agro-Environmental Sciences, Faculty of Agriculture, Kyushu University

PARK, Chan
Department of Landscape Architecture, University of Seoul | Laboratory of Forest Resources Management, Division of Forest Environmental Sciences, Department of Agro-Environmental Sciences, Faculty of Agriculture, Kyushu University

KIM, Ho-Gul
Division of Human Environmental Design, Cheongju University | Laboratory of Forest Resources Management, Division of Forest Environmental Sciences, Department of Agro-Environmental Sciences, Faculty of Agriculture, Kyushu University

他

https://doi.org/10.5109/1955384

出版情報：九州大学大学院農学研究院紀要. 63 (2), pp.211-221, 2018-09-01. Faculty of Agriculture, Kyushu University
バージョン：
権利関係：

J. Fac. Agr., Kyushu Univ., **63** (2), 211–221 (2018)

Assessing Effective Sampling Method and Sample Size for Species Distribution Modeling of Korean Red Pine (*Pinus densiflora*)

Sun–Yong SUNG[1], Dong–Kun LEE[2], Chan PARK[3], Ho–Gul KIM[4], Sung–Ho KIL[5], Hee–Mun CHAE[6], Gwan–Soo PARK[7] and Shoji OHGA*

Laboratory of Forest Resources Management, Division of Forest Environmental Sciences,
Department of Agro–Environmental Sciences, Faculty of Agriculture,
Kyushu University, Fukuoka 811–2415, Japan

(*Received April 18, 2018 and accepted May 8, 2018*)

Sampling method and sample size can alter the performance of species distribution models (SDMs). In this study, we identified an effective sampling method and sample size for modeling Korean red pine (*Pinus densiflora*). We used 3 sampling methods (simple random sampling, stratified sampling, and area–weighted sampling), 7 different sample sizes (30, 50, 100, 200, 500, 1000, and 3000), and 8 SDMs (GLM, GAM, CTA, ANN, GBM, RF, FDA, and MAXENT). The performance of each model was evaluated using the area under the receiver operating characteristic curve. Differences among the models were validated using ANOVA. We found that the area–weighted sampling method was the most effective and stable. As sample size increased, model performance increased in the random and stratified sampling methods. However, performance became saturated as sample size exceeded 200 in the area–weighted sample due to spatial autocorrelation among samples. All models exhibited different levels of performance. The RF and GBM models exhibited the highest performance (AUC = 0.838 and 0.839, respectively), while the ANN model exhibited the lowest performance (AUC = 0.658). Therefore, sampling method and sample size should be carefully considered when selecting SDMs depending on the objective of the study.

Key words: Area–weighted sampling; BIOMOD2; Analysis of variance; Korean red pine

INTRODUCTION

Recently, climate change has begun to influence the distribution of species in space and time (Kelly and Goulden, 2008; Pederson *et al*., 2015), and these spatial and temporal changes in the distribution of species significantly affect ecosystems (Walther *et al*., 2002; Thuiller *et al*., 2006). To assess the potential impact of climate change, empirical and mechanistic models are commonly used. Mechanistic models use physiological or ecological knowledge to model the potential distribution of species under a set of environmental conditions, including competition and establishment amongst others. However, mechanistic models are based on coarse classification trees with heavy parameterization of variable relationships (Ahlström *et al*., 2015; Case and Lawler, 2016). Results from mechanistic models are typically based on plant functional types (PFTs), which are not enough to establish conservation plans or management efforts (Adams *et al*., 2004; Sato *et al*., 2007). Thus, empirical modeling is an effective tool for modeling target species (Park *et al*., 2016).

SDMs are one of the most efficient types of empirical models for understanding the potential impact of anthropogenic activities, as well as natural changes, on species distribution (Elith and Leathwick, 2009). In SDMs, presence data play an important role as the fundamental concept of SDMs is based on the niche of a species within a range of environmental parameters (Beale and Lennon, 2012). Most studies use presence data obtained from the literature or from field surveys owing to logistical and financial constraints. However, without precise presence data, there is no guarantee of the accuracy of the potential distribution of target species predicted by SDMs (Hannemann *et al*., 2016).

SDMs consist of complex procedures that can introduce errors (Araújo and Guisan, 2006; Wiens *et al*., 2009). For example, models run with different sample sizes can have different results (Hernández *et al*., 2006), even under the same environmental conditions. If a sampling method is not properly designed and samples properly selected, the reliability of SDM results decreases (Hirzel and Guisan, 2002). Sampling method is an important factor that affects the performance of SDMs (Zimmermann *et al*., 2010); generally, it employs random sampling of combined datasets. A few studies have examined how samples are selected from vegetation datasets (Shiver and Borders, 1996; Mandallaz, 2008), but sampling from these datasets has different requirements than sampling from animal datasets does since vegetation cannot easily move between habitats in response to environmental

[1] Interdisciplinary Program in Landscape Architecture, Seoul National University, Seoul, South Korea, 08826

[2] Department of Landscape Architecture and Rural Systems Engineering, Seoul National University, Seoul, South Korea, 08826

[3] Department of Landscape Architecture, University of Seoul, Seoul, South Korea, 02504

[4] Division of Human Environmental Design, Cheongju University, South Korea, 28503

[5] Department of Ecological Landscape Architecture Design, Kangwon National University, South Korea, 24341

[6] Department of Forest Environment Protection, Kangwon National University, South Korea, 24341

[7] Department of Forest Resources, Chungnam National University, South Korea, 34167

* Corresponding Author (E–mail: ohga@forest.kyushu–u.ac.jp)

changes. Thus, various sampling methods, sample sizes, and type of SDMs should be carefully evaluated when conducting species distribution modeling on vegetation.

This study had two main objectives: first, to determine an effective sampling method and sample size for modeling the species distribution of Korean red pine (*Pinus densiflora*) and second, to examine how sampling method and sample size affect the performance of each type of SDM. Finally, we discussed the implications that changes in sampling method, sample size, and choice of SDMs have on model performance when applying SDMs to vegetation.

MATERIALS AND METHODS

We conducted our study in South Korea (Fig. 1). More than 60% of South Korea consists of forest cover, which ranks it fourth for total forest area of the OECD countries (OECD, 2011). Most of the forested areas are in northeastern and southern South Korea on high mountains. Forests in South Korea consist of evergreen needle–leaved coniferous forests, deciduous broad–leaved forests, and mixed forests, representing 38%, 33%, and 28% of total forest area, respectively (Korea Forest Service, 2016).

We selected Korean red pine (*Pinus densiflora*) as the target species. This tree is an important species both ecologically and culturally. Korean red pine usually prefers areas below 1300 m altitude with high amounts of radiation. Moreover, almost 95% of the needle–leaved coniferous forest in South Korea are made up Korean red pine. Therefore, modeling the distribution of Korean red pine is directly related to understanding the potential interaction between environmental changes and forest ecosystems.

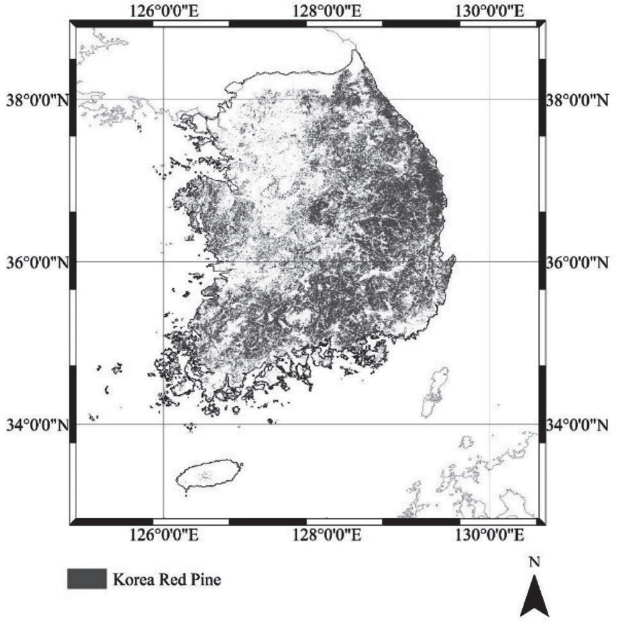


Fig. 1. The distribution of Korean red pine (*Pinus densiflora*) in South Korea.

We selected 360,000 natural Korean red pine forest stands from a 1:5000 forest inventory map from the Korea Forest Service (KFS). The minimum size of natural forest stands was 0.5 ha, and the minimum forest density was 10% forest cover. In the forest inventory map, a forest type was classified—using data from field surveys and aerial images—if a forest stand consisted of one species with a minimum coverage of 75% which resulted in 44 forest types.

We compared 3 sampling methods: random sampling, stratified sampling, and area–weighted sampling (Fig. 2). Using 1:5000 forest inventory map, natural forests and artificial forests were divided according to a

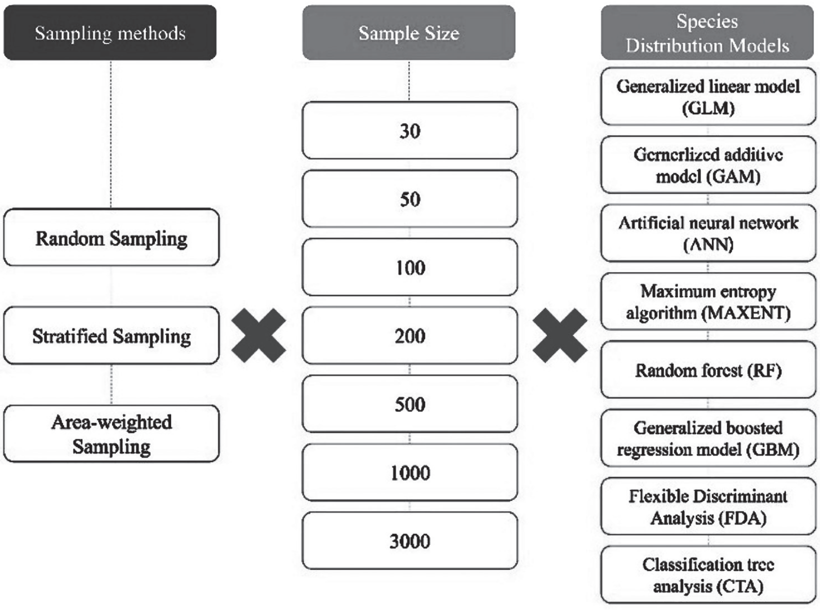

Sampling methods
Random Sampling
Stratified Sampling
Area-weighted Sampling
Sample Size
30
50
100
200
500
1000
3000
Species Distribution Models
Generalized linear model (GLM)
Gernerlized additive model (GAM)
Artificial neural network (ANN)
Maximum entropy algorithm (MAXENT)
Random forest (RF)
Generalized boosted regression model (GBM)
Flexible Discriminant Analysis (FDA)
Classification tree analysis (CTA)

Fig. 2. Study design for evaluating the effect of sampling method, sample size, and species distribution model type on predicted distributions.

Table 1. List of environmental variables

Category	Variables	Data type	Resolution	Source
Vegetation	Forest Type	Feature (SHP)	1:5000	Korea Forest Service
Topography	Altitude	Raster (TIFF)	30 m	National Geography Information Institute
	Slope			
	Radiation			Ministry of Environment
	Distance from water			
	Distance from sea			
Soil	Soil Depth	Feature (SHP)	1:25000	Korea Forest Service
	Soil Organic Matter Content in Layer A			
Climate	Warmth index	Raster (TIFF)	1 km × 1 km	Korea Metrological Administration
	Isothermality			
	Min temperature of coldest month			
	Precipitation of wettest month			
	Precipitation of driest month			
	Climate Zone		30 m	Korea Forest Service

land registration map, afforestation and reforestation map, aerial photos and field survey data by KFS (Korean Forest Research Institute, 2012). To measure the influence of sample size on SDMs, we first selected a random point in a forest stand of the same age, density, and diameter class as a Korean red pine stand. Then, we collected either 30, 50, 100, 200, 500, 1000, or 3000samples from the randomly made sample point with ArcGIS 10.1. For the random sampling method, we randomly selected sample points within the previously created sampling area. For stratified sampling, we selected samples from strata based on the proportion of each climatic zones in each area. Area–weighted sampling was conducted by selecting points from the largest area stands in decreasing order until the number of samples equaled the number of points selected. For example, if we collected 30 samples, we took one each from the 30 largest Korean red pine stands.

We used four environmental coverages—vegetation, climate, topography, and soil (Table 1)—in the SDMs. Vegetation data were based on a 1:5000 detailed forest map from the KFS. Forest age, type, and density were collected for each forest stand. Climate data were derived from the Korea Metrological Administration (KMA), and all climate datasets were statistically downscaled to a 1 km resolution. From the monthly temperature and precipitation data, we generated 19 bioclimatic variables averaged over 10 years, from 2001 to 2010, to match the WorldClim database (Hijmans *et al*., 2005). We also constructed a warmth index (WI) and a coldness index (CI), which are considered efficient indicators for monitoring interactions between climate and species distribution (Kira 1945; Yim 1977) (Equations 1, 2). The WI was calculated for months in which the temperature (t) was greater than 5°C, and the CI was calculated for months in which the temperature was less than 5°C. We used climate zone data derived from KFS for selecting strata for the stratified sampling method.

$$\text{Warmth index (WI)} = \sum (t - 5) \qquad (1)$$

For months in which $t > 5$°C

$$\text{Coldness index (CI)} = -\sum (5 - t) \qquad (2)$$

For months in which $t < 5$°C

Topographical layers included altitude, slope, and aspect. These datasets were derived from digital elevation models (DEMs) from the National Geography Information Institute (NGII). A land cover map from the Ministry of Environment (ME) was used to extract land cover data. Distance from water and distance from the sea (Schulze *et al*., 2005) were calculated using Euclidian distance. Soil depth and soil organic matter content in the A–horizon were extracted from a Korean soil forest map (Brady 2008). All environmental data were resampled with 1 km by 1 km resolution for modeling with the ArcGIS resampling tool. Environmental variables which have discrete characteristics were resampled using the nearest algorithm. Other environmental variables were resampled using bilinear resampling methods.

The package Biomod2 (version 3.1.64) in R (version 3.1.2) was used to model the distribution of Korean red pine (Thuiller *et al*., 2009; R Core Team 2014), which enabled us to run 10 cutting–edge species distribution modeling techniques to describe and model the relationships between Korean red pine and its environment. Biomod2 uses the ecological niche of a particular species based on environmental variables, such as temperature, precipitation, and altitude, to project potential habitat based on current or future environmental variables (Thuiller *et al*., 2015).

There are two categories of SDMs: statistically–based models and machine learning–based models (Table 2).

Table 2. Characteristics of 8 species distribution models and their relative performance (revised from tables in (Franklin 2010b; Thuiller *et al.*, 2010; Kim *et al.*, 2015))

Category	Model	Characteristics	Performance
Statistical	Generalized linear model (GLM)	Flexible mordern regression models	Effective global modeling methods
	Gernerlized additive model (GAM)	Multiple regression but with curve fitting splines or other methods	Performs slightly better than GLM
Machine learning based	Artifi1cial neural network (ANN)	Nonlinear model Using concept of artificial neural network	Performance sometimes worse than statistical model
	Maximum entropy algorithm (MAXENT)	Nonlinear model Using concept of maximum entropy Validated by ROC curve	Perfoms well in data–poor situation
	Random forest (RF)	Estimate a large numver of tree models based on subset of data and averaging result	Ensemble of decision tree have good performance
	Generalized boosted regression model (GBM)		
	Flexible Discriminant Analysis (FDA)	Classification based on mixture models	–
	Classification tree analysis (CTA)	Divisive model	Single decision trees performs poorly

Generalized linear models (GLMs), generalized additive models (GAMs), and multivariate adaptive regression splines (MARS) are all statistically–based models. Machine learning–based models include the generalized boosted regression model (GBM), classification tree analysis (CTA), artificial neural network (ANN), a rectilinear envelope similar to BIOCLIM (SRE), flexible discriminant analysis (FDA), random forest (RF), and maximum entropy (MAXENT). Of these, 8 models were used for the analysis (the SRE and MARS models were excluded as they cannot handle categorical variables).

We conducted a correlation analysis in R to identify auto–correlation among the environmental variables, which were selected with respect to multi–collinearity. If Pearson correlation coefficients were larger than 0.7, we removed relevant variables from the list (Dormann *et al.*, 2013). We also conducted a literature review to select variables potentially important for the distribution of Korean red pine (Takahashi and Okuhara 2012; Park *et al.*, 2016; Nakao *et al.*, 2014).

To analyze the performance of SDMs, we used the area under the receiver operating characteristic (ROC) curve. The ROC curve is an effective method to determine the relationship between the false positive fraction (1–specificity) and the sensitivity for a range of thresholds. A good model has a curve that maximizes sensitivity for low values of 1– specificity (Neovius *et al.*, 2004). The area between the 1:1 line and the curve represents the performance of the model, and this value is called the area under the curve (AUC). Additionally, AUC is an effective model evaluation index and is independent of prevalence (Franklin 2010a). We considered AUC values between: 0.9–1.0, excellent; 0.8–0.9, very good; 0.7–0.8, good; 0.6–0.7, average; and 0.5–0.6, poor (Hansson *et al.*, 2005). An analysis of variation (ANOVA) was conducted using SPSS 18.0 to test the differences in AUC among the sampling methods and SDMs (SPSS Inc 2009).

RESULTS AND DISCUSSION

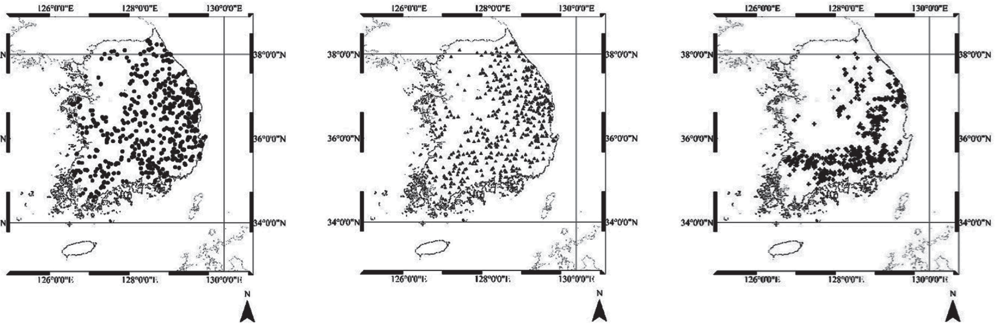
Sampling method caused differences in model performance (Table 3). The area–weighted sampling method performed better than the stratified sampling and random sampling methods. The average AUC value for models based on area–weighted sampling was 0.777, which was considered good, while stratified sampling and random sampling had AUC values of 0.663 and 0.622, respectively, and were thus considered average. Additionally, area–weighted sampling demonstrated stable performance, even across different sample sizes, as shown by the standard deviation of model performance.

The distribution of samples exhibited different spa-

Table 3. Average performance (AUC) of SDMs by sampling method for all sample sizes and SDM types as determined by an analysis of variance

Sampling methods	Average AUC	Standard deviation	F–Value1	Post–Hoc Analysis2
Random sampling	0.622	0.050	1436.904	Random Sampling < Stratified sampling, Area–weighted sampling
Stratified sampling	0.663	0.070		Stratified Sampling > Random Sampling Stratified Sampling < Area–weighted Sampling
Area–weighted Sampling	0.777	0.087		Area–weighted sampling > Random sampling, Stratified sampling

1 * P<0.0001, 2 Games–Howell test. P<0.05



(a) Random sampling (b) Stratified sampling (c) Area-weighted sampling

Fig. 3. Distribution of samples used in species distribution models by sampling method. Each method was used to select 500 samples.

tial patterns based on sampling method (Fig. 3). In the random sampling method, the samples were dispersed in an area distant from the primarily mountainous areas of South Korea. However, samples that were selected using area–weighted sampling were located near these mountainous areas. These differences resulted in varying performance among the sampling methods. We also found that models employing random sampling did not perform well compared to those using stratified sampling, which is consistent with previous studies (Hirzel & Guisan, 2002).

When we analyzed the altitude of each collected sample by sampling method, we found that the area–weighted sampling method showed a different altitude distribution than the random sampling and stratified sampling methods did. Anything from 0–500 m altitude was considered to be a suitable habitat for Korean red pine (Lee and Jo 2003). The area–weighted sampling method showed 88% of its samples within this 0–500 m range, while the random sampling and stratified sampling methods had 80% and 79% of their samples within this range, respectively. This difference led us to include the ecological preferences of Korean red pine which can affect model performance.

Sample size causes differences in modeling performance in all sampling methods (Table 4). Because of

Table 4. Average performance of SDMs by sample size (AUC) and ANOVA result in all SDMs

Sampling methods	Sample size	Average	Standard deviation	F–Value1	Post–Hoc Analysis2
Random sampling	30s	0.610	0.121	40.991	30s < 200s, 500s, 1000s, 3000s ; 30s > 100s
	50s	0.585	0.089		50s < 200s, 500s, 1000s, 3000s
	100s	0.570	0.066		100s < 30s, 100s, 200s, 500s, 1000s, 3000s
	200s	0.642	0.066		200s > 30s, 50s, 100s
	500s	0.643	0.052		500s > 30s, 50s, 100s
	1000s	0.648	0.046		1000s > 30s, 50s, 100s
	3000s	0.653	0.043		3000s > 30s, 50s, 100s
Stratified sampling	30s	0.624	0.104	24.149	30s < 100s, 200s, 500s, 1000s, 3000s
	50s	0.639	0.075		50s < 200s, 500s, 1000s, 3000s
	100s	0.660	0.071		100s > 30s; 100s < 200s, 3000s
	200s	0.681	0.058		200s > 30s, 50s, 100s
	500s	0.669	0.051		500s > 30s, 50s; 500s < 3000s
	1000s	0.675	0.045		1000s > 30s, 50s; 1000s < 3000s
	3000s	0.693	0.043		3000s > 30s, 50s, 100s, 500s, 1000s
Area–weighted sampling	30s	0.762	0.129	4.633	30s < 200s
	50s	0.779	0.116		–
	100s	0.786	0.083		100s > 3000s
	200s	0.798	0.077		200s > 30s, 1000s, 3000s
	500s	0.782	0.058		500s > 3000s
	1000s	0.775	0.060		1000s < 200s
	3000s	0.760	0.045		3000s < 200s, 500s

1 * P<0.0001, 2 Games–Howell test. P<0.05

Table 5. Standard deviation of environmental variables in each sampling size (Area–weighted sampling)

Environmental Variables*	Number of Samples in Area Weighted Sampling						
	30	50	100	200	500	1000	3000
Altitude	141.806	152.128	147.236	163.092	163.467	169.976	180.745
Slope	9.679	8.845	8.409	8.892	9.174	9.426	9.551
Radiation	159.601	164.217	165.517	153.93	146.244	391.282	297.291
Distance from water	827.105	743.953	764.29	738.759	807.033	826.316	847.709
Distance from sea	21,112.97	20,960.28	21,147.27	20,613.09	21,549.58	22,184.47	23,109.21
Soil Depth	80.918	86.526	84.167	83.048	86.047	89.189	91.159
Warmth Index	8.887	9.033	10.243	12.271	12.533	14.52	14.997
Isothermally	1.214	1.417	1.415	1.485	1.607	2.565	2.113
Min temperature of Coldest month	1.46	1.424	1.786	2.077	2.149	2.309	2.451
Precipitation of Wettest month	71.568	68.176	66.438	63.386	65.737	71.361	68.775
Precipitation of driest month	3.848	4.104	3.859	3.946	3.877	3.784	3.653

*Categorical variables are excluded

ANOVA, AUC value significantly increased as sample size increased in random sampling and stratified sampling. But in area–weighted sampling, the AUC value did not significantly increase as sample size increased.

We found that sample size is a critical factor in deciding the reliability of species distribution modeling. Generally, a larger sample size increased model performance (Wisz *et al.*, 2008; Moudrý and Šímová 2012). However, our result showed that as the sampling size increased, the AUC of a given model did not respond in linear form (Table 5). As sampling size increased, the correlation coefficient decreased in log form. Thus, increases in sampling size after a certain number, which was 200 samples in this study, do not increase model performance significantly. When sample size increases, the standard deviation of environmental variables also increases. This could cause a different ecological niche that decreases model performances.

We analyzed the model performance of different SDMs. Average AUC values were higher than 0.7, except for ANN, and AUC values for the GAM, RF, GLM, GBM, and GAM models were higher than 0.8. The RF and GBM models performed well with AUC values of 0.839 and 0.838, respectively, most likely due to the similarities in the algorithms of these two models. Conversely, the AUC for ANN was the lowest of the 8 SDMs (AUC = 0.657). ANN, CTA, and MAXENT models failed in several cases (Fig. 4).

The differences among the 8 SDMs were significant ($P < 0.0001$; Table 6). In the post–hoc analysis, the differences between some of the models were not significant, e.g., the differences between RF and GBM were not significant, and there were no significant differences among FDA, GAM, and GLM. However, the ANN, CTA, and MAXENT models differed significantly from other SDMs.

We found that the similarities between the model algorithms led to similar levels of performance from these models. We found that decision tree–based models (GBM, RF) with ensemble had benefits for modeling the potential distribution of Korean red pine compared to a single decision tree model (CTA). The GBM and RF were both designed to resolve problems with them which were the greedy characteristics of the algorithm (C. Park and Kyong 2003) and the fact that they are over–sensitive to the training data set. Statistical models (GAM, GLM) showed performed well. As the GAM allows for more flexible interaction between variables as functions for GAM can be parametric, non–parametric or splines, the average performance of GAM showed 0.06 higher than GLM but the difference was not significant.

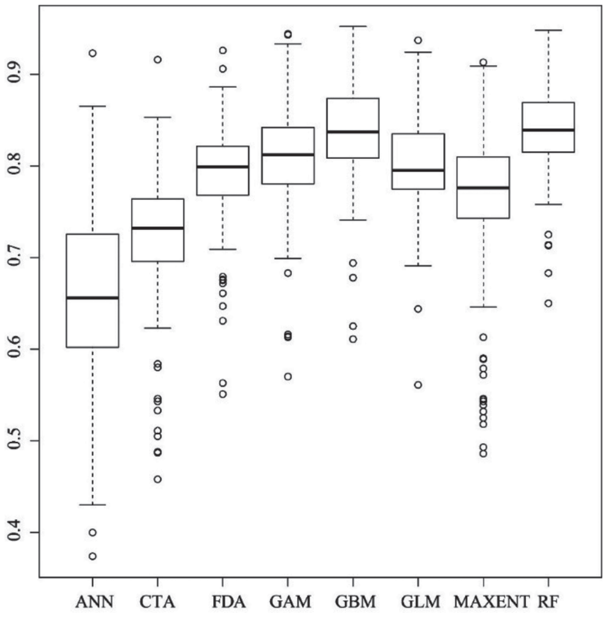


Fig. 4. Box–whisker plot of the performance (AUC) of 8 species distribution models based on the area–weighted sampling method for all sample sizes. (To see the results for all sampling methods and sizes, please see Supplementary Fig. 1–3.)

Table 6. Differences in performance among species distribution model (SDM) algorithms based on an area–weighted sampling method for all sample sizes as determined by an analysis of variance

SDMs	Average	Standard Deviation	F–Value[1]	Post–hoc Analyis[2]
ANN	.658	.088	131.688*	ANN < CTA, FDA, GAM, GBM, GLM, MAXENT, RF
CTA	.725	.071		CTA > ANN CTA < FDA, GAM, GLM, MAXENT, RF
FDA	.791	.053		FDA > ANN, CTA, MAXENT FDA < GBM, RF
GAM	.809	.057		GAM<GBM, MAXENT, RF GAM > ANN,CTA
GBM	.839	.052		GBM > ANN, CTA, FDA, GAM, GLM, MAXENT
GLM	.803	.051		GLM < GBM, RF GLM > ANN, CTA, MAXENT
MAXENT	.760	.089		MAXENT < FDA, GAM, GBM, GLM, RF MAXENT > ANN, CTA
RF	.840	.045		RF > ANN, CTA, FDA, GAM, GLM, MAXENT

The ANN results showed poor performance as the ANN model may lead to overfitting of the data set, which can negatively impact results (Tu 1996).

CONCLUSION

We found that the area–weighted sampling method was an effective tool for applying SDMs to Korean red pine. In addition, we found that selecting suitable sample sizes for SDMs can save time and resources in gathering presence data. In developing countries, a surveying presence dataset throughout the country requires an enormous amount of time and effort. If we can apply effective sampling methods and determine effective sample sizes as demonstrated by this study, we can estimate potential species distribution under recent climate change in time to make adaptation plans to protect ecosystem.

Many kinds of SDMs are used for modeling species distributions. Due to the complexity of these models, it is important to understand the uncertainties inherent in each model. Recent studies (Case and Lawler 2016; Hill *et al.*, 2017) have used two–stage modeling or hybrid modeling techniques to overcome these uncertainties. Species niches may affect model performance because the variation in climatic and environmental variables interact differently (Buisson *et al.*, 2010). Additionally, most models, except Maxent, use pseudo–absence data; thus, true absence data should also be carefully examined. Using environmental factors in SDMs requires further study, as we do not comprehensively understand the interactions among these factors in the context of models. These uncertainties can then cause uncertainties in policy and the decision–making process during planning and conservation.

AUTHOR CONTRIBUTIONS

S. Y. Sung designed the research and wrote whole manuscript. D. K. LEE designed the research and supervised whole manuscript. H. G. KIM contributed designing modeling process. C. PARK and S. H. KIL developed experiment design and discussion for manuscript. H. M. CHAE and G. S. PARK revised manuscript and discussion. and S. OHGA designed the study, supervised the research. All authors assisted in editing the manuscript and approved the final version.

ACKNOWLEDGEMENT

The authors declare no conflict of interest. This work supported by the BK21 Plus Project in 2016 (Seoul National University Interdisciplinary Program in Landscape Architecture, Global Leadership Program toward Innovative Green Infrastructure); This study was supported by 2017 Research Grant From Kangwon National University.

REFERENCES

Adams B., A. White and T. M. Lenton 2004 An analysis of some diverse approaches to modelling terrestrial net primary productivity. *Ecol Modell*. **177**(3–4): 353–391

Ahlström A, J. Xia, A. Arneth, Y. Luo and B. 2015 Smith Importance of vegetation dynamics for future terrestrial carbon cycling. *Environ Res Lett*. **10**(5): 54019

Araújo M. B. and Guisan A. 2006 Five (or so) challenges for species distribution modelling. *J Biogeogr*. **33**(10): 1677–1688

Beale C. M. and J. J. Lennon 2012 Incorporating uncertainty in predictive species distribution modelling. *Philos Trans R Soc B Biol Sci*. **367**(1586): 247–258

Brady N. C. 2008 The Nature and Properties of Soils. Upper Saddle River, N.J., Prentice Hall

Buisson L., W. Thuiller, N. Casajus, S. Lek and G. Grenouillet 2010 Uncertainty in ensemble forecasting of species distribution. *Glob Chang Biol*.,**16**(4): 1145–1157

Case M. J. and J. J. Lawler 2016 Integrating mechanistic and empirical model projections to assess climate impacts on tree species distributions in northwestern North America. *Glob Chang Biol*. **23**(5): 2005–2015

Dormann C. F., J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carreé, J.R. Marquéz, B. Gruber, B. Lafourcade, P.J. Leitão, T. Münkemüller, C. McClean, P. E. Osborne, B. Reineking, B. Schröder, A.K. Skidmore, D. Zurell and S. Lautenbach 2013

Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*. **36**(1): 27–46

Elith J. and J. R. Leathwick 2009 Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annu Rev Ecol Evol Syst*. **40**: 677–697

Franklin J. 2010a Mapping species distributions. Spatial inference and prediction. *Ecol Biodivers Conserv*. **53**(9): 340

Franklin J. 2010b Mapping Species Distributions: Spatial Inference and Prediction. Cambridge: Cambridge University Press

Hannemann H., K. J. Willis and M. Macias–Fauria 2016 The devil is in the detail: unstable response functions in species distribution models challenge bulk ensemble modelling. *Glob Ecol Biogeogr*. **25**(1): 26–35

Hansson S. L., A. Svanströmröjvall, M. Rastam, C. Gillberg, C. Gillberg and H. Anckarsäter 2005 Psychiatric telephone interview with parents for screening of childhood autism – Tics, attention–deficit hyperactivity disorder and other comorbidities (A–TAC): Preliminary reliability and validity. *Br J Psychiatry*. **187**(3): 262–267

Hernández P. A., C. H. Graham, L. L. Master and D. L. Albert 2006 The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*. **29**(5): 773–785

Hijmans R. J., S. E. Cameron, J. L. Parra, P.G. Jones and A. Jarvis 2005 Very high resolution interpolated climate surfaces for global land areas. *Int J Climatol*. **25**(15): 1965–1978

Hill L., A. Hector, G. Hemery, S. Smart, M. Tanadini and N. Brown 2017 Abundance distributions for tree species in Great Britain: A two–stage approach to modeling abundance using species distribution modeling and random forest. *Ecol Evol*. **7**(4): 1043–1056

Hirzel A and A. Guisan 2002 Which is the optimal sampling strategy for habitat suitability modelling. *Ecol Modell*. **157**(2–3): 331–341

Kelly A. E. and M. L. Goulden 2008 Rapid shifts in plant distribution with recent climate change. *Proc Natl Acad Sci U S A*. **105**(33): 11823–11826

Kim H. G., D.K. Lee, C. Park, S. Kil, Y. Son and J.H. Park 2015 Evaluating landslide hazards using RCP 4.5 and 8.5 scenarios. *Environ Earth Sci*. **73**(3): 1385–1400

Kira T. 1945 A New Classification of Climate in Eastern Asia as the Basis for Agricultural Geography. Kyoto: Horticulturla Institute, Kyoto University

Korea Forest Service. 2016 Statistical Yearbook of Forestry. Daejeon: Korea Forest Service

Korean Forest Rearch Institute. 2012 Developement of Manufacturing and Renewaling Methods for Large Scale Forest Inventory Map by Using Digital Airbone Image. Seoul

Lee C. and B. Jo 2003 Pine, Pine Forest. Seoul: National Institute of Forest Science

Mandallaz D. 2008 Sampling Techniques for Forest Inventories. Boca Raton, FL: Chapman & Hall/CRC

Moudrý V. and P. Šímová 2012 Influence of positional accuracy, sample size and scale on modelling species distributions: a review. *Int J Geogr Inf Sci*. **26**(11): 2083–2095

Nakao K., M. Higa, I. Tsuyama, C.T. Lin, S.T. Sun, J.R. Lin, C.R. Chiou, T.Y. Chen, T. Matsui and N. Tanaka 2014 Changes in the potential habitats of 10 dominant evergreen broad–leaved tree species in the Taiwan–Japan archipelago. *Plant Ecol*. **215**(6): 639–650

Neovius M. G., Y. M. Linné, B. S. Barkeling and S. O. Rossner 2004 Sensitivity and specificity of classification systems for fatness in adolescents. *Am J Clin Nutr*. **80**(3): 597–603

OECD. 2011 Forest area as share of land area, OECD countries, BRIICS, 2008, 1990, in Towards Green Growth: Monitoring Progress. OECD Publishing

Park C. and K. Y. Kyong 2003 A Decision Tree Algorithm using Genetic Programming. *Korean Commun Stat*. **10**(3): 845–857

Park S. U., K. A. Koo and W. S. Kong 2016 Potential Impact of Climate Change on Distribution of Warm Temperate Evergreen Broad–leaved Trees in the Korean Peninsula. *J Korean Geogr Soc*. **51**(2): 201–217

Pearson, R. G., W. Thuiller, M. B. Araújo, E. Martinez–Meyer, L. Brotons, C. McClean, L. Miles, P. Segurado, T. P. Dawson and D. C. Lees. 2006 Model–Based Uncertainty in Species Range Prediction *J Biogeogr*. **33**(10): 1704–1711

Pederson, N., A. W. D'Amato, J. M. Dyer, D. R. Foster, D. Goldblum, J. L. Hart, A. E. Hessl, L. R. Iverson, S. T. Jackson, D. Martin–Benito, B. C. McCarthy, R. W. McEwan, D. J. Mladenoff, A. J. Parker, B. Shuman and J. W. Williams 2015 Climate remains an important driver of post–European vegetation change in the eastern United States. *Glob Chang Biol*. **21**(6): 2105–2110

R Core Team. 2014 R: A language and environment for statistical computing

Sato H., A. Itoh and T. Kohyama 2007 SEIB–DGVM: A new Dynamic Global Vegetation Model using a spatially explicit individual–based approach. *Ecol Modell*. **200**(3–4): 279–307

Schulze E. D., E. Beck and K. Müller–Hohenstein 2005 Plant Ecology. Berlin: Springer

Shiver B. D and B. E. Borders 1996 Sampling Techniques for Forest Resource Inventory. (Borders BE (Bruce E, ed.). New York: Wiley

SPSS Inc. 2009 PASW Statistics for Windows, Version 18.0

Takahashi K. and I. Okuhara 2012 Comparison of climatic effects on radial growth of evergreen broad–leaved trees at their northern distribution limit and co–dominating deciduous broad–leaved trees and evergreen conifers. *Ecol Res*. **27**(1): 125–132

Thuiller W., B. Lafourcade, R. Engler and M. B. Araújo 2009 BIOMOD – A platform for ensemble forecasting of species distributions. *Ecography*. **32**(3): 369–373

Thuiller W., D. Georges and R. Engler 2015 biomod2 Package Manual

Thuiller W., Lafourcade B., Araujo M. 2010 The Presentation Manual for BIOMOD

Tu J. V. 1996 Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J Clin Epidemiol*. **49**(11): 1225–1231

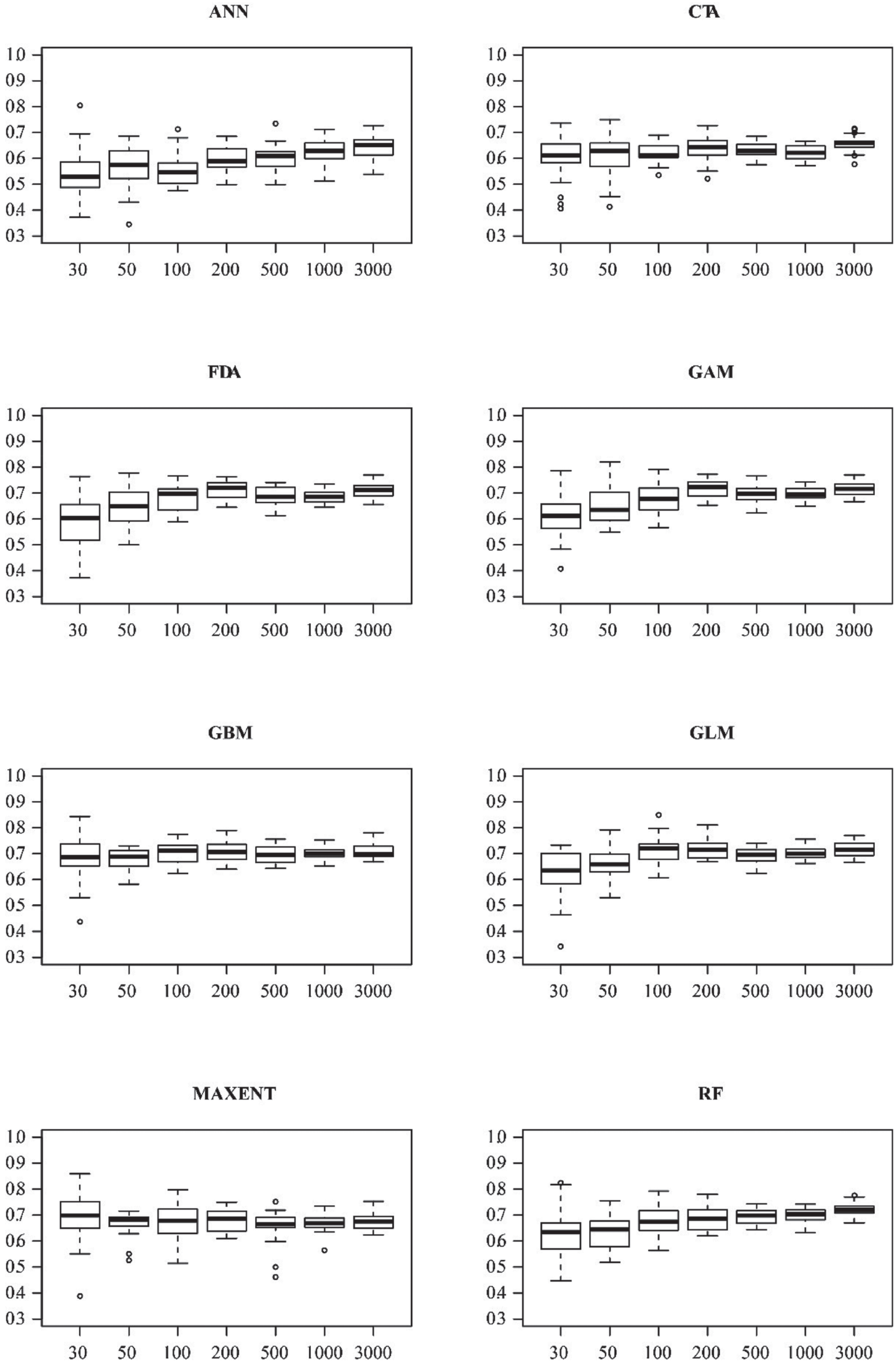
Walther G. R., E. Post, P. Convey, A. Menzel, C. Parmesan, T. J. C. Beebee, J. Fromentain, O. Hoegh–Guldberg and F. Bairlein 2002 Ecological responses to recent climate change. –*Nature* **416**(6879): 389–395

Wiens J. A., D. Stralberg, D. Jongsomjit, C. A. Howell and M. A. Snyder 2009 Niches, models, and climate change: Assessing the assumptions and uncertainties. *Proc Natl Acad Sci*. **106**(Supplement 2): 19729–19736

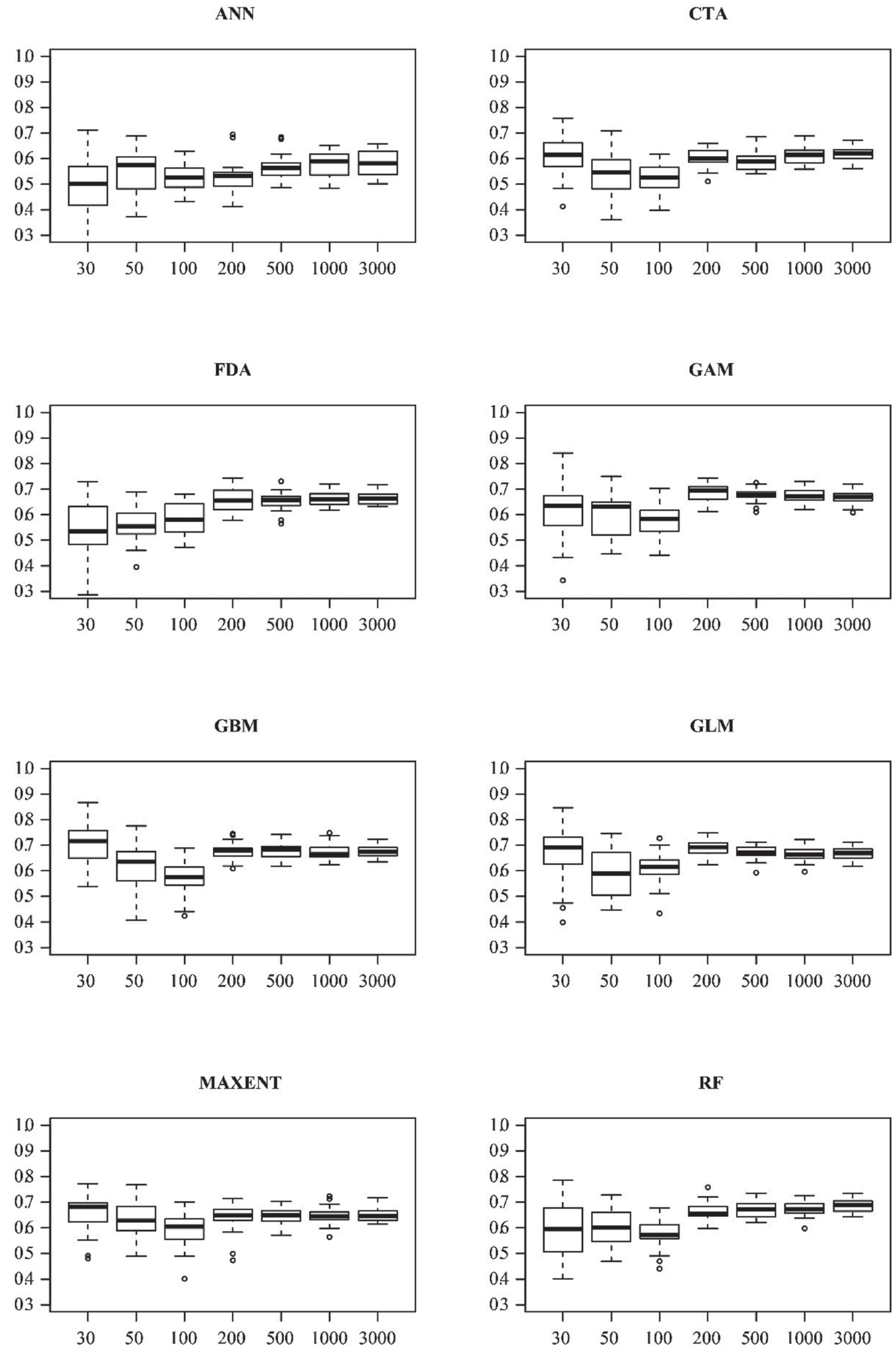
Wisz M. S., R. J. Hijmans, J. Li, A. T. Peterson, C.H. Graham, A. Guisan and N.P.S. Distribution 2008 Effects of sample size on the performance of species distribution models. *Divers Distrib*. **14**(5): 763–773

Yim Y. J. 1977 Distribution of Forest Vegetation and Climate in the Korean Penninsula: III. Distribution of Tree Species Along the Thermal Gradient. *Japanese J Ecol*. **27**(3): 177–189

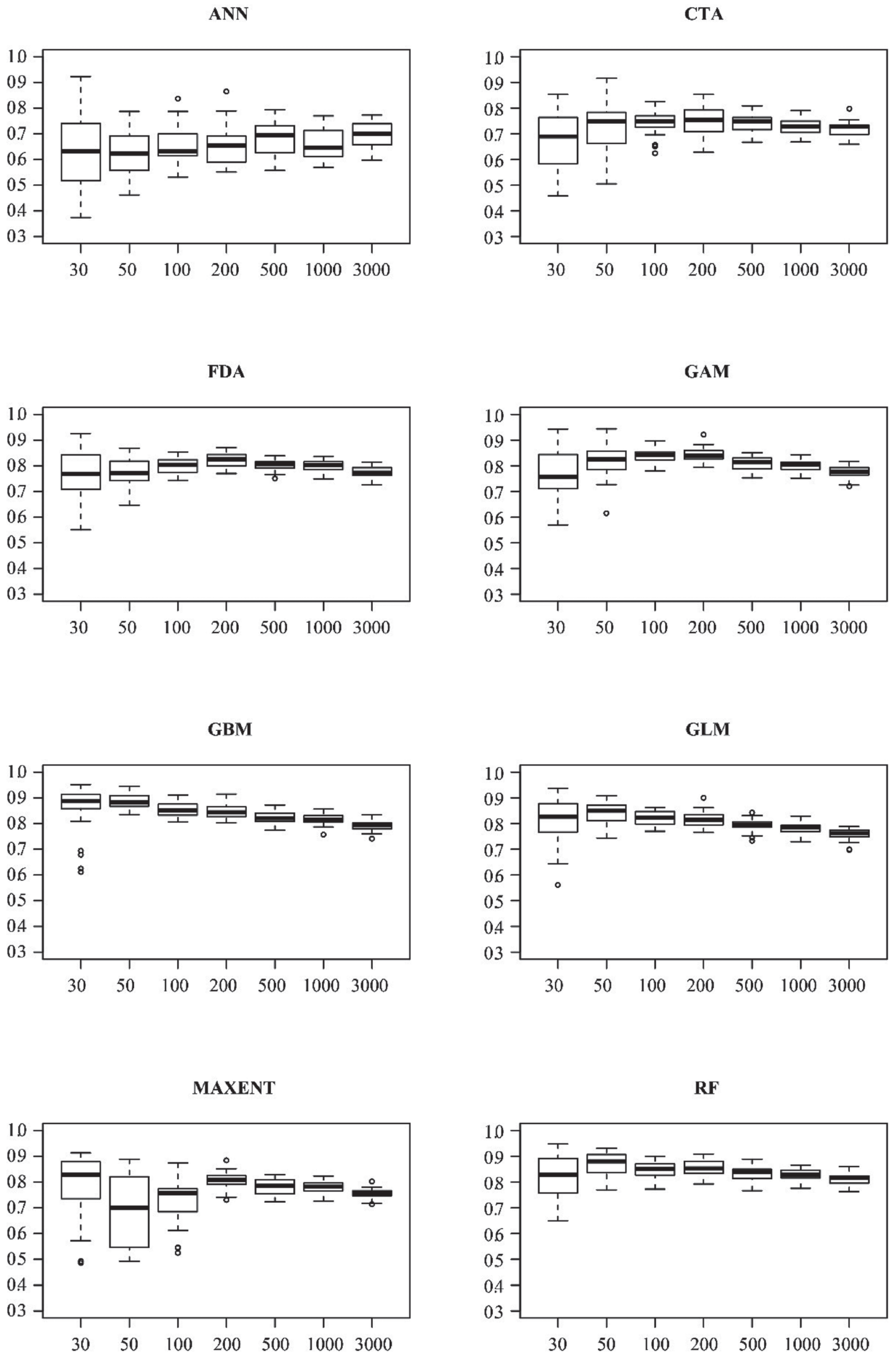
Zimmermann N. E., T. C. Edwards, C. H. Graham, P. B. Pearman and J. C. Svenning 2010 New trends in species distribution modelling. *Ecography*. **33**(6): 985–989



Supplementary Fig. 1. The performance of SDMs by sampling size (random sampling; x–axis: AUC, y–axis: sample size)



Supplementary Fig. 2. The performance of SDMs by sampling size (stratified sampling; x–axis: AUC, y–axis: sample size)



Supplementary Fig. 3. The performance of SDMs by sampling size (area–weighted sampling; x–axis: AUC, y–axis: sample size)