International Journal of Geographical Information Science, 2014
Vol. 28, No. 8, 1723–1739, http://dx.doi.org/10.1080/13658816.2014.891222

Taylor & Francis
Taylor & Francis Group

How many predictors in species distribution models at the landscape scale? Land use versus LiDAR-derived canopy height

Gentile Francesco Ficetola[a]*, Anna Bonardi[a], Caspar A. Mücher[b], Niels L.M. Gilissen[c] and Emilio Padoa-Schioppa[a]

[a]*Dipartimento di Scienze dell'Ambiente e del Territorio, e di Scienze della Terra, Università degli Studi di Milano Bicocca, Milano, Italy;* [b]*Alterra, Wageningen UR, Wageningen, The Netherlands;* [c]*Dienst Vastgoed Defensie, Wageningen, The Netherlands*

(*Received 12 March 2013; accepted 29 January 2014*)

At the local spatial scale, land-use variables are often employed as predictors for ecological niche models (ENMs). Remote sensing can provide additional synoptic information describing vegetation structure in detail. However, there is limited knowledge on which environmental variables and how many of them should be used to calibrate ENMs. We used an information-theoretic approach to compare the performance of ENMs using different sets of predictors: (1) a full set of land-cover variables (seven, obtained from the LGN6 Dutch National Land Use Database); (2) a reduced set of land-cover variables (three); (3) remotely sensed laser data optimized to measure vegetation structure and canopy height (LiDAR, light detection and ranging); and (4) combinations of land cover and LiDAR. ENMs were built for a set of bird species in the Veluwe Natura 2000 site (the Netherlands); for each species, 26–214 records were available from standardized monitoring. Models were built using MaxEnt, and the best performing models were identified using the Akaike's information criterion corrected for small sample size (AICc). For 78% of the bird species analysed, LiDAR data were included in the best AICc model. The model including LiDAR only was the best performing one in most cases, followed by the model including a reduced set of land-use variables. Models including many land-use variables tended to have limited support. The number of variables included in the best model increased for species with more presence records. For all species with 33 records or less, the best model included LiDAR only. Models with many land-use variables were only selected for species with >150 records. Test area under the curve (AUC) scores ranged between 0.72 and 0.92. Remote sensing data can thus provide regional information useful for modelling at the local and landscape scale, particularly when presence records are limited. ENMs can be optimized through the selection of the number and identity of environmental predictors. Few variables can be sufficient if presence records are limited in number. Synoptic remote sensing data provide a good measure of vegetation structure and may allow a better representation of the available habitat, being extremely useful in this case. Conversely, a larger number of predictors, including land-use variables, can be useful if a large number of presence records are available.

Keywords: birds; ecological niche models; land use; habitat suitability modelling; model performance; model selection; selection of variables

1. Introduction

Correlative ecological niche models (ENMs; often referred to as species distribution models) analyse relationships between species distribution data and environmental

*Corresponding author. Email: francesco.ficetola@unimib.it features. ENMs allow the assessment of suitability of a given area for one or multiple species and provide important information on ecological factors determining species distributions (Sillero 2011). The output of ENMs is increasingly used for multiple purposes, including the identification of conservation priorities, the prediction of species invasions and analyses of the impact of environmental changes on biodiversity (Elith and Leathwick 2009).

ENMs can be performed at many spatial scales, ranging from local (e.g. one single reserve) to regional, continental and global. ENMs require both species distribution data (e.g. presence/absence, abundance or presence-only data) and relevant environmental data. Among environmental data, abiotic variables (e.g. climate) tend to be more important in models analysing distribution at broad scales, such as in continental or global models, while variables representing habitat, landscape, vegetation or biotic interactions can be more important at finer spatial scales (Soberon and Nakamura 2009, Boulangeat *et al.* 2012). Processes acting at multiple scales can also interact among them: for example, a large region can have a suitable climate for a given species but, within this region, the target species can attain positive fitness only in the areas with certain landscape features, or with appropriate resources (Anadón *et al.* 2006, Soberon and Nakamura 2009, Ficetola *et al.* 2010, Boulangeat *et al.* 2012, Brambilla and Ficetola 2012, Gallien *et al.* 2012). Models implemented at the landscape scale are particularly relevant to guide management planning in protected areas. For instance, land-use features that are positively associated with species of conservation concern could be favoured in such cases, while those increasing the risk of invasion by alien species could be limited (Brambilla *et al.* 2010, Ficetola *et al.* 2010).

Nevertheless, the identification of relevant environmental variables for modelling species distributions is a complex task that is frequently underestimated (Seoane *et al.* 2004b, Synes and Osborne 2011, Williams *et al.* 2012). The number of predictors to be included in an ENM depends in part on the number of observation data points that are used to calibrate the model (Rushton *et al.* 2004), but many studies include a very large number of predictors, independently of the number of training data. For example, several studies analysing climatic suitability used the 19 'bioclimate' variables of the WorldClim data set as predictors (Hijmans *et al.* 2005). However, over-fitting the model is a risk of including too many predictors and may limit the ability of models to perform predictions under different conditions (i.e. model transferability) (Peterson and Nakazawa 2008, Rödder *et al.* 2009, Synes and Osborne 2011). The situation may be even more complex for models that consider habitat variables. Frequently, landscape variables are derived from land-use and land-cover (LULC) digital databases. Researchers identify the land-use classes present within the study area. The percentage cover of a given land-use class, or the dominant class in a given cell, is then used as environmental predictor in the models (e.g. Seoane *et al.* 2004b, Brambilla *et al.* 2010, Ficetola *et al.* 2010, Morán-Ordóñez *et al.* 2012). LULC databases can distinguish a large number of classes, and it is not always easy to identify *a priori* how many (and which) land-use categories are actually relevant and should be included in the analyses. Furthermore, remote sensing techniques are a potential tool to map the vegetation structure more accurately and at a high resolution, providing information on environmental features that are relevant for species distribution and fitness. This information can be added to ENMs in addition to standard land-use categories, and might even substitute the traditional data (Seoane *et al.* 2004a, Morán-Ordóñez *et al.* 2012, Tattoni *et al.* 2012, Bunce *et al.* 2013).

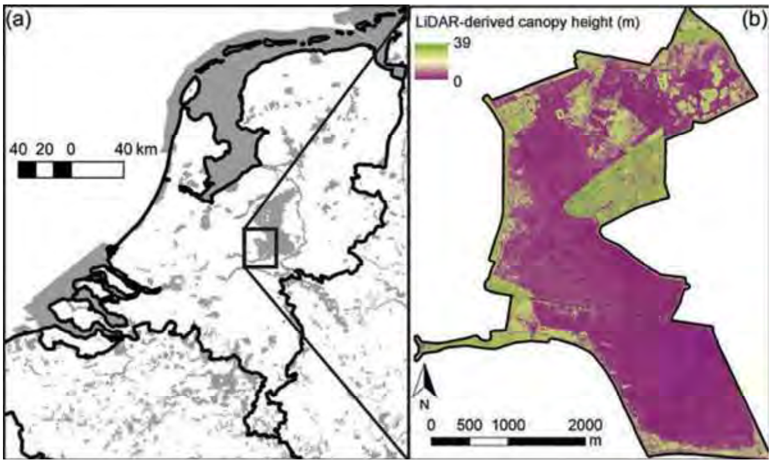

Figure 1. Study area. (a) Location of the Veluwe site within the Netherlands; the grey shading represents the Natura 2000 network. (b) LiDAR-derived canopy height values within the Veluwe site.

Nevertheless, there are limited guidelines on the best approach to ENMs at the landscape scale. Is it better to use a large number of LULC classes, even if some of them might be poorly represented in the landscape? Or is it better to consider only a limited number of coarse or dominant land-use categories? Is it better to use other synoptic information obtained from remote sensing (such as normalized difference vegetation index, surface reflectance or LiDAR (light detection and ranging); Lefsky *et al.* 2002, Seoane *et al.* 2004a, Goetz *et al.* 2007, Morán-Ordóñez *et al.* 2012, Tattoni *et al.* 2012), or is it better to use land-cover maps classifying the land use into discrete categories that may represent habitats?

In this study, we used an information-theoretic approach to identify the landscape variables that are more appropriate to build ENMs for birds in a protected site of the Netherlands. The study area is small and there are no climatic or topographic gradients (Figure 1); for this reason, the factors limiting the birds' distributions are likely landscape-related. We compared five approaches to the selection of environmental variables as predictors into ENMs at landscape scale: (1) models using a relatively large number of traditional land-use variables; (2) models using a small number of land-use variables; (3) models excluding land-use variables, and considering only canopy height data obtained through remote sensing (LiDAR; Lefsky *et al.* 2002, Vierling *et al.* 2008). Furthermore, we built models combining traditional land-use variables and LiDAR data: (4) models using both LiDAR and a large number of land-use variables and (5) models using both LiDAR and a small number of land-use variables.

2. Material and methods

2.1. Study area, environmental and species distribution data

This case study has been conducted within the framework of Biodiversity Multi-SOurce Monitoring System: From Space To Species (BIO_SOS) project (www.biosos.eu), a 3-year EU-PF7 research project aiming at developing a pre-operational system for an effective monitoring of changes in the land cover and habitats within and along the borders of protected areas, to judge their effectiveness in protecting and conserving the regions from human impacts.

The study area, Ederheide and Ginkelse heide, is located within the Dutch Natura 2000 site 'Veluwe' (NL9801023 + NL3009017) in the Province of Gelderland (Figure 1). The Veluwe covers approximately 91,200 ha; it is constituted by sand dune areas alternated with heathlands and dry forests and moulded by a long history of intensive land use. The Ederheide and Ginkelse heide spread over an area of approximately 1000 ha; heathland is the dominant land cover (Mücher *et al.* 2013).

Land-use/land-cover data were obtained from the sixth version of Landelijk Grondgebruiksbestand Nederland (LGN6) – Dutch National Land Use database with a 25 m spatial resolution, which is based on the integration of satellite images (e.g. Landsat) collected in 2007/2008 with Dutch topographical maps and databases of geographical data and natural areas (Hazeu *et al.* 2011). The full set of land-use variables used in ENMs included seven categories: broadleaved forest, coniferous forest, heathland, grassland, sparse vegetation, built-up and shifting sand. We also considered a reduced set of land-use variables, including only the three most representative classes within the study area: broadleaved forest, coniferous forest and heathland. To calculate the cover of the land-use classes, the study area was partitioned in 20 m × 20 m cells. For each cell, we measured land-use variables as the average cover of the habitat categories considered, calculated in a 100 m radius from the cell centre (Brambilla and Ficetola 2012). For LiDAR, we calculated the average value in a 100 m radius from the cell centre. Therefore, land cover was represented by seven continuous variables, each representing the percentage cover within 100 m around each cell.

We used data obtained from LiDAR as it can measure the three-dimensional distribution of plant canopies, and thus be used to estimate the structural features of vegetation (Lefsky *et al.* 2002, Vierling *et al.* 2008). LiDAR data have been used in the Netherlands since early 2000 for the construction of detailed elevation models. The recently acquired AHN-2 (Actueel Hoogtebestand Nederland) has a height precision of 5 cm and 10 measured points per square metre. The original data from AHN-2 used in this study were acquired in spring 2010 by Fugro Aerial Mapping BV. Fugro used the FLI-MAP 400 system and is carried on board of a helicopter, integrated with high-resolution photograph and video camera and a GPS system; the average number of points per square metre was approximately 15. The absolute accuracy for a single point is 3 cm or better. The canopy height model (CHM) was derived from the LiDAR LAS files using the multiscale curvature classification (MCC)-LiDAR software (Evans and Hudak 2007) and LAStools (rapidlasso GmbH, Gilching, Germany; http://www.lastools.org/) for the classification of ground points.

Bird data were collected by the fauna specialists of the Netherlands Ministry of Defence in scheduled surveys of standardized monitoring. The goal of this monitoring was to obtain results that represent the spatial distribution of the sampled species in the study area as a tool for habitat management. Monitoring was performed using the standard territory mapping method which is among the best and most commonly used methods to obtain these data (Bibby *et al*. 2000, Gregory *et al.* 2004). The study area was sampled in a homogenous way and with a complete coverage, in order to have a constant survey effort over the whole area and obtain non-biased distribution data for the species. The fieldwork was executed four times in the morning during the period 15 March–15 June 2009 (breeding season of birds within the study area). The location of all detected birds was recorded, resulting in spatial data on the distribution of bird territories of each species. Although not all the individuals can be detected, the homogeneity of the survey will result in a relative spatial distribution over the area that represents the actual distribution for each species (Bibby *et al.* 2000, Gregory *et al.* 2004).

2.2. *Statistical analyses*

Before running analyses, we calculated correlation between environmental variables. Correlation coefficients between independent variables $|r| > 0.7$ can make it difficult the interpretation of species/habitat relationships. Furthermore, we used variance inflation factor (VIF) to evaluate whether multicollinearity occurs in our models. VIF values >10 indicate that multicollinearity may make the interpretation of the effect of variables within models problematic (Zuur *et al.* 2010). Nevertheless, significance of individual variables was not the focus of this study; correlation between variables is not a major problem when the focus understands the processes within a given area, and Akaike's information criterion corrected for small sample size (AICc)-based model selection has been shown to have good performance even in presence of collinearity (Murtaugh 2009).

We used maximum entropy modelling (MaxEnt) (Phillips *et al.* 2006, Elith *et al.* 2011) to build models relating bird occurrence data to the land-use and LiDAR data. MaxEnt is a presence-background machine-learning approach that assesses the suitability in a given cell on the basis of environmental features in that cell; comparative analyses showed that MaxEnt is among the most efficient approaches to ENMs (Elith *et al.* 2006, 2011). The program establishes flexible relationships between the dependent (species presence) and independent variables, and is well suited to evaluate complex or non-linear relationships. MaxEnt analyses the realized niche of species (Sillero 2011), and this approximation is often similar to that of correlative models using presence and absence records (Elith *et al.* 2011); MaxEnt output represents the suitability of habitats (Sillero 2011). MaxEnt shows good performance even with limited sample size (Wisz *et al.* 2008); therefore, we built models for the bird species for which we obtained 26 or more records within the study area (Table 1). Models were run with linear, quadratic and hinge features using default regularization settings, as they optimize the ability of the model to predict independent test data (Phillips and Dudík 2008, Warren and Seifert 2011). The output of MaxEnt models is a monotone scale (i.e. order preserving), but it

Table 1. Species with at least 26 presence points, and results of comparison between models; models are listed according to increasing AICc values.

		Models				Autocorrelation	
Species	*N*	Variables included	AICc	ΔAICc	*w*	*I*	*P*
Alauda arvensis	214	All variables + LiDAR	3941.00	0.00	0.999	0.024	0.486
		All land-use variables	3954.47	13.47	0.001		
		Reduced land use + LiDAR	3957.78	16.78	0.000		
		Reduced land use	3964.42	23.42	0.000		
		LiDAR only	4019.19	78.19	0.000		
Anthus pratensis	154	All variables + LiDAR	2866.74	0.00	0.643	−0.039	0.384
		Reduced land use	2867.94	1.20	0.353		
		All land use variables	2877.18	10.43	0.003		

(*Continued*)

Table 1. (Continued).

		Models				Autocorrelation	
Species	N	Variables included	AICc	ΔAICc	w	I	P
		Reduced land use + LiDAR	2879.84	13.09	0.001		
		LiDAR only	2934.75	68.00	0.000		
Anthus trivialis	156	LiDAR only	2927.81	0.00	0.784	0.02	0.581
		Reduced land use + LiDAR	2930.38	2.58	0.216		
		All variables + LiDAR	2950.00	22.19	0.000		
		Reduced land use	2956.09	28.28	0.000		
		All land-use variables	2963.01	35.20	0.000		
Saxicola rubicola	75	LiDAR only	1393.56	0.00	0.997	−0.144	0.019
		Reduced land use + LiDAR	1405.01	11.45	0.003		
		Reduced land use	1420.11	26.56	0.000		
		All variables + LiDAR	1420.57	27.01	0.000		
		All land-use variables	1437.10	43.55	0.000		
Turdus philomelos	26	LiDAR only	1393.56	0.00	0.997	0.181	0.048
		Reduced land use	1405.01	11.45	0.003		
		All land-use variables	1420.11	26.56	0.000		
		Reduced land use + LiDAR	1420.57	27.01	0.000		
		All variables + LiDAR	1437.10	43.55	0.000		
Lophophanes cristatus	39	Reduced land use	717.32	0.00	0.789	−0.038	0.771
		LiDAR only	719.96	2.64	0.211		
		Reduced land-use + LiDAR	731.92	14.59	0.001		
		All land use variables	742.77	25.45	0.000		
		All variables + LiDAR	761.92	44.60	0.000		
Certhia brachydactyla	33	LiDAR only	556.73	0.00	0.999	0.005	0.823
		Reduced land use	571.24	14.52	0.001		
		Reduced land use + LiDAR	583.87	27.14	0.000		
		All land-use variables	602.37	45.65	0.000		
		All variables + LiDAR	619.07	62.35	0.000		
Carduelis cannabina	33	LiDAR only	641.61	0.00	0.906	−0.042	0.767
		Reduced land use + LiDAR	646.64	5.03	0.073		
		Reduced land use	649.21	7.60	0.020		
		All variables + LiDAR	657.26	15.65	0.000		
		All land-use variables	686.98	45.37	0.000		
Emberiza citrinella	58	Reduced land use	1107.41	0.00	0.997	−0.037	0.681
		All land use variables	1119.26	11.85	0.003		
		LiDAR only	1129.01	21.60	0.000		
		Reduced land use + LiDAR	1132.15	24.74	0.000		
		All variables + LiDAR	1140.66	33.26	0.000		

Note: *N*, number of occurrences; AICc, Akaike's Information Criterion corrected for small sample size; ΔAICc, difference in AICc units from the best model; *w*, model AICc weight; *I*, Moran's *I* and associated significance.

does not represent the probability of presence of species (Elith *et al.* 2011). Some researchers have proposed approaches to estimate prevalence and derive presence probability on the basis of presence/background data (Li *et al.* 2011, Royle *et al.* 2012), but this remains an open and controversial field of research (Phillips and Elith 2013).

For each species, we considered five MaxEnt models, using different sets of independent variables: (1) using all the seven land-use variables; (2) using only the three dominant land-use variables (i.e. heathland, coniferous and broadleaved forest cover); (3) using LiDAR vegetation height data only; (4) using the seven land-use variables and LiDAR; (5) using the three more represented land-use variables and LiDAR. We then used an information-theoretic approach, based on AICc to identify the best model for each species. AICc trades off explanatory power versus model complexity; parsimonious models explaining more variation have the lowest AICc value and are considered to be the 'best-AICc model'. Simulations suggested that AICc allows to identify the models with highest generality and transferability better than using other approaches such as cross-validation (Warren and Seifert 2011). Models were then ranked on the basis of their ΔAICc, which represents the difference in AICc units between the best model and the model of interest. Models with a low rank according to AICc are therefore the ones with the best performance. For each model, we also calculated the AICc weight *w*, which represents the average support of the model; models with high weight have the highest support (Burnham and Anderson 2002). We calculated AICc values of MaxEnt models using ENMTools 1.3 (Warren *et al.* 2010, Warren and Seifert 2011). Analyses showed that other approaches, such as the area under the curve (AUC) of the receiver operator plot, do not allow a reliable model selection for presence/background models (Smith 2013).

AICc values are calculated using .lambdas files of individual models (Warren *et al.* 2010), and were here calculated for models run using all available data for training. However, MaxEnt models sample points from the background, assuming that presence records are random samples of localities where a given species is present (Elith *et al.* 2011). For each species, we therefore run five replicated models, each time using 80% of presence data to calibrate models, and setting apart 20% of data (test data) (Nogués-Bravo 2009). Compared to cross-validation approaches using more replicates (e.g. 10 replicates), running five replicate models has the advantage that a larger number of data are used for testing at each replicate. As a measure of discrimination capacity, for each model we calculated AUC, averaged over the five replicated runs, as well as its standard deviation. Models with AUC = 0.5 discriminate no better than random, and discrimination improves as AUC approaches a value of 1; low values of standard deviation indicate that models using different training and background points have similar results (Manel *et al.* 2001). Running replicated models also allowed us to use cross-validation to assess predictive performance. In replicated models, the test data were used to assess predictive performance, each time using a different set of test data (Nogués-Bravo 2009). The AUC for the test data was calculated and averaged over the five runs. Furthermore, for each of the replicates we used a *Z*-test comparing observed frequencies of correct and incorrect predictions to evaluate if our models predict distribution significantly better than expected by chance. We converted the MaxEnt suitability scores to binary values, assuming that a cell is suitable for a given species if it has suitability larger than the equal training sensitivity plus specificity threshold (Bartel and Sexton 2009).

Spatial autocorrelation, arising for instance from clustering of presence points, might influence the results of ENMs (Veloz 2009). We therefore assessed whether our data are affected by spatial autocorrelation. First, for each species we randomly generated a number of pseudo-absence points equal to the number of presence points. We then extracted the suitability value, as predicted by the best-AICc model, for the presence and pseudo-absence points of each species. We calculated residuals as the difference between observed presence/pseudo absence (1/0) and the predicted suitability extracted at these points. We then used Moran's I to assess spatial autocorrelation of residuals. A Gabriel's graph was used to define the set of neighbours for each point (Legendre and Legendre 1998); Moran's I was then calculated in SAM 4.0 (Rangel *et al.* 2010). As nine tests were performed, significance values were adjusted using sequential Bonferroni's correction, and we set $\alpha = 0.006$ (Legendre and Legendre 1998).

3. Results

We obtained data for nine bird species with between 26 and 214 records per species (Table 1): sky lark *Alauda arvensis*; meadow pipit *Anthus pratensis*; tree pipit *Anthus trivialis*; European stonechat *Saxicola rubicola*; song thrush *Turdus philomelos*; crested tit *Lophophanes cristatus*; short-toed treecreeper *Certhia brachydactyla*; linnet *Carduelis cannabina* and yellowhammer *Emberiza citrinella*. Out of these species, the sky lark is listed in the annex II of the EU Bird Directive (2009/147/EC) and is a SPEC3 species (species of European conservation concern not concentrated in Europe but with an unfavourable conservation status); the crested tit and the linnet are SPEC2 species (concentrated in Europe and with an unfavourable conservation status, see BirdLife International 2004). Pairwise correlations between environmental variables generally showed $|r| < 0.7$, indicating lack of multicollinearity problems. However, LiDAR-derived canopy height was strongly and positively related to coniferous forests, and strongly and negatively related to heathland. Furthermore, there was a strong, negative correlation between coniferous forest and heathland (Table 2).

For seven of the nine species (i.e. 78% of species), LiDAR was included in the best AICc model. For 56% of species, the best AICc model included LiDAR only and did not include any land-use variable. The reduced set of land-use variables was included into the best model for 22% of species, while for 22% of species the best model

Table 2. Pairwise Pearson's correlation between environmental variables.

	Shifting sand	Sparse vegetation	Built-up	Coniferous forest	Deciduous forest	Grassland	Heathland
LiDAR	−0.02	−0.04	0.09	**0.90**	0.65	0.15	**−0.91**
Shifting sand		−0.02	−0.02	0.00	−0.04	−0.01	−0.01
Sparse vegetation			0.03	−0.06	−0.02	0.10	−0.17
Built-up				0.05	0.09	0.21	−0.25
Coniferous forest					0.39	0.06	**−0.88**
Deciduous forest						0.28	−0.67
Grassland							−0.31

Note: In bold, correlations > 0.7.

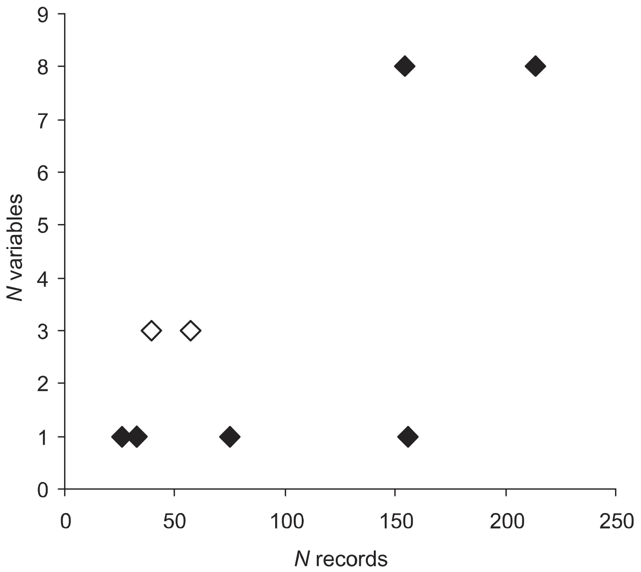

Figure 2. Relationship between number of records per species and number of predictors included in the best AICc model. Filled diamonds: LiDAR included in the best model; empty diamonds: only land-use variables included into the best model.

included both the full set of land-use variables and LiDAR data. Species with more presence points tended to include a larger number of variables in the best model (Pearson's correlation = 0.73, P = 0.026; Figure 2). For all the species with less than 38 records, the best model included LiDAR as the unique predictor; up to three variables were included in best models of species with 38–75 records, while in two species with >150 records the best model included eight predictors (Figure 2). VIF was >10 for some variables in the best models of three species: *A. arvensis* (variables: heathland and sparse vegetation), *L. cristatus* (variables: coniferous forest and heathland) and *E. citrinella* (variables: coniferous forest and heathland). LiDAR did not show VIF > 10 in any of the best models. For all species, Moran's I of residuals was small ($-0.144 \leq I \leq 0.18$), and spatial autocorrelation was not significant after Bonferroni's correction (Table 1). Species occurrences and suitability maps for the study species are shown in Figure 3.

Overall, the models including LiDAR as the sole predictor tended to be the ones with best performance, as they generally showed low rank (low rank is better) (Figure 4a) and had the highest average weight (Figure 4b). Conversely, the models including all the nine land-use variables were consistently those with poorest performance, as they showed high rank and low AICc weight (Figure 4).

The five replicated runs of each species yielded very similar results for all species. For training data, standard deviation of AUC across the five runs was always ≤0.025, suggesting minor differences among replicated runs (Table 3). Cross-validation indicated that the models for all species showed AUC ≥ 0.72; models predicted test data significantly better than expected by chance (Table 2). Predictive performance was not correlated to the number of data available for each species (r = –0.30, P = 0.44) nor to the number of predictors included into the best AICc model (r = –0.19, P = 0.62).

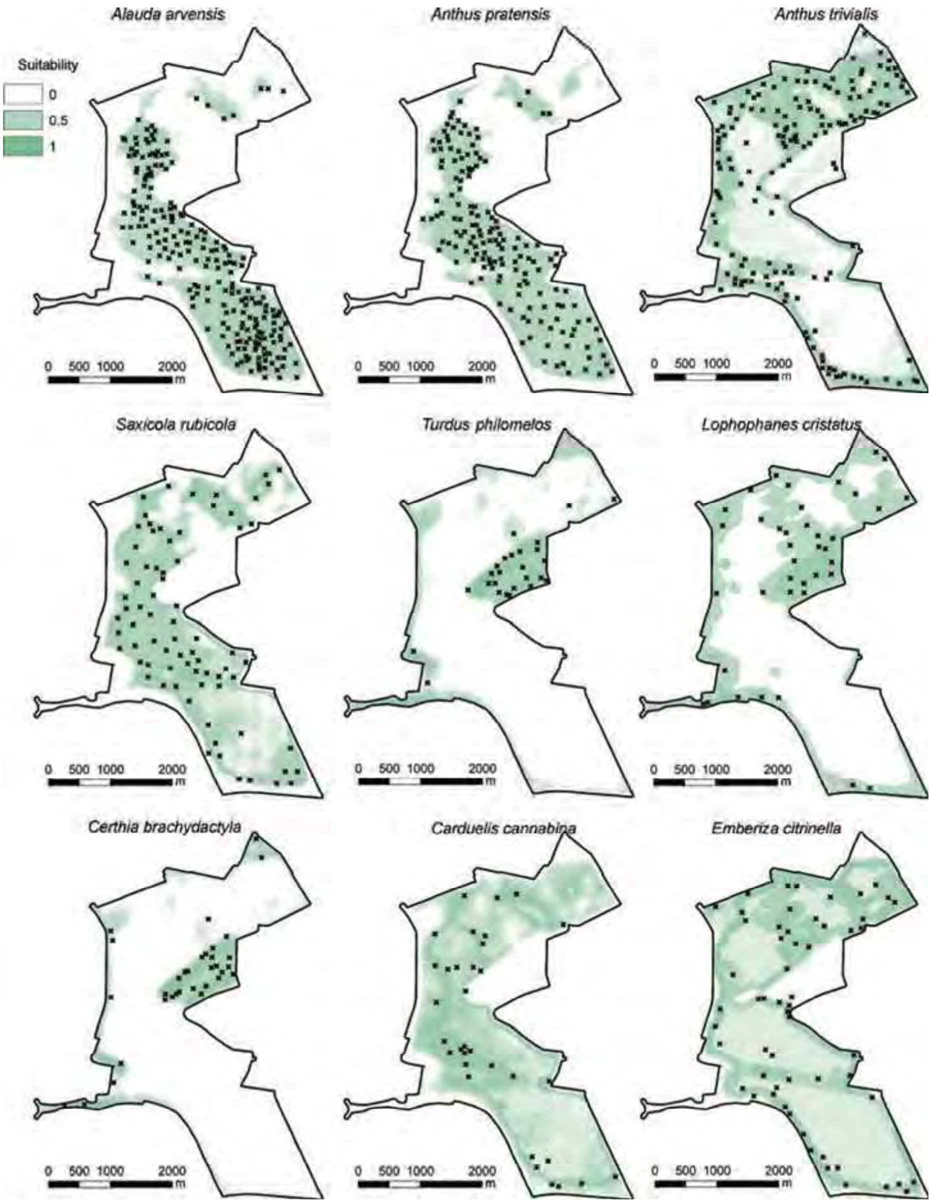

Figure 3. Suitability map for the nine species. The presence points used for model calibration are also shown.

4. Discussion

Land-use and vegetation variables are often used as predictors in ENMs, but limited guidelines are available on which environmental predictors and how many of them would allow building the models at the landscape scale (Seoane *et al.* 2004a, 2004b). Both remotely sensed land cover and LiDAR-derived vegetation structure data can be useful, but the use of a small number of predictors describing the habitat requirements of species results in ENMs with the best performance (see Rödder *et al.* 2009 for similar results with

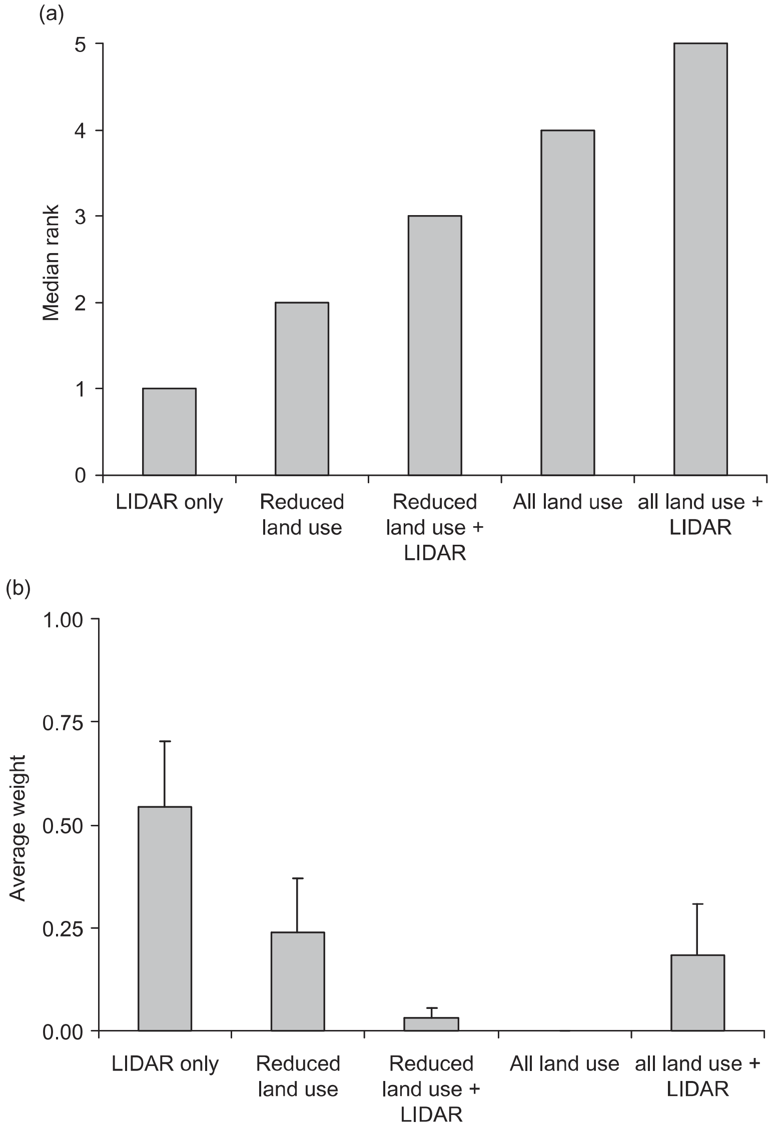

Figure 4. Relative importance of the five candidate models, averaged across species. In (a) the importance is assessed as the median rank of the model according to AICc (lower is better); in (b) the importance is assessed as the average AICc weight of the model (higher is better). For weight, error bars are standard errors of the mean.

climatic variables). Traditional land-use maps can certainly provide important information, but the LiDAR-derived CHM improves the performance of ENMs for birds. Actually, for 56% of the nine species, the LiDAR CHM data alone performed better than any of the traditional land-cover classes (Table 1, Figure 4). Furthermore, for species in which both LiDAR and land-use variables were included in the best model, LiDAR was always among the most important variables (Table 3). The small number of species and the size of the study area limit the generality of conclusions about the species. Nevertheless, our conclusions on the utility of LiDAR are expected to hold over larger

Table 3. For each species, we report the environmental variables included in the best-AICc model explaining more variation (percentage contribution >10%; contribution of variables reported in parentheses), the model performance with training data and the predictive performance calculated using cross-validation.

Species	Important variables	Training AUC	SD	Test AUC	SD	Prediction of test data (*P*)
Alauda arvensis	LiDAR (57%), coniferous forest (38%)	0.815	0.007	0.800	0.025	<0.001
Anthus pratensis	Coniferous forest (62%), LiDAR (39%)	0.802	0.006	0.777	0.029	<0.001
Anthus trivialis	LiDAR*	0.761	0.003	0.752	0.012	<0.001
Saxicola rubicola	LiDAR*	0.784	0.008	0.777	0.037	<0.001
Turdus philomelos	LiDAR*	0.906	0.005	0.906	0.019	<0.001
Lophophanes cristatus	Coniferous forest (99%)	0.859	0.008	0.785	0.027	<0.001
Certhia brachydactyla	LiDAR*	0.922	0.007	0.920	0.027	<0.001
Carduelis cannabina	LiDAR*	0.767	0.025	0.746	0.088	<0.001
Emberiza citrinella	Heatland (92%)	0.759	0.024	0.716	0.085	<0.001

Note: *LiDAR was the only variable included into the model, so percentage importance of variables was not estimated.

areas, because LiDAR is a more direct measure of the habitat used by the birds than are the choropleth-based land-use maps.

Vegetation structure affects bird distribution in multiple ways, by providing shelter, potential sites for nest and foraging habitats. Therefore, variables that accurately describe vegetation structure are excellent predictors of bird distribution (Seoane *et al.* 2004b, Goetz *et al.* 2007). Conversely, LULC maps are usually produced by governmental agencies for general purposes, and they are often not detailed enough to accurately describe the habitat required by species. ENMs obtained from land-cover maps often have good performance, but this is at least in part related to the large number of parameters included in the models (Seoane *et al.* 2004a, 2004b). Variables representing vegetation structure and obtained from remote sensing, such as CHMs, may provide a more accurate representation of the habitat actually available for species and therefore help build better models, and plant height is a variable that can influence birds more directly than land-cover class (Seoane *et al.* 2004a, Bradbury *et al.* 2005, Morán-Ordóñez *et al.* 2012, Tattoni *et al.* 2012). For instance, for species living in open habitats (e.g. pipits, genus *Anthus*), fine scale variation of vegetation structure determines environmental differences, with important consequences for habitat selection. The tree pipit, *A. trivialis*, requires high places within the breeding territories, which are used as song and lookout posts, and avoids areas without shrubs, small woods or isolated trees. Conversely, *A. pratensis* prefers open grasslands with dense, low vegetation, and avoids both areas with very short grass and areas with shrubs and trees (Kumstatova *et al.* 2004). The very fine vertical resolution of LiDAR (5 cm) can better capture the variation of habitats determining the distribution of these species. Land-use classes may provide better results for birds that are specialist of habitats easily categorized, such as the crested tit *Lophophanes cristatus* which is a specialist of coniferous forests (Table 3)

(Atiénzar *et al.* 2009). In this case, the land-use classes provide an adequate representation of habitats, while LiDAR does not improve land-use information, as coniferous and deciduous forests with the same canopy height may be confounded. Additional advantages of remote sensing data, such as LiDAR, is that they synthesize complex information on vegetation structure in one or very few variables, and this results in parsimonious ENMs using a few parameters.

Vegetation attributes and structure information help to understand ecological functions and habitat availability, because they provide a synoptic measure of vegetation structure (Tattoni *et al.* 2012). These measures of canopy metrics and forest structure have been proved to be strong predictors of species richness or presence/absence for birds in several studies (Goetz *et al.* 2007, Vierling *et al.* 2008, Tattoni *et al.* 2012), and make available information even in difficult terrain (Hyde *et al.* 2005). The correlation between LiDAR-derived estimates of vegetation structure and bird distribution has been demonstrated in multiple habitats (Bradbury *et al.* 2005, Goetz *et al.* 2007), and may even provide indication about territories and breeding success (Bergen *et al.* 2009). LiDAR has a great potential for effective habitat monitoring and management of endangered species (Graf *et al.* 2009), and LiDAR-based habitat classification may surpass results obtained with optical data (Korpela *et al.* 2009). The result of habitat analysis obtained with LiDAR may also be enhanced when used in combination with spectral data (Hyde *et al.* 2006, Clawges *et al.* 2008). Overall, LiDAR remote sensing shows considerable efficacy for habitat mapping/characterization and wildlife management, allowing fine detail even across broad areas.

The number of parameters in the best models tended to increase with sample size (Figure 2). Only one parameter (i.e. LiDAR) was included in the best model for the three species with less than 38 records, while the full set of land-use variables was selected for species with many records only (Table 1, Figure 2). In our study system, three of the nine species had a relatively low number of records (26–33, Table 1). The production of good ENMs with low sample size is dependent on the modelling method, and MaxEnt is among the techniques with best performance with low sample sizes (Alvarez and Brito 2006, Pearson *et al.* 2007, Wisz *et al.* 2008, Ficetola *et al.* 2009). This is particularly frequent for rare or endangered species, or in invasive species at the early stages of invasions. Actually, species that are invasive or of conservation concern are those species for which the output of ENMs is particularly helpful for management planning (Pearson *et al.* 2007, Ficetola *et al.* 2009); therefore, the conclusions obtained for species with few records may have relevant implications.

The number of variables that should be included in ENMs can dramatically decrease with sample size, and very few predictors (1–3) may be included in models if the number of records is small. This probably occurs because sample size greatly influences the statistical power of analyses, and models with many parameters and limited presence points receive high penalties during the calculation of AICc (Burnham and Anderson 2002). Furthermore, rare species are often those for which few presence points are available. Rare species frequently are habitat specialists and might therefore be predictable from a few habitat variables, while widespread and more abundant species occupy a broader range of habitats, and thus more combinations of variables might be required to predict their occurrence. Synoptic predictors such as LiDAR CHM may be particularly relevant when presence points are limited, as one single environmental variable provides a comprehensive and accurate measure of vegetation structure, and can allow to build more parsimonious models. Nevertheless, it should be noted that our study focused on nine species. Analyses considering more species are needed to assess the generality of our results. Furthermore, it is also possible that it is not the number of predictors which is important here, but their quality. LiDAR is a good proxy for vegetation structure that may be more important than specific land uses for many bird species.

Remote sensing data can provide extremely useful information for ENMs (Bergen *et al.* 2009, Cord and Rödder 2011, Morán-Ordóñez *et al.* 2012, Sillero *et al.* 2012, Tattoni *et al.* 2012) that can be important at both the local and broad spatial scales. At the local scale, habitat features are often the major determinants of species distribution, and accurate measures of habitat structure can be extremely important. At broad scales, comprehensive habitat maps are rarely available and often have coarse resolution: remote sensing allows to obtain useful and consistent measures of habitat availability (Cord and Rödder 2011). The *a priori* identification of the appropriate number and identity of predictors can greatly improve ENMs (Peterson and Nakazawa 2008, Rödder and Lötters 2009). Our study suggests that few variables can be enough to build ENMs if presence records are limited. Remote sensing data provide a good measure of vegetation structure and may allow a better representation of the available habitat, therefore improving our ability to model and understand species distribution.

Aknowledgements

The comments of S. Laffan and thee anonymous reviewers improved an earlier version of the manuscript. We would also like to thank Fugro BV for the early provision of the original LiDAR data from AHN-2 for the Netherlands.

Funding

GFF was funded by a fellowship by University of Milano-Bicocca and AB was funded by the EU 7FP BIO_SOS project.

References

Alvarez, F. and Brito, J.C., 2006. Habitat requirements and potential areas of occurrence for the Pine Marten in North-western Portugal: conservation implications. *In*: M. Santos-Reis *et al.*, eds. *Martes in carnivore communities*. Alberta: Alpha Wildlife, 27–43.

Anadón, J.D., *et al.*, 2006. Factors determining the distribution of the spur-thighed tortoise *Testudo graeca* in south-east Spain: a hierarchical approach. *Ecography*, 29, 339–346. doi:10.1111/j.2006.0906-7590.04486.x

Atiénzar, F., *et al.*, 2009. Nesting habitat requirements and nestling diet in the Mediterranean populations of Crested Tits *Lophophanes cristatus*. *Acta Ornithologica*, 44, 101–108. doi:10.3161/000164509X482678

Bartel, R.A. and Sexton, J.O., 2009. Monitoring habitat dynamics for rare and endangered species using satellite images and niche-based models. *Ecography*, 32, 888–896. doi:10.1111/j.1600-0587.2009.05797.x

Bergen, K.M., *et al.*, 2009. Remote sensing of vegetation 3-D structure for biodiversity and habitat: review and implications for lidar and radar spaceborne missions. *Journal of Geophysical Research-Biogeosciences*, 114, 13. doi:10.1029/2008JG000883

Bibby, C.J., *et al.*, 2000. *Bird census techniques*. London: Academic Press.

BirdLife International, 2004. *Birds in the European Union: a status assessment*. Wageningen: Birdlife International.

Boulangeat, I., Gravel, D., and Thuiller, W., 2012. Accounting for dispersal and biotic interactions to disentangle the drivers of species distributions and their abundances. *Ecology Letters*, 15, 584–593. doi:10.1111/j.1461-0248.2012.01772.x

Bradbury, R.B., *et al.*, 2005. Modelling relationships between birds and vegetation structure using airborne LiDAR data: a review with case studies from agricultural and woodland environments. *Ibis*, 147, 443–452. doi:10.1111/j.1474-919x.2005.00438.x

Brambilla, M., *et al.*, 2010. Glorious past, uncertain present, bad future? Assessing effects of land-use changes on habitat suitability for a threatened farmland bird species. *Biological Conservation*, 143, 2770–2778. doi:10.1016/j.biocon.2010.07.025

Brambilla, M. and Ficetola, G.F., 2012. Species distribution models as a tool to estimate reproductive parameters: a case study with a passerine bird species. *Journal of Animal Ecology*, 81, 781–787. doi:10.1111/j.1365-2656.2012.01970.x

Bunce, R.G.H., *et al.*, 2013. The significance of habitats as indicators of biodiversity and their links to species. *Ecological Indicators*, 33 (special issue: Biodiversity Monitoring), 19–25. doi:10.1016/j.ecolind.2012.07.014

Burnham, K.P. and Anderson, D.R., 2002. *Model selection and multimodel inference: a practical information-theoretic approach*. New York: Springer Verlag.

Clawges, R., *et al.*, 2008. The use of airborne lidar to assess avian species diversity, density, and occurrence in a pine/aspen forest. *Remote Sensing of Environment*, 112, 2064–2073. doi:10.1016/j.rse.2007.08.023

Cord, A. and Rödder, D., 2011. Inclusion of habitat availability in species distribution models through multi-temporal remote-sensing data? *Ecological Applications*, 21, 3285–3298. doi:10.1890/11-0114.1

Elith, J., *et al.*, 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29, 129–151. doi:10.1111/j.2006.0906-7590.04596.x

Elith, J., *et al.*, 2011. A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, 17, 43–57. doi:10.1111/j.1472-4642.2010.00725.x

Elith, J. and Leathwick, J.R., 2009. Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology Evolution and Systematics*, 40, 677–697. doi:10.1146/annurev.ecolsys.110308.120159

Evans, J.S. and Hudak, A.T., 2007. A multiscale curvature algorithm for classifying discrete return LiDAR in forested environments. *IEEE Transactions on Geoscience and Remote Sensing*, 45, 1029–1038. doi:10.1109/TGRS.2006.890412

Ficetola, G.F., *et al.*, 2010. Knowing the past to predict the future: land-use change and the distribution of invasive bullfrogs. *Global Change Biology*, 16, 528–537. doi:10.1111/j.1365-2486.2009.01957.x

Ficetola, G.F., Thuiller, W., and Padoa-Schioppa, E., 2009. From introduction to the establishment of alien species: bioclimatic differences between presence and reproduction localities in the slider turtle. *Diversity and Distributions*, 15, 108–116. doi:10.1111/j.1472-4642.2008.00516.x

Gallien, L., *et al.*, 2012. Invasive species distribution models – how violating the equilibrium assumption can create new insights. *Global Ecology and Biogeography*, 21, 1126–1136. doi:10.1111/j.1466-8238.2012.00768.x

Goetz, S., *et al.*, 2007. Laser remote sensing of canopy habitat heterogeneity as a predictor of bird species richness in an eastern temperate forest, USA. *Remote Sensing of Environment*, 108, 254–263. doi:10.1016/j.rse.2006.11.016

Graf, R.F., Mathys, L., and Bollmann, K., 2009. Habitat assessment for forest dwelling species using LiDAR remote sensing: capercaillie in the Alps. *Forest Ecology and Management*, 257, 160–167. doi:10.1016/j.foreco.2008.08.021

Gregory, R.D., Gibbons, D.W., and Donald, P.F., 2004. Bird census and survey techniques. *In*: W.J. Sutherland, I. Newton, and R.E. Green, eds. *Bird ecology and conservation: a handbook of techniques*. Oxford: Oxford University Press.

Hazeu, G.W., *et al.*, 2011. A Dutch multi-date land use database: identification of real and methodological changes. *International Journal of Applied Earth Observation and Geoinformation*, 13, 682–689. doi:10.1016/j.jag.2011.04.004

Hijmans, R.J., *et al.*, 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25, 1965–1978. doi:10.1002/joc.1276

Hyde, P., *et al.*, 2005. Mapping forest structure for wildlife habitat analysis using waveform lidar: validation of montane ecosystems. *Remote Sensing of Environment*, 96, 427–437. doi: 10.1016/j.rse.2005.03.005

Hyde, P., *et al.*, 2006. Mapping forest structure for wildlife habitat analysis using multi-sensor (LiDAR, SAR/InSAR, ETM+, Quickbird) synergy. *Remote Sensing of Environment*, 102, 63–73. doi:10.1016/j.rse.2006.01.021

Korpela, I., *et al.*, 2009. Airborne small-footprint discrete-return LiDAR data in the assessment of boreal mire surface patterns, vegetation, and habitats. *Forest Ecology and Management*, 258, 1549–1566. doi:10.1016/j.foreco.2009.07.007

Kumstatova, T., *et al.*, 2004. Habitat preferences of tree pipit (*Anthus trivialis*) and meadow pipit (*A. pratensis*) at sympatric and allopatric localities. *Journal of Ornithology*, 145, 334–342. doi:10.1007/s10336-004-0048-3

Lefsky, M.A., *et al.*, 2002. Lidar remote sensing for ecosystem studies. *Bioscience*, 52, 19–30. doi:10.1641/0006-3568(2002)052[0019:LRSFES]2.0.CO;2

Legendre, P. and Legendre, L., 1998. *Numerical ecology*. Amsterdam: Elsevier.

Li, W., Guo, Q., and Elkan, C., 2011. Can we model the probability of presence of species without absence data? *Ecography*, 34, 1096–1105. doi:10.1111/j.1600-0587.2011.06888.x

Manel, S., Williams, H.C., and Ormerod, S.J., 2001. Evaluating presence–absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology*, 38, 291–931.

Morán-Ordóñez, A., *et al.*, 2012. Satellite surface reflectance improves habitat distribution mapping: a case study on heath and shrub formations in the Cantabrian Mountains (NW Spain). *Diversity and Distributions*, 18, 588–602. doi:10.1111/j.1472-4642.2011.00855.x

Mücher, C.A., *et al.*, 2013. Quantifying structure of Natura 2000 heathland habitats using spectral mixture analysis and segmentation techniques on hyperspectral imagery. *Ecological Indicators*, 33 (special issue: Biodiversity Monitoring), 71–81. doi:10.1016/j.ecolind.2012.09.013

Murtaugh, P.A., 2009. Performance of several variable-selection methods applied to real ecological data. *Ecology Letters*, 12, 1061–1068. doi:10.1111/j.1461-0248.2009.01361.x

Nogués-Bravo, D., 2009. Predicting the past distribution of species climatic niches. *Global Ecology and Biogeography*, 18, 521–531. doi:10.1111/j.1466-8238.2009.00476.x

Pearson, R.G., *et al.*, 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. *Journal of Biogeography*, 34, 102–117. doi:10.1111/j.1365-2699.2006.01594.x

Peterson, A.T. and Nakazawa, Y., 2008. Environmental data sets matter in ecological niche modelling: an example with *Solenopsis invicta* and *Solenopsis richteri*. *Global Ecology and Biogeography*, 17, 135–144.

Phillips, S.J., Anderson, R.P., and Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190, 231–259. doi:10.1016/j.ecolmodel.2005.03.026

Phillips, S.J. and Dudík, M., 2008. Modeling of species distributions with MaxEnt: new extensions and a comprehensive evaluation. *Ecography*, 31, 161–175. doi:10.1111/j.0906-7590.2008.5203.x

Phillips, S.J. and Elith, J., 2013. On estimating probability of presence from use-availability or presence-background data. *Ecology*, 94, 1409–1419. doi:10.1890/12-1520.1

Rangel, T.F.L.V.B., Diniz-Filho, J.A.F., and Bini, L.M., 2010. SAM: a comprehensive application for spatial analysis in macroecology. *Ecography*, 33, 46–50. doi:10.1111/j.1600-0587.2009.06299.x

Rödder, D. and Lötters, S., 2009. Niche shift versus niche conservatism? Climatic characteristics of the native and invasive ranges of the Mediterranean house gecko (*Hemidactylus turcicus*). *Global Ecology and Biogeography*, 18, 674–687. doi:10.1111/j.1466-8238.2009.00477.x

Rödder, D., *et al.*, 2009. Alien invasive slider turtle in unpredicted habitat: a matter of niche shift or of predictors studied? *PLoS ONE*, 4, e7843. doi:10.1371/journal.pone.0007843

Royle, J.A., *et al.*, 2012. Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. *Methods in Ecology and Evolution*, 3, 545–554. doi:10.1111/j.2041-210X.2011.00182.x

Rushton, S.P., Ormerod, S.J., and Kerby, G., 2004. New paradigms for modelling species distributions? *Journal of Applied Ecology*, 41, 193–200. doi:10.1111/j.0021-8901.2004.00903.x

Seoane, J., Bustamante, J., and Díaz-Delgado, R., 2004a. Are existing vegetation maps adequate to predict bird distributions? *Ecological Modelling*, 175, 137–149. doi:10.1016/j.ecolmodel.2003.10.011

Seoane, J., Bustamante, J., and Díaz-Delgado, R., 2004b. Competing roles for landscape, vegetation, topography and climate in predictive models of bird distribution. *Ecological Modelling*, 171, 209–222. doi:10.1016/j.ecolmodel.2003.08.006

Sillero, N., 2011. What does ecological modelling model? A proposed classification of ecological niche models based on their underlying methods. *Ecological Modelling*, 222, 1343–1346. doi:10.1016/j.ecolmodel.2011.01.018

Sillero, N., *et al.*, 2012. The significance of using satellite imagery data only in Ecological Niche Modelling. *Acta Herpetologica*, 7, 221–237.

Smith, A.B., 2013. On evaluating species distribution models with random background sites in place of absences when test presences disproportionately sample suitable habitat. *Diversity and Distributions*, 19, 867–872. doi:10.1111/ddi.12031

Soberon, J. and Nakamura, M., 2009. Niches and distributional areas: concepts, methods, and assumptions. *Proceedings of the National Academy of Sciences of the United States of America*, 106, 19644–19650. doi:10.1073/pnas.0901637106

Synes, N.W. and Osborne, P.E., 2011. Choice of predictor variables as a source of uncertainty in continental-scale species distribution modelling under climate change. *Global Ecology and Biogeography*, 20, 904–914. doi:10.1111/j.1466-8238.2010.00635.x

Tattoni, C., Rizzolli, F., and Pedrini, P., 2012. Can LiDAR data improve bird habitat suitability models? *Ecological Modelling*, 245, 103–110. doi:10.1016/j.ecolmodel.2012.03.020

Veloz, S.D., 2009. Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. *Journal of Biogeography*, 36, 2290–2299. doi:10.1111/j.1365-2699.2009.02174.x

Vierling, K.T., *et al.*, 2008. Lidar: shedding new light on habitat characterization and modeling. *Frontiers in Ecology and the Environment*, 6, 90–98. doi:10.1890/070001

Warren, D.L., Glor, R.E., and Turelli, M., 2010. ENMTools: a toolbox for comparative studies of environmental niche models. *Ecography*, 33 (3), 607–611.

Warren, D.L. and Seifert, S.N., 2011. Ecological niche modeling in MaxEnt: the importance of model complexity and the performance of model selection criteria. *Ecological Applications*, 21, 335–342. doi:10.1890/10-1171.1

Williams, K.J., *et al.*, 2012. Which environmental variables should I use in my biodiversity model? *International Journal of Geographical Information Science*, 26 (11), 2009–2047. doi:10.1080/13658816.2012.698015

Wisz, M.S., *et al.*, 2008. Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, 14, 763–773. doi:10.1111/j.1472-4642.2008.00482.x

Zuur, A.F., Ieno, E.N., and Elphick, C.S., 2010. A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, 1, 3–14. doi:10.1111/j.2041-210X.2009.00001.x