sciendo

SOMMERFELTIA 39 (2019)

How much is enough? Influence of number of presence observations on the performance of species distribution models

Bente Støa, Rune Halvorsen, Jogeir N. Stokland and Vladimir I. Gusarov

Bente Støa, Rune Halvorsen and Vladimir I. Gusarov, Natural History Museum, University of Oslo, P.O.Box 1172, Blindern, 0318 Oslo, Norway

Jogeir N. Stokland, The Norwegian Forest and Landscape Institute, P.O.Box 115, N-1431 Ås, Norway

Rune Halvorsen, Corresponding author. Tel.: +47-95477287, Email address: rune.halvorsen@nhm.uio.no

Bente Støa, Rune Halvorsen, Jogeir N. Stokland and Vladimir I. Gusarov 2018. How much is enough? Influence of number of presence observations on the performance of species distribution models. – Sommerfeltia 39: 1-28. Oslo. ISBN 978-82-7420-053-5. ISSN 0800-6865. DOI: 10.2478/som-2019-0001.

Species distribution modeling (SDM) can be useful for many applied purposes, e.g., mapping and monitoring of rare and endangered species. Sparse presence data are a recurrent, major obstacle to precise modeling of species distributions. Thus, knowing the minimum number of presences required to obtain reliable distribution models is of fundamental importance for applied use of SDM. This study uses a novel approach to assess the critical sample size (CSS) sufficient for an accurate prediction of species distributions with Maximum Entropy Modeling (MaxEnt). Large presence datasets for thirty insect species, ranging from generalists to specialists regarding their responses to main bioclimatic gradients, were used to produce reference distribution models. Models based on replicated subsamples of different size drawn randomly from the full dataset were compared to the reference model using the index of vector similarity (*IVS*). Two thresholds for *IVS* were determined based on comparison of nine reference models to random null models. The threshold values correspond to 0.95 and 0.99 probability that a

DOI: 10.2478/som-2019-0001
© Bente Støa, Rune Halvorsen, Jogeir N. Stokland, Vladimir I. Gusarov & Natural History Museum, University of Oslo

model outperforms a random null model in terms of similarity to the reference dataset. For 90% of the species, clearly nonrandom models were obtained with less than 10 presence observations, and for 97% of the species with less than 15 presence observations. We conclude that the number of presence observations required to produce nonrandom models is generally low and, accordingly, that even sparse datasets may be useful for distribution modelling.

Keywords: Hellinger distance, Maxent, Sample size, Species distribution modeling, Thresholds for critical sample sizes

Contents

Introduction 4
Material and methods 5
Study area 5
Species occurrence datasets 5
Explanatory variables 8
Subsampling presences from the full datasets 8
Distribution modelling by the software MAXENT 9
Assessment of similarity between models 10
Determination of the critical sample size 10
Statistical analysis of the critical sample size 12
Results 12
Discussion 19
Comparisons between our results and results from other studies 19
Selecting threshold values for determining critical sample size 20
Use of subsampling of full datasets for determination of critical sample size 21
Relationships between critical sample size and species properties 23
Does the general critical sample sixze exist? 24
Concluding remarks 24
Acknowledgements 25
References 25

INTRODUCTION

Distribution modeling (DM) is a rapidly evolving field of ecology (Franklin 2009, Lobo et al. 2010, Peterson et al. 2011). Many methods are available for predicting species distributions using geo-referenced species presence data together with wall-to-wall covering data for relevant environmental explanatory variables (Elith et al. 2006, Guisan and Zimmermann 2000) .

Most of the data available for species DM are of the presence-only type, i.e., datasets from which observations of species absence are lacking (Franklin 2009, Stokland et al. 2011). Accordingly, many of the commonly used DM methods are adapted to use presence-only data (Franklin 2009). One of the currently most popular methods is MaxEnt (Elith et al. 2011, Halvorsen 2013, Phillips et al. 2006, Phillips and Dudík 2008) which, based upon the maximum entropy principle (Jaynes 1957a, 1957b), estimates a probability distribution for the modeled target over the set of all *n* grid cells of a rasterized study area (Elith et al. 2011, Phillips et al. 2006, Phillips and Dudík 2008). MaxEnt has proved to give models with acceptable predictive ability even when few presence records are available, i.e., when the sample size is small (Elith et al. 2006, Hernandez et al. 2006, Marini et al. 2010, Peterson et al. 2011, Rebelo and Jones 2010).

Distribution models can be useful for many applied purposes, e.g., mapping and monitoring of rare and endangered species (Edvardsen et al. 2011, Engler et al. 2004, Guisan et al. 2006, Le Lay et al. 2010, Lomba et al. 2010, Marini et al. 2010) . Because rare species, by definition, are known from few localities and/or have restricted spatial distributions, sparsity of presence data is a major obstacle for modeling their distribution (Feeley and Silman 2011a, 2011b, Kamino et al. 2012, Lim et al. 2002, Papeş and Gaubert 2007). Thus, an important task of DM methodology is to determine a critical sample size (CSS), i.e., the minimum number of presence observations normally required to obtain reliable distribution models. In particular, knowledge of the CSS is important when rare species, typically known from few localities, are modeled. Assessment of a CSS has been addressed in several studies, without any general conclusions reached so far. Recommendations from these studies span from 5–10 (Hernandez et al. 2006) via 10–30 (Mateo et al. 2010, Stockwell and Peterson 2002, Wisz et al. 2008) to more than 200 presence observations (Hanberry et al. 2012) needed. The contrasting recommendations given in these studies may result from differences with respect to study organisms, study areas, spatial resolution (grid-cell size), modeling method and environmental variables used in the modeling, as well as from different criteria being used to assess acceptability of distribution models.

This study uses Norwegian insects to examine the influence of the amount of available data on the performance of distribution models. Insects are the group represented by the largest number of species (2013 out of 4599) on the current Norwegian Red List (Kålås et al. 2010); very well represented in the IUCN (International Union for Conservation of Nature) categories DD (Data Deficient) and NE (Not Evaluated). The IUCN-approved criteria for evaluation of Red List status (IUCN 2001, SPWG 2006) emphasize the probability of extinction within a certain time span and the rate of population decline over a given number of generations. However, these criteria are in general difficult to apply to insects, in particular to rare and poorly known species, for which population decline cannot be observed directly. Accordingly, the supplementary criteria based on changes in the species' geographic range are often used for insects. Distribution models facilitate use of these criteria and may therefore, potentially, be of great importance for future Red List assessments. So far, however, DM methods have been applied to insects much less often than to plants and vertebrates (Guisan and Thuiller 2005) , and most studies of insects using DM methods have addressed vectors of human diseases (López-Cárdenas et al. 2005, Peterson et al. 2005) or introduced species (Fitzpatrick et al. 2007, Roura-Pascual et al. 2004). Some studies have, however, applied DM methods to insect data for conservation

purposes. This is exemplified by the studies of threatened beetles in Tasmania (Meggs et al. 2004) and on the Iberian peninsula (Chefaoui et al. 2005). The CSS required to obtain acceptable distribution models for insects seems so far to have been addressed only in one study, of the butterfly *Danaus plexippus* (Hernandez et al. 2006). The large uncertainty with respect to assessment of a CSS for insects is particularly unfortunate because insects comprise more than half of all described species (Chapman 2009) and many insects are represented by few presence records (New 2009). The effect of sample size (number of presence observations of the modeled target) on reliability of predictions by distribution models has high general and applied importance because DM methodology is an emergent and promising tool for routine use in Red List assessments.

The aim of this study is threefold: (1) to propose and explore a novel method for determining the CSS (critical minimum number of presence observations required to obtain reliable distribution models); (2) to assess the CSS for different species of Norwegian insects; and (3) to discuss the extent to which a generally applicable CSS value can be determined for use with MaxEnt.

MATERIAL AND METHODS

STUDY AREA

Our study area is the mainland of Norway, which covers 323,782 km^2 and spans from 58 to 71 °N and from 5 to 31 °E. This area is particularly well suited for exploring properties of distribution models because it contains strong bioclimatic gradients in temperature (corresponding to vegetation zones from the boreonemoral to the high alpine) and oceanicity (corresponding to vegetation sections from the strongly oceanic to the slightly continental) (Moen 1999, Wollan et al. 2008). The main land-cover types in Norway are non-forested land (46%; mainly situated above and north of the tree line), forest (38%) and mires and lakes (6 % each). The high topographical and geological diversity in Norway, and the variation in intensity of human land-use, e.g. for agricultural purposes, bring about variation in environmental conditions and species composition over a large range of spatial scales (Halvorsen 2012).

SPECIES OCCURRENCE DATASETS

For this study we selected thirty insect species, ten from each of the following orders: Coleoptera, Diptera and Lepidoptera (Table 1). The species were selected to represent contrasting distribution patterns, from restricted to broad, based on the species' ecological range (the method used to quantify the ecological range is described in the final section of the Methods chapter). For every combination of order and distribution pattern we selected the species with the largest number of available presence observations.

Most of the presence observation data used in this study were extracted from the database of the insect collections at the Natural History Museum, University of Oslo, Norway. Additional

Table 1
List of modeled species with their systematic position, ecological range in Norway, number of presences in full datasets, and range of subsamples generated.

Species	Family	Order	Full dataset size	Subsamples	Relative fraction of oceanicity x relative fraction of temperature	Ecological range
Ips acuminatus	Curculionidae	Coleoptera	65	5, 10 ..., 60	0.2785	broad
Leptura maculata	Cerambycidae	Coleoptera	58	5, 10 ..., 55	0.0928	restricted
Meligethes aeneus	Nitidulidae	Coleoptera	60	5, 10 ..., 55	0.0337	restricted
Otiorhynchus nodosus	Curculionidae	Coleoptera	82	5, 10 ..., 75	0.2980	broad
Pogonocherus hispidus	Cerambycidae	Coleoptera	44	5, 10 ..., 40	0.0565	restricted
Rhagium mordax	Cerambycidae	Coleoptera	111	5, 10 ..., 100	0.1810	intermediate
Rhagonycha limbata	Cantharidae	Coleoptera	113	5, 10 ..., 100	0.3161	broad
Selatosomus aeneus	Elateridae	Coleoptera	84	5, 10 ..., 80	0.2196	broad
Strophosoma capitatum	Curculionidae	Coleoptera	46	5, 10 ..., 40	0.1228	intermediate
Tetrops praeusta	Cerambycidae	Coleoptera	41	5, 10 ..., 35	0.1364	intermediate
Conops quadrifasciatus	Conopidae	Diptera	133	5, 10 ..., 100, 120	0.1060	restricted
Dioctria hyalipennis	Asilidae	Diptera	95	5, 10 ..., 90	0.0728	restricted
Eristalis arbustorum	Syrphidae	Diptera	137	5, 10 ..., 100, 120	0.2890	broad
Eristalis interrupta	Syrphidae	Diptera	92	5, 10 ..., 85	0.2182	broad
Eristalis intricaria	Syrphidae	Diptera	194	5, 10 ..., 100, 120 ..., 180	0.2387	broad
Eristalis pertinax	Syrphidae	Diptera	141	5, 10 ..., 100, 120	0.1462	intermediate
Laphria flava	Asilidae	Diptera	73	5, 10 ..., 70	0.1237	intermediate
Neoitamus socius	Asilidae	Diptera	93	5, 10 ..., 90	0.0792	restricted
Sicus ferrugineus	Conopidae	Diptera	210	5, 10 ..., 100, 120 ..., 180	0.2136	intermediate
Volucella bombylans	Syrphidae	Diptera	103	5, 10 ..., 95	0.1656	intermediate
Aporia crataegi	Pieridae	Lepidoptera	57	5, 10 ..., 50	0.1622	intermediate
Glaucopsyche alexis	Lycaenidae	Lepidoptera	84	5, 10 ..., 80	0.0389	restricted

Table 1 (Continued)

Species	Family	Order	Full dataset size	Subsamples	Relative fraction of oceanicity x relative fraction of temperature	Ecological range
Heterothera serraria	Geometridae	Lepidoptera	49	5, 10 …, 45	0.1251	intermediate
Lasiocampa trifolii	Lasiocampidae	Lepidoptera	45	5, 10 …, 40	0.0170	restricted
Parnassius apollo	Papilionidae	Lepidoptera	88	5, 10 …, 85	0.1585	intermediate
Pieris napi	Pieridae	Lepidoptera	346	5, 10 …, 100, 120 …, 320	0.2750	broad
Thecla betulae	Lycaenidae	Lepidoptera	43	5, 10 …, 40	0.0163	restricted
Xanthoroe annotinata	Geometridae	Lepidoptera	112	5, 10 …, 100	0.0520	restricted
Xanthoroe decoloraria	Geometridae	Lepidoptera	238	5, 10 …, 100, 120 …, 220	0.0593	restricted
Zygaena exulans	Zygaenidae	Lepidoptera	169	5, 10 …, 100, 120, 140	0.1150	restricted
Pieris napi$_j$; j = 1, ..., 10	Pieridae	Lepidoptera	246[a]	5, 10…,95	0.2750	broad
Species narrow$_j$; j = 1, ..., 10	N/A, simulated	N/A, simulated	170[a]	5, 10…, 95	0.0163	restricted

[a], Number of presences in the dataset used for the reference model, see section 2.4 for details

data were obtained from the insect collections at Bergen Museum (University of Bergen, Norway) and the Museum of Natural History and Archaeology (Norwegian University of Science and Technology, Trondheim). Voucher specimens for all species were examined and only specimens for which the identity was confirmed and that could be geo-referenced with an uncertainty of 1 km or less were used for this study and subjected to data analysis.

The study area was rasterized by applying a 1 × 1 km grid including a total of 302,484 grid cells. Presence observations were assigned to grid cells based on their geographic coordinates. Multiple presences in one grid cell were treated as one single presence observation. Presences in grid cells with the center points outside the Norwegian border or in the sea were excluded from the analyses if situated more than 1 km off the coast or border. Otherwise they were moved to the nearest adjacent grid cell. The number of presences in the full datasets for each of the thirty investigated species ranged from 41 to 346 (Table 1).

EXPLANATORY VARIABLES

We used two continuous environmental (bioclimatic) explanatory variables for all models: a step-less oceanicity gradient and a step-less temperature gradient (Bakkestuen 2008). Both variables were available for all 1 × 1 km grid cells in the rasterized study area. The two variables were obtained by PCA ordination of 54 climatic, topographical, hydrological and geological variables recorded or modeled for the entire mainland of Norway, followed by a subsequent rotation of the PCA axes to maximize the fit to the division of Norway by Moen (1999) into vegetation sections (reflecting the oceanicity gradient; PCA axis 1) and vegetation zones (reflecting the summer temperature gradient; PCA axis 2). These two explanatory variables, which will be referred to as oceanicity and temperature, respectively, are the two most important regional bioclimatic gradients in Norway (Bakkestuen 2008, Moen 1999). Together, they explain 63% of the variation in the set of 54 variables subjected to PCA ordination (Bakkestuen 2008).

SUBSAMPLING PRESENCES FROM THE FULL DATASETS

For each species effects of sample size on the predictive performance of distribution models were explored by generating random subsamples of different sizes from the full dataset. Subsamples ranged in size from 5 presences, with an increment of 5, up to 100 and, if applicable, with an increment of 20 beyond 100 presences, until the size of the full dataset was reached (Table 1). Twenty replicate subsamples were generated for each combination of species and subsample size.

For a given species, the random subsamples are not independent of the full dataset. This lack of independence is however not assumed to influence the properties of the distribution models at small sample sizes (See Discussion). To test this assumption, two *additional datasets* were prepared as described below.

Additional dataset I was obtained from data for the species with the largest number of presences in our study (346), the widespread generalist (in terms of ecological tolerance) butterfly *Pieris napi*. The full dataset for this species was randomly split into two subsets, with 246 and 100 presences. This procedure was repeated ten times, resulting in ten subset pairs (referred to as $Pieris\ napi_j$; j = 1, ..., 10). This repeated random splitting procedure was chosen for evalu-

ation of distribution models because it is among the best methods in cases where independent presence-absence evaluation datasets cannot be obtained (Austin 2007, Halvorsen 2012, Raes and ter Steege 2007, Veloz 2009). The larger subset in each pair was used as reference while the smaller was used to randomly generate 380 subsamples: 20 subsamples of each size from 5 to 95 in increments of 5.

Additional dataset II was simulated to represent a locally common specialist species, i.e., a species with narrow ecological tolerance, represented by many presence observations from a restricted distribution area. This combination of properties was not encountered among the thirty species selected for the analyses. The following simulation procedure was used: (1) The full datasets for all the thirty species were pooled. (2) A subset of presences was selected from the pooled dataset according to the following conditions: $0.00028 \leq$ oceanicity ≤ 0.00141 and $0.00323 \leq$ temperature ≤ 0.00416. These ranges of environmental values correspond exactly to those of the species with the narrowest bioclimatic envelope in our study, the butterfly *Thecla betulae*. This bioclimatic envelope for this simulated species, referred to as 'Species narrow', included 270 occurrence points. As for *Pieris napi*, the total dataset of 270 presences of Species narrow was randomly split 10 times into paired subsets referred to as Species narrow$_j$; $j = 1, ..., 10$. The larger subset, with 170 presence observations, was used as reference while the smaller, with 100 presence observations, was used to generate random subsamples as described for Additional dataset I above.

DISTRIBUTION MODELING BY THE SOFTWARE MAXENT

Distribution models for all datasets listed in Table 1 were obtained using Maxent software, version 3.3.1 (Phillips et al. 2006), which performs distribution modeling by the MaxEnt method (note the distinction between the software Maxent and the modeling method MaxEnt). This method is based on the maximum entropy principle (Jaynes 1957a, 1957b) and has been described as a machine learning method (Elith et al. 2011, Phillips et al. 2006, Phillips and Dudík 2008) or as a maximum likelihood estimation method (Halvorsen 2013, Renner and Warton 2013). MaxEnt estimates the probability distribution of maximum entropy for the modeled target, i.e., the distribution which is most spread out or closest to uniform, subject to a set of constraints that represent our incomplete information about the target distribution (Phillips et al. 2004). With access only to presence data, the true prevalence of a modeled species (the proportion of grid cells in the study area occupied by the species) is neither known nor possible to estimate, and distribution models therefore provide estimates for the probability of finding the modeled target under given ecological conditions, *relative* to other conditions. Model predictions are therefore relative predicted probabilities of presence (RPPP values; Halvorsen 2012).

All MaxEnt models were obtained using default Maxent program values for all settings and options (e.g., Phillips and Dudík 2008) except output format, which was set to "raw". The raw output format consists of a set of values that sum to unity for the total training dataset of presence and uninformed background observations. With default settings Maxent is run with automatic generation of derived variables ["auto features" in the terminology of Phillips et al. (2006)] from the (in our study) two explanatory variables. This opens for derived variables of up to five types to be generated by the program from each supplied explanatory variable, depending on the number of presences in the set subjected to modeling: linear variables (all sets); quadratic variables (sets with ≥ 10 presences); hinge variables (≥ 15 presences), and product and threshold variables (≥ 80 presences). The default model selection method in Max-

ent software, the shrinkage method referred to as $\ell 1$-regularisation (Hastie et al. 2009) or lasso penalty (Tibshirani 1996), was used with the default regularization multiplier = 1. Each model was trained on a dataset consisting of the presence observations and 10,000 randomly selected uninformed background observations (i.e., observation units for which nothing is known about presence or absence of the target species; Halvorsen 2012). All models were projected to all 302,484 grid cells in the study area. The sum of predictions for all grid cells therefore summed to 302,484/ (ca. 10,000) ≈ 30.

ASSESSMENT OF SIMILARITY BETWEEN MODELS

Predictions from each MaxEnt model were converted into a vector of RPPP values. Vectors based on replicated subsamples were compared, grid cell by grid cell, with the vector based on the respective reference model obtained using the full dataset. Reference models for *Pieris napi*$_j$ and Species narrow$_j$ were obtained by use of the larger subset in each pair of subsets.

The Index of Vector Similarity (*IVS*) was used to measure the similarity between RPPP vectors in each pair. *IVS* is derived from the Hellinger distance (Van der Vaart 1998) between the two corresponding RPPP vectors. After normalizing the vector of RPPP values to a sum of 1, *IVS* was calculated as follows:

$$IVS(px,py) = 1 - \frac{1}{\sqrt{2}} \sqrt{\Sigma(\sqrt{px_i} - \sqrt{py_i})^2},$$

where px_i and py_i are corresponding values of two RPPP vectors. *IVS* ranges from 0, when two vectors are completely dissimilar, to 1 when in every pair the two vectors are identical (i.e. all $px_i = py_i$).

DETERMINATION OF THE CRITICAL SAMPLE SIZE

The critical sample size (CSS) was determined with reference to two threshold values of IVS by use of sets of random predictions generated as follows: (1) Ten datasets were created for each of seven sample sizes (5, 20, 40, 60, 100, 200 and 300) by randomly selecting the required number of grid cells from the 302,484 cells covering Norway. (2) MaxEnt models, referred to as null models, were obtained for each of these 70 datasets, using default Maxent settings. (3) Each null model was compared with the models obtained for full datasets for nine species (*D. hyalipennis*, *E. arbustorum*, *E. pertinax*, *I. acuminatus*, *L. maculata*, *P. apollo*, *P. napi*, *R. mordax* and *T. betulae*) representing all combinations of three insect orders and three ecological range types (broad, intermediate and restricted). In order to decide if the threshold *IVS* values had to be set separately for each group of species, separate one-way ANOVAs with *IVS* as response variables and random dataset size and distribution type as predictors were conducted. *IVS* ($F_{1,628}$ = 0.039, p = 0.844) was not significantly related to subsample size (Fig. 1) but differed significantly among the three distribution types ($F_{2,627}$ = 567.1, p < 10-10; Fig. 1), being lower for the species with restricted ecological range and higher for the species with broad ecological range. (4) The 0.95 and 0.99 quantiles in the distribution of 630 pooled *IVS* values, calculated for all combinations of the nine species and seventy datasets, were used as thresholds defining

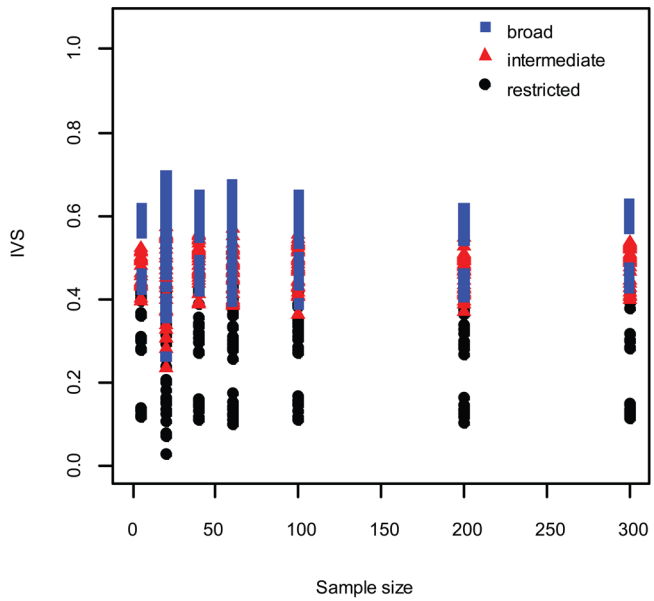


Fig. 1. Distribution of *IVS* in comparisons between seventy null models based on random datasets (ten replicates of seven sample sizes) and reference models for nine species.

nonrandom and clearly nonrandom models, respectively. The threshold values were *IVS* = 0.622 and 0.652, respectively, for nonrandom (0.95 quantile) and clearly nonrandom (0.99 quantile) models. The significant relationship between *IVS* and distributional type makes the CSS estimates based upon the general threshold conservative when applied to species with restricted and intermediate distributions.

Two critical sample sizes, defined as the number of presences required to obtain nonrandom and clearly nonrandom models (CSS_n and CSS_{cn}), were determined for all species (including the 20 'species' *Pieris napi*$_j$ and Species narrow$_j$) as follows: (1) For each subsample size (5, 10, ...), the 5%-percentile in the distribution of twenty *IVS* values, obtained by comparing the models based on subsamples with the reference model, was plotted as a function of the number of presence observations (which will be referred to as the 'sample size'). (2) CSS_n and CSS_{cn} were determined as the number of presences at which the graph that joins subsequent values of *IVS* crosses the threshold values for nonrandom and clearly nonrandom models, respectively. (3) In cases where the graph was not monotonously ascending, the graph was smoothed until it became monotonous by applying a three-point moving averages approach with weights of 1, 2 and 1 for three consecutive points (subsample sizes). (4) CSS values were rounded off to the nearest integer value. The CSS can be interpreted as the number of presence observations needed for a model to produce predictions that are, in 19 out of 20 and 99 out of 100 cases, respectively, more similar to predictions from the reference model (based on the full dataset) than to the predictions of random models.

STATISTICAL ANALYSIS OF THE CRITICAL SAMPLE SIZE

General patterns of variation in *IVS* with increasing sample size were modeled by non-linear regression analyses (Crawley 2007), using *IVS* as response variable and subsample size as independent variable. We fit the asymptotic exponential function, $y = 1 - ae^{-bx}$, with two parameters and the horizontal asymptote $y = 1$ (corresponding to the maximum possible value for *IVS*); a is 1 minus the *y* intercept, and *b* is the rate constant.

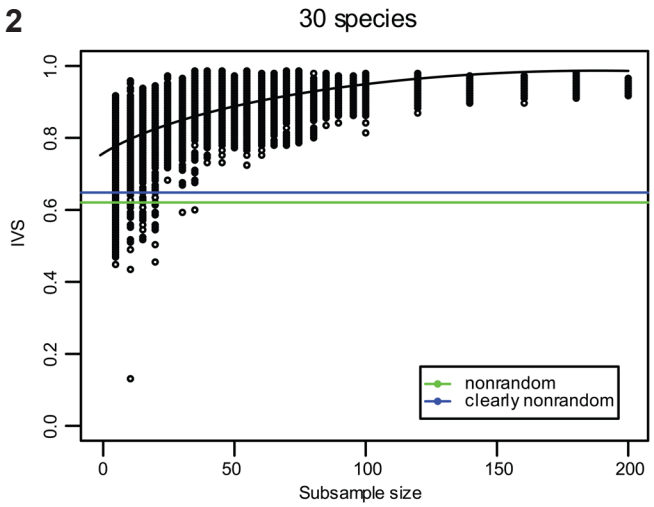
Generalized linear models (GLMs) with identity link and normal errors (standard linear regression analyses and ANOVAs; Crawley, 2007), and the non-parametric Wilcoxon-Mann-Whitney tests (Crawley 2007) were used to explore relationships between each of the two alternative response variables, CSS_n and CSS_{cn}, and properties of the species (three independent variables): the species' relative ecological range, taxonomic affiliation (order) and the number of presence observations in the full dataset. The relative ecological range of a species was estimated using the full dataset for the species as follows: (1) The smallest interval along each of the oceanicity and the temperature gradients containing 90% of the species presences was found. (2) The widths of both intervals were rescaled as the fraction of the range along the respective gradients spanned by all grid cells (in Norway). (3) The product of the two fractions was used as an estimate of the relative ecological range of the species. This relative ecological range was used to assign species to distribution types (see Table 1). Species with restricted, intermediate and broad distribution had relative ecological ranges of 0.0163–0.1150, 0.1228–0.2135 and 0.2182–0.3161 respectively.

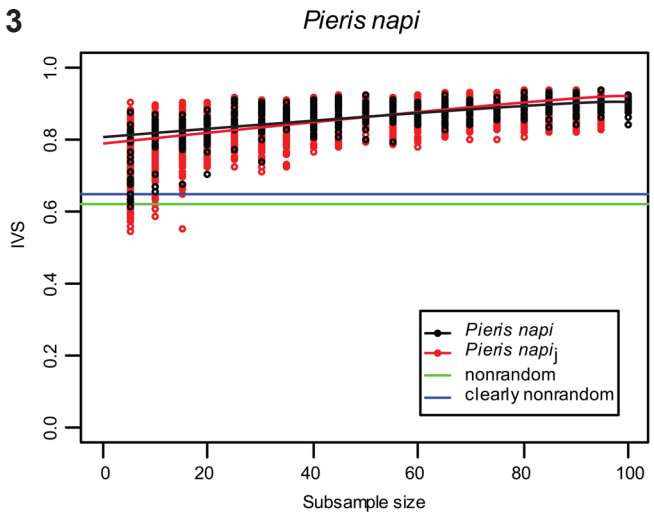
RESULTS

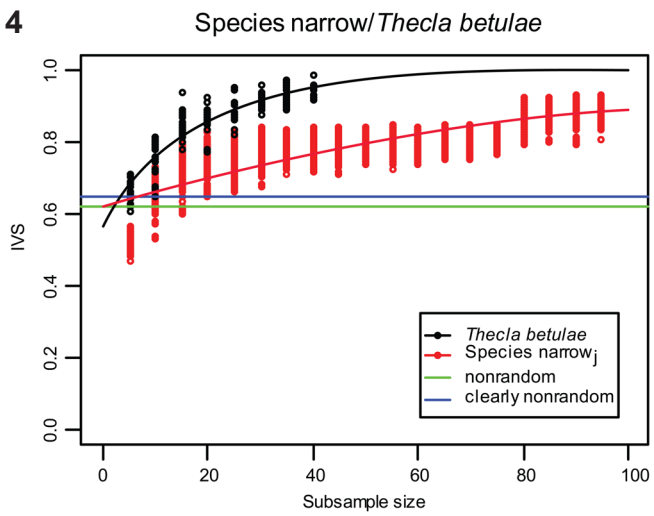
Comparisons between subsample-based models and reference models for the respective full datasets revealed substantial variation in *IVS* (range = 0.131–0.990, mean = 0.872) among the 9860 models obtained for all combinations of subsample sizes and species. MaxEnt models based on subsamples became more similar to corresponding models based on the full datasets with increasing subsample size, for all species pooled (Fig. 2) and for each species analyzed separately (Figs 5-28), including *P. napi*$_j$ and Species narrow$_j$ (Fig. 3 and Fig. 4).

IVS was generally higher (range = 0.549–0.939, mean = 0.846) for the widespread generalist species *P. napi*$_j$ (j = 1, ... , 10) than for Species narrow$_j$ (range = 0.470–0.931; mean = 0.785) (Wilcoxon test: $p < 1{\cdot}10^{-10}$). A comparison of *IVS* values for *P. napi*$_j$ and Species narrow$_j$ with those for *P. napi* and *T. betulae* (which determine the relative ecological range on which Species narrow$_j$ is based), respectively, revealed higher values when subsamples were drawn

Figs 2–4. Relationship between model similarity, expressed by IVS, and subsample size (the number of presence observations). Models for each subsample size are compared with the respective reference models based on full datasets. (2) All 30 species pooled; (3) *Pieris napi*$_j$ and *Pieris napi*; (4) Species narrow and *Thecla betulae*. The black and red lines show non-linear regression curves (two-parameter asymptotic exponential functions are fitted) for the species listed in the legend. The green and blue lines indicate the nonrandom and clearly nonrandom thresholds, respectively (see text for more details).


2
30 species
IVS
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
Subsample size
nonrandom
clearly nonrandom


3
Pieris napi
IVS
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Subsample size
Pieris napi
Pieris napij
nonrandom
clearly nonrandom


4
Species narrow/Thecla betulae
IVS
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Subsample size
Thecla betulae
Species narrowj
nonrandom
clearly nonrandom

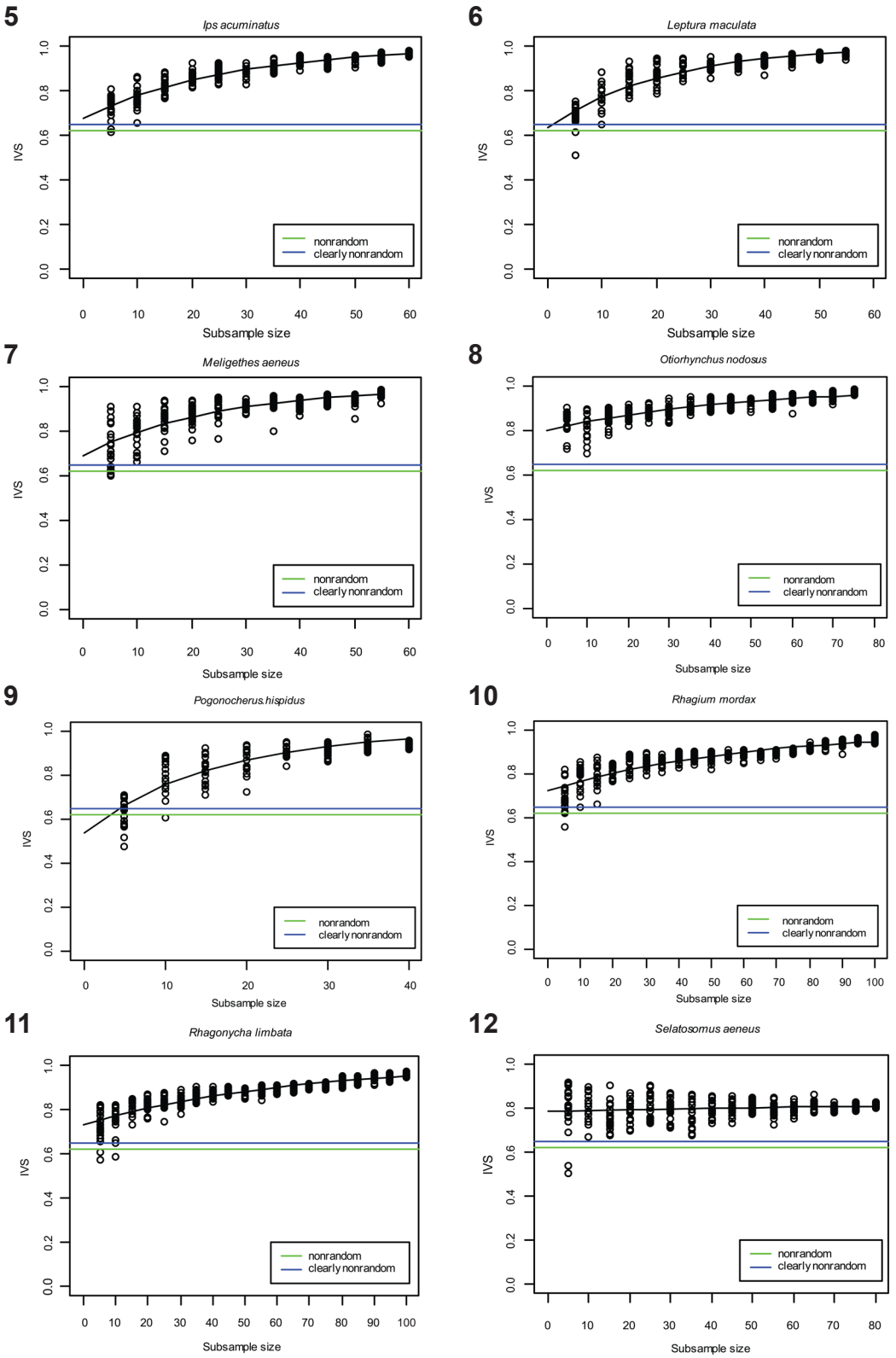
from the full dataset than when they were drawn from the corresponding independent dataset (all four Wilcoxon two-sample paired tests: $p < 1.0 \cdot 10^{-10}$).

The CSS_n and CSS_{cn} ranged from 4 to 21 (mean = 7) and from 4 to 22 (mean = 8), *IVS* (range = 0.131–0.990, mean = 0.872) varied accordingly, depending on the species (see Table 2 for details). CSS_n and CSS_{cn} were higher for Species narrow$_j$, than for *P. napi*$_j$ (Table 2), as confirmed by Wilcoxon tests (p = 0.0002 and 0.0026, respectively). Comparisons between corresponding CSS_n and CSS_{cn} values for the 30 species as one group and the 10 *P. napi*$_j$ as another group revealed no significant differences (Wilcoxon paired two-sample tests, see Table 3). Comparisons between the 30-species group and Species narrow$_j$ showed that CSS_n was higher for Species narrow$_j$ than for the 30 species (p = 0.0001). CSS_{cn} was also higher for Species narrow$_j$ than for

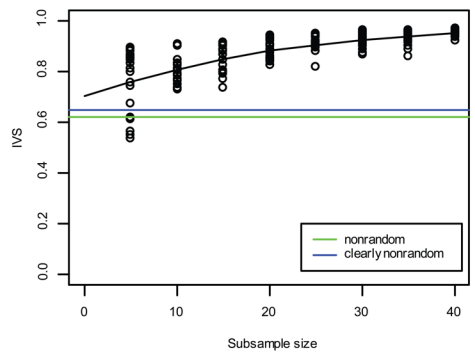
Table 2. CSS needed to exceed the threshold values of *IVS* for nonrandom and clearly nonrandom models.

Species	CSS_n	CSS_{cn}	Species	CSS_n	CSS_{cn}
Ips acuminatus	5	6	*Pieris napi*	5	8
Leptura maculata	5	7	*Thecla betulae*	6	8
Meligethes aeneus	6	8	*Xanthoroe annotinata*	7	8
Otiorhynchus nodosus	4	4	*Xanthoroe decoloraria*	4	5
Pogonocherus hispidus	8	9	*Zygaena exulans*	4	4
Rhagium mordax	5	7	*Pieris napi 1*	5	5
Rhagonycha limbata	7	10	*Pieris napi 2*	6	7
Selatosomus aeneus	7	8	*Pieris napi 3*	9	11
Strophosoma capitatum	7	8	*Pieris napi 4*	8	11
Tetrops praeusta	7	8	*Pieris napi 5*	6	8
Conops quadrifasciatus	5	6	*Pieris napi 6*	6	8
Dioctria hyalipennis	5	7	*Pieris napi 7*	8	10
Eristalis arbustorum	8	10	*Pieris napi 8*	6	7
Eristalis interrupta	12	13	*Pieris napi 9*	6	7
Eristalis intricaria	7	7	*Pieris napi 10*	6	7
Eristalis pertinax	6	8	Species narrow 1	10	12
Laphria flava	4	5	Species narrow 2	12	14
Neoitamus socius	8	9	Species narrow 3	9	10
Sicus ferrugineus	11	12	Species narrow 4	13	15
Volucella bombylans	5	5	Species narrow 5	10	11
Aporia crataegi	21	22	Species narrow 6	9	10
Glaucopsyche alexis	6	7	Species narrow 7	10	15
Heterothera serraria	8	9	Species narrow 8	10	16
Lasiocampa trifolii	5	6	Species narrow 9	9	9
Parnassius apollo	8	9	Species narrow 10	9	11

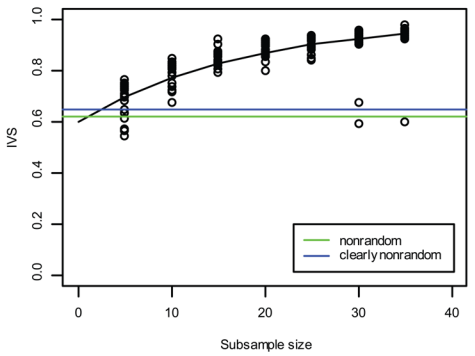
Figs. 5-28. Relationship between IVS and subsample size (the number of presence observations). The black lines show the non-linear regression curves (two-parameter asymptotic exponential function). The green lines show the nonrandom threshold and the blue line shows the clearly nonrandom threshold. See text for more details.


5
Ips acuminatus
6
Leptura maculata
7
Meligethes aeneus
8
Otiorhynchus nodosus
9
Pogonocherus hispidus
10
Rhagium mordax
11
Rhagonycha limbata
12
Selatosomus aeneus
IVS
Subsample size
nonrandom
clearly nonrandom

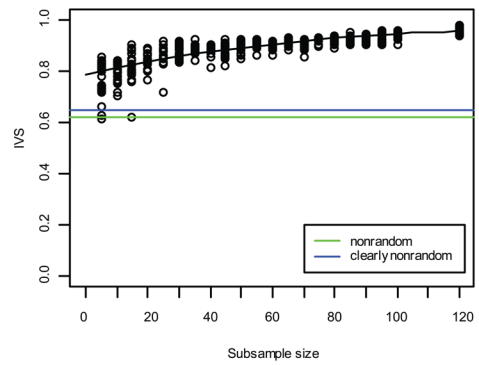
13



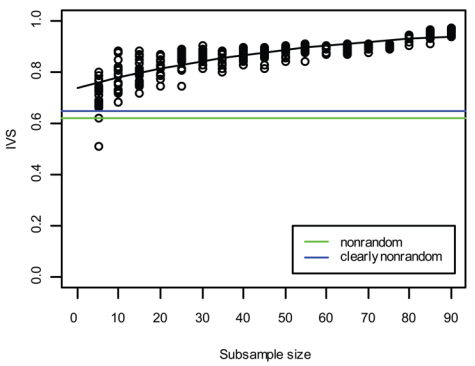
14



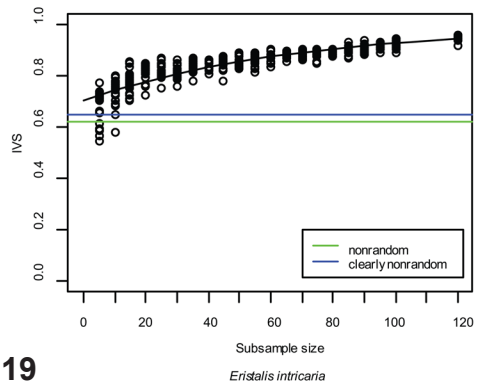
15



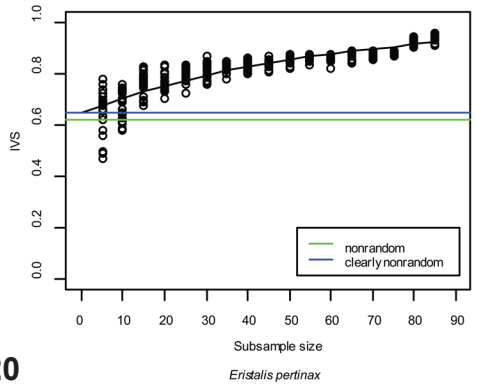
16



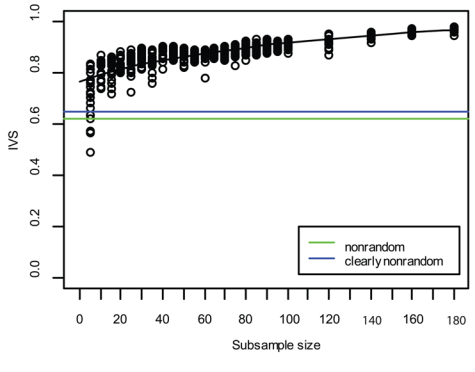
17



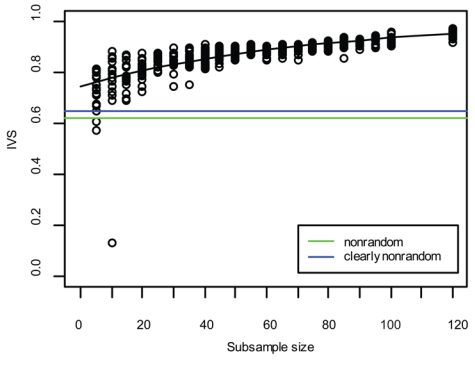
18

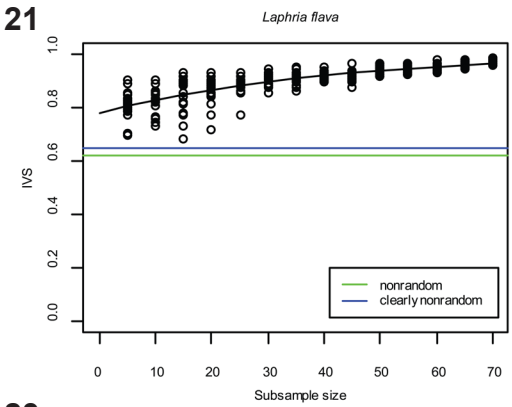


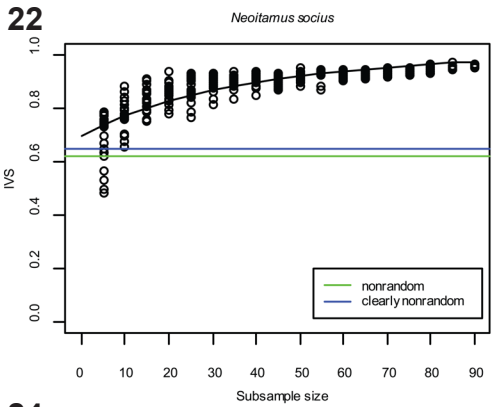
19

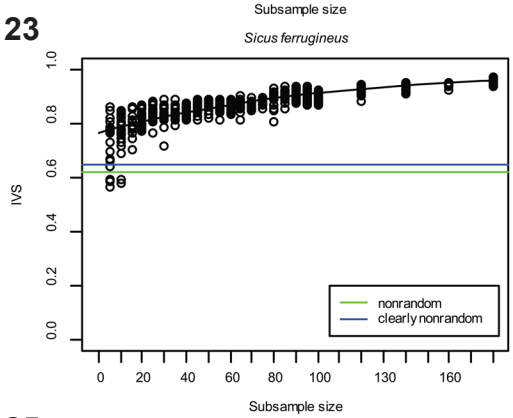


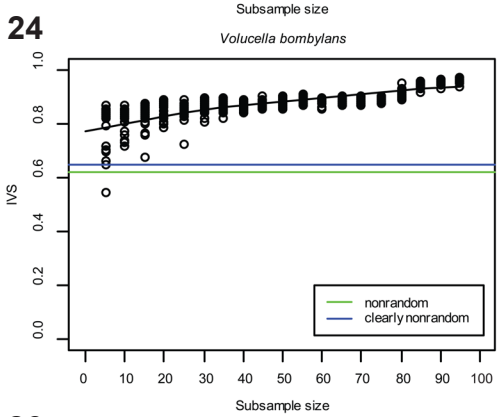
20

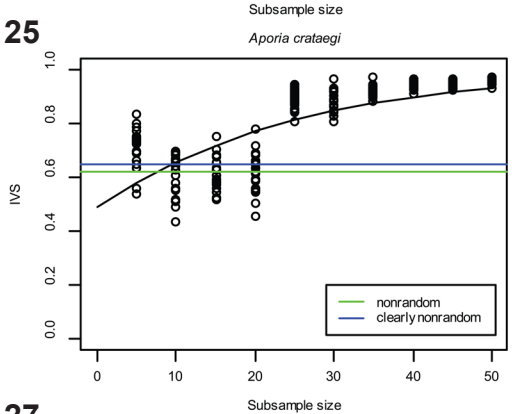


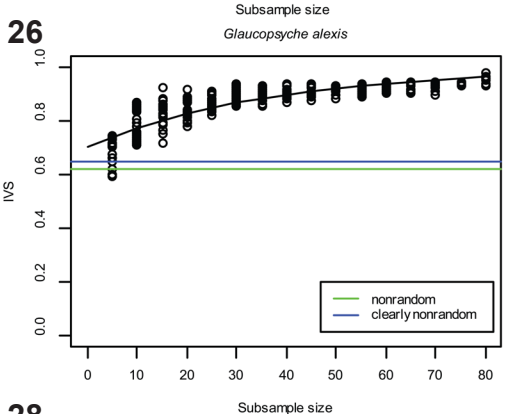

21
Laphria flava
IVS
Subsample size
nonrandom
clearly nonrandom

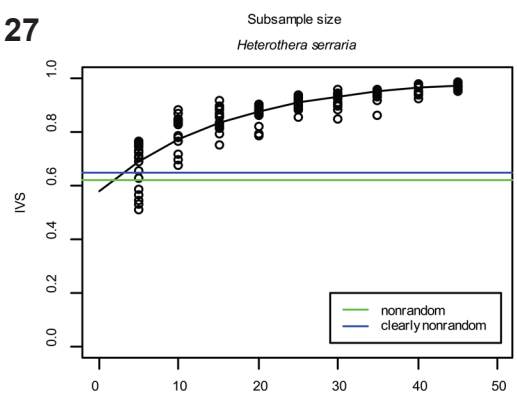

22
Neoitamus socius
IVS
Subsample size
nonrandom
clearly nonrandom


23
Sicus ferrugineus
IVS
Subsample size
nonrandom
clearly nonrandom


24
Volucella bombylans
IVS
Subsample size
nonrandom
clearly nonrandom


25
Aporia crataegi
IVS
Subsample size
nonrandom
clearly nonrandom


26
Glaucopsyche alexis
IVS
Subsample size
nonrandom
clearly nonrandom


27
Heterothera serraria
IVS
Subsample size
nonrandom
clearly nonrandom

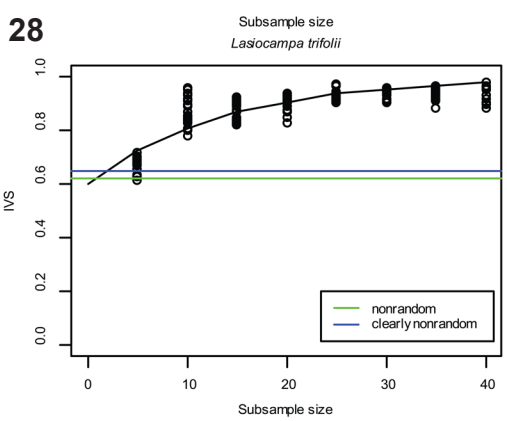

28
Lasiocampa trifolii
IVS
Subsample size
nonrandom
clearly nonrandom

Table 3. Wilcoxon tests for comparisons among CSS_n and CSS_{cn}, between *P. napi*$_j$ and Species narrow$_j$, *P. napi*$_j$ and the 30 species, and Species narrow$_j$ and the 30 species. *N* = number of species/reference models. *M* = mean sample size for obtaining nonrandom/ clearly nonrandom models, *SD* = standard deviation, *p* = *p* value of the Wilcoxon test.

Model similarity measure	Model property	*P. napi*$_j$			Species narrow$_j$			30 species			p comparison *P. napi*$_j$ – Species narrow$_j$	p comparison *P. napi*$_j$ – 30 species	p comparison Species narrow$_j$ – 30 species
		N	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*			
IVS	Nonrandom	10	6.6	1.3	10	10.1	1.4	30	6.9	3.3	0.0002	0.5777	0.0001
IVS	Clearly nonrandom	10	8.1	2.0	10	12.3	2.5	30	8.1	3.6	0.0026	0.7754	0.0001

the 30 species (p = 0.0001) (Table 3).

CSS_n and CSS_{cn} neither differed significantly among species with different relative ecological ranges (Table 4), nor among species that belonged to different taxonomic orders (Table 5), nor among species with different numbers of presence observations in the full dataset (Table 4).

DISCUSSION

COMPARISONS BETWEEN OUR RESULTS AND RESULTS OF OTHER STUDIES

The main result of our study, that distributions are modeled more accurately when sample size increases, agrees with results of previous studies (Cumming 2000, Hernandez et al. 2006, Kadmon et al. 2003, Pearce and Ferrier 2000, Reese et al. 2005, Stockwell and Peterson 2002, Wisz et al. 2008). Furthermore, our results indicate that nonrandom models are obtained for a

Table 4. Linear regression analyses with critical sample size (response variables CSS_n and CSS_{cn}) and species' relative ecological range and full dataset size (number of presence observations in the full dataset) as predictor variables.

Response variable	Explanatory variable	Coefficient	Standard error	p-value
CSS_n	Species' relative ecological range	4.810	6.900	0.491
CSS_{cn}	Species' relative ecological range	5.539	7.006	0.436
CSS_n	Full dataset size	-0.008	0.009	0.401
CSS_{cn}	Full dataset size	-0.006	0.009	0.536

Table 5. Results from ANOVA with critical sample size (response variables CSS_n and CSS_{cn}) and taxonomic order (explanatory variable). *df* = model degree of freedom. *F* and *p* refer to F-test of the hypothesis of no difference between taxonomic orders.

Response variable	Explanatory variable	*F*	*df*	*p*
CSS_n	Taxonomic order	0.411	2 and 27	0.667
CSS_{cn}	Taxonomic order	0.263	2 and 27	0.771

majority of species with as few as 10 presence observations, although with some variation in the critical sample size (CSS) among species. The relatively small spread of CSS values obtained in our study does, however, contrast with the considerable disagreement among the above-mentioned studies with respect to the minimum sample size required for models to be reliable. Extremes in this respect are Pearson et al. (2007), who conclude that as few as 5 presences can be sufficient for MaxEnt to produce acceptable predictions of the distribution of geckos endemic to Madagascar, and Hanberry et al. (2012) who conclude that Random Forest models based on less than 200 presences may fail to be accurate. Sample sizes in between these two extremes are recommended by others. Thus, Stockwell and Peterson (2002), using GARP, find that 10 presence observations are in general sufficient to obtain a mean prediction accuracy of 90% of the maximum obtainable accuracy, while 50 presences result in near-maximal accuracy. Hernandez et al. (2006), using MaxEnt, find that useful results can be obtained with sample sizes as small as 5–10 presence observations. Mateo et al. (2010), on the other hand, find that the predictive power is considerably improved when sample size exceeds 18–20 presence observations, and Wisz et al. (2008) find that none among several compared methods, the overall best-performing method MaxEnt included, consistently provide good predictions with sample sizes smaller than 30 presence. Accordingly, Wisz et al. (2008) advise against the use of predictions from models obtained with fewer than 30 presence observations for practical conservation purposes.

Many factors have been claimed to influence the sample size required to obtain accurate distribution models, including the spatial resolution and extent of the study area (Loe et al. 2012), the modeling method (Dupin et al. 2011), and characteristics of the modelled species (Hernandez et al. 2006). Moreover, the criteria for considering a model as 'acceptable', 'useful' or 'reliable', which are frequently used as qualifiers (cf. Araújo & Guisan et al. 2006), will inevitably affect assessments of a minimum required sample size. Most published DM studies apply the AUC [the Area Under the receiver operating Curve (Fielding and Bell 1997, Pearce and Ferrier 2000)] as a model performance criterion. However, used with training data only, AUC is an internal model performance assessment criterion that measures the ability of the model to distinguish between presence and background observations in the training dataset and not the model's predictive capability in general, e.g. its ability to predict presence in other areas (Halvorsen 2012). The purpose of the current paper is to investigate how the modeled distribution is affected by changes in the number of presence observations used to make the model. We therefore restrict ourselves to discussing variation in similarity between models based on subsamples and reference models obtained from full datasets, using the similarity index, *IVS*.

SELECTING THRESHOLD VALUES FOR DETERMINING CRITICAL SAMPLE SIZE

To avoid the common practice of arbitrarily setting the thresholds for separating useful models from useless models, we determine thresholds based on comparisons with randomly generated models (i.e., models based upon sets of randomly selected grid cells). The rationale for this approach is that a model performing better than random (i.e., the model is more similar to the reference model than a randomly generated model) provides useful information about the modeled species. Significance levels of 0.05 and 0.01 are commonly used in tests of statistical hypotheses. For this reason, we selected two threshold values, based on 0.95 and 0.99 quantiles in the distribution of similarity indices calculated between a random model and a reference model, for assessment of the CSS.

We determine thresholds for *IVS* by comparisons between null models (based on randomly generated datasets of different sizes) and reference models for nine species that represent three

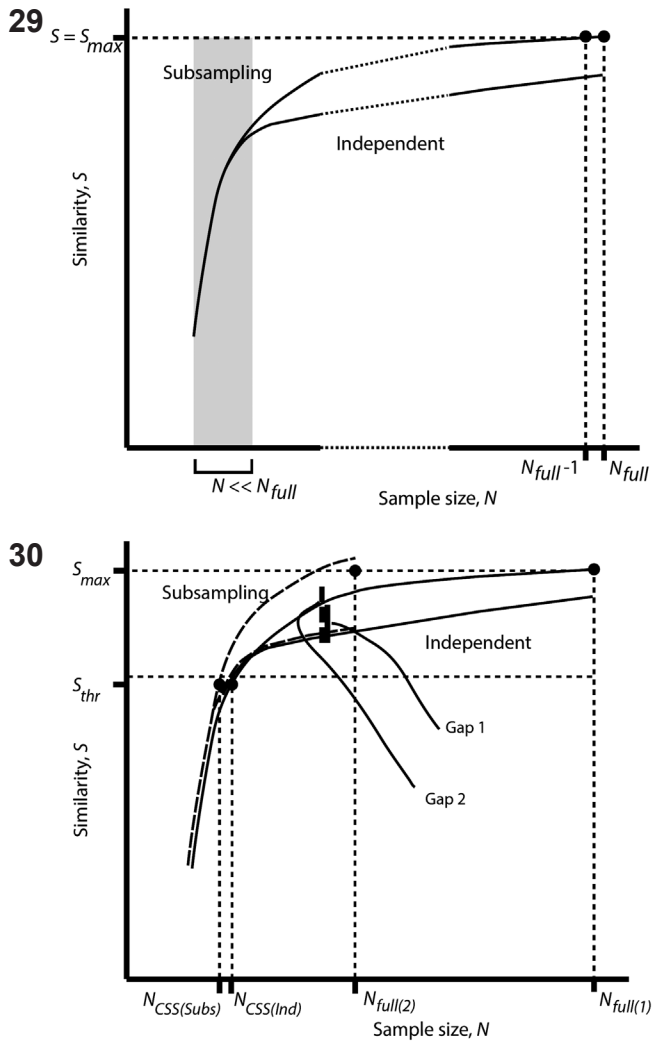
classes of ecological ranges; restricted, intermediate and broad. We find that species distribution patterns affect the threshold *IVS* values (see Results), being significantly lower for species with restricted distributions than for species with broader distributions. This result is likely to be due to the models based upon randomly chosen presences from all over the modeled area being more similar to the models based upon presence records for broadly distributed species than to models for species with restricted distributions. However, a threshold that applies to a broadly distributed species is definitely applicable also to a species with restricted distribution. Accordingly, we used pooled *IVS* values for all the nine species to determine threshold values for non-randomness that were to be used for comparisons between subsamples and reference models for all species and for CSS estimation. This represents a non-arbitrary and conservative criterion for estimating CSS that does not require prior knowledge of a species' distribution and that may thus also apply to rare or poorly known species, for which few presence observations are available.

USE OF SUBSAMPLING OF FULL DATASETS FOR DETERMINATION OF CRITICAL SAMPLE SIZE

Random subsamples of presence records for a species are not independent from the full dataset from which the subsamples are drawn. Accordingly, distribution models obtained from subsamples may be similar to reference models obtained for the full dataset for two reasons: (1) because they represent the same species, i.e., they indicate similar responses to the main underlying environmental complex-gradients; and (2) because they make use of the same presence records. If explanation (2) prevails, CSS estimates obtained by the subsampling method used in this paper are likely to underestimate the real CSS values, and conclusions about the reliability of distribution models based upon small datasets may be overly optimistic. Theoretically, the relative roles of the two explanations are expected to depend on subset size, as illustrated in Fig. 29. in which two model similarity curves (S), corresponding to the two sampling approaches, with and without subsampling, are compared. At large sample size N, the models based on subsamples from the full dataset become more and more similar to the reference model as N approaches N_{full} until the subsample and the full dataset become identical when $N = N_{full}$ (Fig. 29). Explanation (2) then prevails. Conversely, smaller subsets are expected to become increasingly independent of the full dataset when N is low compared to N_{full}. Explanation (1) then prevails. However, without independent datasets, it is difficult to evaluate the relative contributions of explanations (1) and (2) to the similarity between the reference model and the model based on a subsample [Fig. 29 illustrates a situation in which explanation (1) completely prevails at low N].

Our study rests upon the assumption that, at small N, the two S curves in Fig. 29 will be very close to each other or even coincide. This assumption is tested using two independent datasets. Our results for *Pieris napi* (Fig. 3) show that already at $N = 100 \approx 0.5 \cdot N_{full}$, the two S curves are very close to merging, indicating that when $N << N_{full}$, eventual non-independence of subsamples from the full dataset does not affect our results. The CSS estimates obtained in this study are in the range of 5–20 presences, well within the range of subsample sizes in which the S curves of subsamples and independent samples coincide or are close to coinciding. We therefore conclude that our results are not influenced by non-independence of subsamples and the full dataset.

Our results based on the pair *Thecla betulae* and Species narrow$_j$ differ from those for *Pieris napi* in that the subsample S curve (for *T. betulae*) and the independent S curve (for Species narrow$_j$) do not converge, even at $N = 5$. However, the two curves correspond to two 'species'



Figs 29-30. Similarity between the models based on the full dataset (i.e., the reference model) and smaller samples as a function of sample size. (6) Similarity between the models based on subsamples and the full dataset (Subsampling *S* curve), compared to similarity between the models based on independent samples and the full dataset (Independent *S* curve). N_{full} is the size of the full dataset. S_{max} is the maximum possible similarity, e.g. when the full dataset is compared with itself. At $N = N_{full} - 1$, a subsample includes all the points of the full dataset except one. The resulting model is thus very similar to the reference model ($S_{Nfull\,-\,1} \approx S_{max}$). When $N << N_{full}$, the two *S* curves are expected to merge. (7) Similarity between the reference models and the models based on smaller samples, for two full datasets of different size ($N_{full(1)} > N_{full(2)}$). Solid line indicates the larger dataset, dashed line indicates the smaller dataset. The gap between the Subsampling *S* curve and the Independent *S* curve is larger (Gap 2 > Gap 1) when the full dataset is smaller ($N_{full(1)}$). As a result, the threshold of the similarity measure S_{thr} may intercept the Subsampling and independent *S* curves at different sample size *N* resulting in different CSS estimates, $N_{CSS(Ind)} > N_{CSS(Subs)}$.

that are ecologically different because Species narrow$_j$ was simulated by sampling within the tolerance limits of *T. betulae* along the two environmental complex-gradients used in this study. A bivariate uniform ('rectangular') distribution in underlying ecological space is a rather crude representation of a species' overall ecological response to two main gradients compared with the expected bivariate unimodal response (cf. Halvorsen 2012). Thus, the set of presence records for Species narrow$_j$ is likely to include unrealistically many presences near the tolerance limits of *T. betulae*, whereby Species narrow$_j$ obtains a broader optimum range and hence stands out as having broader tolerance to the two gradients than *T. betulae*. This is likely to explain the lack of convergence of the *S* curves for the two 'species'.

The size of full datasets for the 30 insect species studied in this paper varies from 41 to 346 presence observations. In theory, the smallest among the full datasets could be so small that even CSS estimates based upon the smallest subsamples were affected by non-independence. Indeed, the gap between the subsampling *S* curve and the independent *S* curve at a given *N* may be expected to be wider when the full dataset is smaller (Fig. 30). Thus, when the full dataset is small, the subsampling *S* curve may cross the *S*-threshold (S_{thr}) line at a smaller sample size N than the independent *S* curve ($N_{css(Subs)} < N_{css(Ind)}$). If this is the case, a negative correlation between CSS and the size of the full dataset is expected. The lack of a significant relationship between CSS and the size of the full dataset in our study (Table 4) indicates that even the smallest of the full datasets were sufficiently large to provide unbiased estimates of CSS.

RELATIONSHIPS BETWEEN CRITICAL SAMPLE SIZE AND SPECIES PROPERTIES

We find no significant relationships between the taxonomic affiliation or relative ecological range on the one hand and the CSS estimates for the thirty species on the other. Differences between insect orders could have been expected, because within the same order species may share life-history and life-cycle properties, dispersal capabilities and other biologically important characteristics. The fact that in our results CSS is affected neither by taxonomic affiliation nor ecological range suggests that our generally low CSS estimate is a robust result, of general validity. Some studies have, however, reported differences in species characteristics, such as distribution patterns, to be of importance for the predictability of species distributions by DM methods (Guisan et al. 2007, Hernandez et al. 2006, Mateo et al. 2010, Stokland et al. 2011). These studies conclude that specialist species, i.e., species with narrow distributions in environmental space and restricted distributions in geographical space, are easier targets for spatial prediction modeling (SPM; Halvorsen 2012) than generalist species, i.e., species with a broader distribution in the two conceptual spaces, and, accordingly, that fewer presence observations are needed to obtain acceptable distribution models for the former. The argument underpinning this view is that generalists have broader ecological requirements and that more presences are needed to represent the entire range of suitable environmental conditions for such species. Our results do, however, suggest that distributions also of broadly distributed species may be reasonably accurately predicted with few presence observations. Contrary to the reasoning above, however, our analyses of the simulated specialist species, Species narrow$_j$, and the widespread species, *P. napi*$_j$, result in a higher CSS for the specialist species.

DOES THE GENERAL CRITICAL SAMPLE SIZE EXIST?

Two important issues have to be taken into account when the CSS results for individual species are interpreted: firstly, that the distribution of observed presences for the species along the underlying gradient in question, expressed by the observed frequency-of-presence (FoP) curve, has to be adequately captured by the available presence observations; and, secondly, that the derived variables that represent the gradient, and/or the modelling method, must be able to fit a function that adequately describes the main features of the FoP curves (Støa et al. 2015). Species responses are typically unimodal, but hinge-shaped, linear, or other types of truncated curve shapes may also occur (Halvorsen 2012, Whittaker 1956). The reasons for the different conclusions in the literature regarding required minimum sample size is likely to be a combination of the two issues raised above and the fact that different criteria for minimum acceptable sample size are used. Our results suggest that 15 presence observations are often sufficient to capture the general response of the species to important environmental variables.

CONCLUDING REMARKS

The present study is part of a project with a broader goal of assessing if distribution modeling is a useful tool for studying rare and endangered insect species. Because that available information about the distribution of rare species is often limited to a few recorded presence observations, it is important to know the minimum number of presence records necessary for generating potentially useful models. As expected, we find that the probability that distribution models trained by use of real data outperform random models increase with the number of presence observations in the dataset. However, our results also show that nonrandom models are obtained in most cases when very few presences are available. Using *IVS*, we find that only 3 out of the 30 modeled species require more than 10 presence observations for a nonrandom model to be obtained, and that more than 15 presence observations are required for only 1 out of 30 species. We therefore propose, as a rule of the thumb, that MaxEnt modeling may be applied to species when at least 10–15 presence observations are available. Furthermore, we consider this to be a conservative recommendation because nonrandom models are obtained for many species with as few as 5 presences. All nonrandom distribution models may be useful by providing insights about species' distributions that cannot easily be obtained from simple inspection of the primary distribution data.

We conclude that sample size is an important factor that may influence the reliability of distribution models. Our study indicates that valuable information about a species' distribution can be obtained even when few presence records are available for distribution modelling.

ACKNOWLEDGEMENTS

We thank Eirik Rindal for valuable help with data processing and for formatting all the figures, Sabrina Mazzoni for helping with GIS-related work, Leif Aarvik for suggesting some of the model species for this study and Anders Endrestøl for extracting data from the insect database at the Natural History Museum in Oslo. We are also grateful to Kaare Aagaard (Museum of Natural History and Archaeology, Norwegian University of Science and Technology, Trondheim) and Bjarte Jordal (Bergen Museum, University of Bergen) for giving us access to collections under their care and allowing us to extract occurrence information for selected species from these collections. Finally, thank Oddvar Pedersen and Einar Timdal for their help in converting geographic coordinates from MGRS to standard UTM format. This study was supported by the Norwegian Research Council MILJØ2015 grant 183318 to Vladimir Gusarov.

REFERENCES

Austin, M. 2007. Species distribution models and ecological theory, a critical assessment and some possible new approaches. Ecological Modelling 200,1–19.

Bakkestuen, V., Erikstad, L., Halvorsen, R. 2008. Step-less models for regional environmental variation in Norway. Journal of Biogeography 35,1906–1922.

Chapman, A. D. 2009. Numbers of living species in Australia and the world. Department of the Environment, Water, Heritage and the Arts, Canberra, Australia.

Chefaoui, R. M., J. Hortal, and J. M. Lobo. 2005. Potential distribution modelling, niche characterization and conservation status assessment using GIS tools, a case study of Iberian *Copris* species. Biological conservation 122,327–338.

Crawley, M. J. 2013. The R book. Second edition. Wiley, Chichester, UK.

Cumming, G. S. 2000. Using between-model comparisons to fine-tune linear models of species ranges. Journal of Biogeography 27,441–455.

Dupin, M., P. Reynaud, V. Jarošík, R. Baker, S. Brunel, D. Eyre, J. Pergl, and D. Makowski. 2011. Effects of the Training Dataset Characteristics on the Performance of Nine Species Distribution Models, Application to *Diabrotica virgifera virgifera*. PLoS ONE 6,e20957.

Edvardsen, A., V. Bakkestuen, and R. Halvorsen. 2011. A fine-grained spatial prediction model for the red-listed vascular plant *Scorzonera humilis*. Nordic Journal of Botany 29,495–504.

Elith, J., C. H. Graham, R. P. Anderson, M. Dudik, 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. M. Overton, A. T. Peterson, S. J. Phillips, K. Richardson, R. Scachetti-Pereira, R. E. Schapire, J. Soberon, S. Williams, M. S. Wisz, and N. E. Zimmermann. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29,129–151.

Elith, J., S. J. Phillips, T. Hastie, M. Dudík, Y. E. Chee, and C. J. Yates. 2011. A statistical explanation of MaxEnt for ecologists. Diversity and Distributions 17,43–57.

Engler, R., A. Guisan, and L. Rechsteiner. 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. Journal

of Applied Ecology 41,263–274.
Feeley, K. J., and M. R. Silman. 2011a. The data void in modeling current and future distributions of tropical species. Global Change Biology 17,626–630.
Feeley, K. J., and M. R. Silman. 2011b. Keep collecting, accurate species distribution modelling requires more collections than previously thought. Diversity and Distributions 17,1132–1140.
Fielding, A. H., and J. F. Bell. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24,38–49.
Fitzpatrick, M. C., J. F. Weltzin, N. J. Sanders, and R. R. Dunn. 2007. The biogeography of prediction error, why does the introduced range of the fire ant over-predict its native range? Global Ecology and Biogeography 16,24–33.
Franklin, J. 2009. Mapping species distributions, spatial inference and prediction. Cambridge University Press, Cambridge.
Guisan, A., O. Broennimann, R. Engler, M. Vust, N. G. Yoccoz, A. Lehmann, and N. E. Zimmermann. 2006. Using niche-based models to improve the sampling of rare species. Conservation Biology 20,501–511.
Guisan, A., and W. Thuiller. 2005. Predicting species distribution, offering more than simple habitat models. Ecology letters 8,993–1009.
Guisan, A., and N. E. Zimmermann. 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135,147–186.
Guisan, A., N. E. Zimmermann, J. Elith, C. H. Graham, S. Phillips, and A. T. Peterson. 2007. What matters for predicting the occurrences of trees, Techniques, data, or species' characteristics? Ecological Monographs 77,615–630.
Halvorsen, R. 2012. A gradient analytic perspective on distribution modelling. Sommerfeltia 35,1–165.
Halvorsen, R. 2013. A maximum likelihood explanation of MaxEnt, and some implications for distribution modelling. Sommerfeltia 36,1–165.
Hanberry, B. B., H. S. He, and D. C. Dey. 2012. Sample sizes and model comparison metrics for species distribution models. Ecological Modelling 227,29–33.
Hastie, T., R. Tibshirani, and J. Friedman. 2009. The elements of statistical learning. Data mining, inference, and prediction. Second Edition. Springer, New York.
Hernandez, 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 29,773–785.
IUCN. 2001. IUCN Red List categories and criteria, version 3.1. IUCN (International Union for Conservation of Nature, Gland, Switzerland.
Jaynes, E. T. 1957a. Information theory and statistical mechanics. Physical Review 106,620–630.
Jaynes, E. T. 1957b. Information theory and statistical mechanics 2. Physical Review 108,171–190.
Kadmon, R., O. Farber, and A. Danin. 2003. A systematic analysis of factors affecting the performance of climatic envelope models. Ecological Applications 13,853–867.
Kamino, L. H. Y., J. R. Stehmann, S. Amaral, P. De Marco, T. F. Rangel, M. F. de Siqueira, R. De Giovanni, and J. Hortal. 2012. Challenges and perspectives for species distribution modelling in the neotropics. Biology Letters 8,324–326.
Kålås, J. A., Å. Viken, S. Henriksen, S. Skjelseth, and (Eds.). 2010. The 2010 Norwegian Red List for Species. Norwegian Biodiversity Information Centre, Trondheim, Norway.
Le Lay, G., R. Engler, E. Franc, and A. Guisan. 2010. Prospective sampling based on model ensembles improves the detection of rare species. Ecography 33,1015–1027.

Lim, B. K., A. T. Peterson, and M. D. Engstrom. 2002. Robustness of ecological niche modeling algorithms for mammals in Guyana. Biodiversity and Conservation 11,1237–1246.

Lobo, J. M., A. Jiménez-Valverde, and J. Hortal. 2010. The uncertain nature of absences and their importance in species distribution modelling. Ecography 33,103–114.

Loe, L. E., C. Bonenfant, E. L. Meisingset, and A. Mysterud. 2012. Effects of spatial scale and sample size in GPS-based species distribution models, are the best models trivial for red deer management? European Journal of Wildlife Research 58,195–203.

Lomba, A., L. Pellissier, C. Randin, J. Vicente, F. Moreira, J. Honrado, and A. Guisan. 2010. Overcoming the rare species modelling paradox, a novel hierarchical framework applied to an Iberian endemic plant. Biological conservation 143,2647–2657.

López-Cárdenas, J., F. E. G. Bravo, P. M. S. Schettino, J. C. G. Solorzano, E. R. Barba, J. M. Mendez, V. Sánchez-Cordero, A. T. Peterson, and J. Ramsey. 2005. Fine-scale predictions of distributions of Chagas disease vectors in the state of Guanajuato, Mexico. Journal of medical entomology 42,1068–1081.

Marini, M., M. Barbet-Massin, L. Lopes, and F. Jiguet. 2010. Predicting the occurrence of rare Brazilian birds with species distribution models. Journal of Ornithology 151,857–866.

Mateo, R. G., A. M. Felicísimo, and J. Muñoz. 2010. Effects of the number of presences on reliability and stability of MARS species distribution models, the importance of regional niche variation and ecological heterogeneity. Journal of Vegetation Science 21,908–922.

Meggs, J. M., S. A. Munks, R. Corkrey, and K. Richards. 2004. Development and evaluation of predictive habitat models to assist the conservation planning of a threatened lucanid beetle, *Hoplogonus simsoni*, in north-east Tasmania. Biological conservation 118,501–511.

Moen, A. 1999. National atlas of Norway, Vegetation. Norwegian Mapping Authority, Hønefoss.

New, T. R. 2009. Insect species conservation. Cambridge University Press, Cambridge, UK.

Papeș, M., and P. Gaubert. 2007. Modelling ecological niches from low numbers of occurrences, assessment of the conservation status of poorly known viverrids (Mammalia, Carnivora) across two continents. Diversity and Distributions 13,890–902.

Pearce, J., and S. Ferrier. 2000. Evaluating the predictive performance of habitat models developed using logistic regression. Ecological Modelling 133,225–245.

Pearson, R. G., C. J. Raxworthy, M. Nakamura, and A. T. Peterson. 2007. Predicting species distributions from small numbers of occurrence records, a test case using cryptic geckos in Madagascar. Journal of Biogeography 34,102–117.

Peterson, A. T., C. Martínez-Campos, Y. Nakazawa, and E. Martínez-Meyer. 2005. Time-specific ecological niche modeling predicts spatial dynamics of vector insects and human dengue cases. Transactions of the Royal Society of Tropical Medicine and Hygiene 99,647–655.

Peterson, A. T., J. Soberón, R. G. Pearson, R. P. Anderson, E. Martínez-Meyer, M. Nakamura, and M. B. Araújo. 2011. Ecological niches and geographic distributions. Princeton University Press, Princeton and Oxford.

Phillips, S. J., R. P. Anderson, and R. E. Schapire. 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190,231–259.

Phillips, S. J., and M. Dudík. 2008. Modeling of species distributions with Maxent, new extensions and a comprehensive evaluation. Ecography 31,161–175.

Phillips, S. J., M. Dudík, and R. E. Schapire. 2004. A maximum entropy approach to species distribution modeling.in Proceedings of the twenty-first international conference on machine learning. ACM, New York.

Raes, N., and H. ter Steege. 2007. A null-model for significance testing of presence-only species distribution models. Ecography 30,727–736.

Rebelo, H., and G. Jones. 2010. Ground validation of presence-only modelling with rare species, a case study on barbastelles *Barbastella barbastellus* (Chiroptera, Vespertilionidae). Journal of Applied Ecology 47,410–420.

Reese, G. C., K. R. Wilson, J. A. Hoeting, and C. H. Flather. 2005. Factors affecting species distribution predictions, A simulation modeling experiment. Ecological Applications 15,554–564.

Renner, I. W., and D. I. Warton. 2013. Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. Biometrics 69,274–281.

Roura-Pascual, N., A. V. Suarez, C. Gómez, P. Pons, Y. Touyama, A. L. Wild, and A. T. Peterson. 2004. Geographical potential of Argentine ants (*Linepithema humile* Mayr) in the face of global climate change. Proceedings of the Royal Society of London. Series B, Biological Sciences 271,2527–2535

SPWG. 2006. Guidelines for Using the IUCN Red List Categories and Criteria. Version 6.2. IUCN Gland, Switzerland; Cambridge, UK.

Stockwell, D. R. B., and A. T. Peterson. 2002. Effects of sample size on accuracy of species distribution models. Ecological Modelling 148,1–13.

Stokland, J. N., R. Halvorsen, and B. Støa. 2011. Species distribution modelling—Effect of design and sample size of pseudo-absence observations. Ecological Modelling 222,1800–1809.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological),267–288.

Van der Vaart, A. W. 1998. Asymptotic statistics. Cambridge University Press, Cambridge, UK.

Veloz, S. 2009. Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. Journal of Biogeography 36,2290–2299.

Whittaker, R. H. 1956. Vegetation of the great smoky mountains. Ecological Monographs 26,1–80.

Wisz, M. S., R. J. Hijmans, J. Li, A. T. Peterson, C. H. Graham, and A. Guisan. 2008. Effects of sample size on the performance of species distribution models. Diversity and Distributions 14,763–773.

Wollan, A. K., V. Bakkestuen, H. Kauserud, G. Gulden, and R. Halvorsen. 2008. Modelling and predicting fungal distribution patterns using herbarium data. Journal of Biogeography 35,2298–2310.