Ecological Modelling 312 (2015) 166–174

Contents lists available at ScienceDirect

Ecological Modelling

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

A framework for species distribution modelling with improved pseudo-absence generation

Maialen Iturbide[a,*,1], Joaquín Bedia[a], Sixto Herrera[b,c], Oscar del Hierro[d], Miriam Pinto[d], Jose Manuel Gutiérrez[a]

[a] *Meteorology Group, Institute of Physics of Cantabria, Universidad de Cantabria-CSIC, 39005 Santander, Spain*
[b] *Meteorology Group, Department of Applied Mathematics and Computer Science, Universidad de Cantabria, 39005 Santander, Spain*
[c] *Predictia Intelligent Data Solutions, S.L. CDTUC Fase A, Planta 2–203, Avda. los Castros s/n, 39005 Santander, Spain*
[d] *NEIKER-Tecnalia, Basque Institute for Agricultural Research and Development, 48160 Derio, Spain*

ARTICLE INFO

Article history:
Received 3 January 2015
Received in revised form 29 April 2015
Accepted 17 May 2015
Available online 10 June 2015

Keywords:
Ecological niche
Quercus
Environmental profiling
Sampling methods
Threshold distance

ABSTRACT

Species distribution models (SDMs) are an important tool in biogeography and phylogeography studies, that most often require explicit absence information to adequately model the environmental space on which species can potentially inhabit. In the so-called *background pseudo-absences* approach, absence locations are simulated in order to obtain a complete sample of the environment. Whilst the commonest approach is random sampling of the entire study region, in its multiple variants, its performance may not be optimal, and the method of generation of pseudo-absences is known to have a significant influence on the results obtained. Here, we compare a suite of classic (random sampling) and novel methods for pseudo-absence data generation and propose a generalizable three-step method combining environmental profiling with a new technique for background extent restriction. To this aim, we consider 11 phylogenetic groups of Oak (*Quercus* sp.) described in Europe. We evaluate the influence of different pseudo-absence types on model performance (area under the ROC curve), calibration (reliability diagrams) and the resulting suitability maps, using a cross-validation approach. Regardless of the modelling algorithm used, random-sampling models were outperformed by the methods that incorporate environmental profiling of the background, stressing the importance of the pseudo-absence generation techniques for the development of accurate and reliable SDMs. We also provide an integrated modelling framework implementing the methods tested in a software package for the open source R environment.

1. Introduction

Species distribution models (SDMs) constitute rules that associate known presence locations of biological entities with the characteristics of their environment to predict its potential distribution in the geographic space (Guisan and Zimmermann, 2000; Elith et al., 2006). SDM building techniques can be broadly classified into two types: *profile* and *group discrimination* techniques. The first group refers to those modelling approaches that rely solely on known presences to infer the potential distribution of the species, while group discrimination techniques require information of the environmental range where the species do not occur, that is, absence data. Group discrimination techniques have gained popularity in recent years, as they have been reported to yield better results than profile techniques (Engler et al., 2004; Chefaoui and Lobo, 2008; Elith et al., 2006; Mateo et al., 2010). However, in part due to the great effort involved in true absence sampling, most of the available datasets for predictive modelling (generally natural history collections, see e.g. Araújo and Williams, 2000) are lacking explicit absence data. Thus, in most cases discrimination techniques are used, requiring the environmental characterization of the sites of presence in front of a background sample (pseudo-absence data) that characterizes the available environment in the study region.

Although the strong influence of the pseudo-absence generation process has been shown in previous studies, comparative analyses addressing the suitability of different methods, some of them quite novel, are scarce in the literature (Zaniewski et al., 2002; Phillips et al., 2009; Lobo et al., 2010), and there is not a consensus on the way in which pseudo-absences should be generated. In fact, several previous studies addressing this issue (e.g. Hengl et al., 2009; Wisz

[*] Corresponding author. Tel.: +34 942 20 20 64.
E-mail address: miturbide@ifca.unican.es (M. Iturbide).
[1] Edificio Juan Jordá, Avda Los Castros s/n, 39005 Santander, Spain.

http://dx.doi.org/10.1016/j.ecolmodel.2015.05.018
0304-3800/© 2015 Elsevier B.V. All rights reserved.

and Guisan, 2009; Stokland et al., 2011; Senay et al., 2013) propose contradictory solutions. As such, the inclusion of reliable pseudo-absences in model evaluation remains an open issue.

The most simple and widely applied method of generating pseudo-absences is random selection of the entire study area (e.g. Gastón and García-Viñas, 2011; Hanspach et al., 2011; Domisch et al., 2013). A search in the SCOPUS database containing the terms "habitat suitability", "niche modelling" and "background data", "pseudo-absence" or "presence-only", narrowed to the journals of the first quartile and the topic "environmental sciences" for the period 2009–July 2014, yielded a total of 64 articles from which roughly 80% used presence-only datasets. Of them, the 92% used randomly generated pseudo-absences within the study area, either explicitly (38%), or implicitly (54%) via the MAXENT algorithm (see e.g. Barbet-Massin et al., 2012; Jiménez-Valverde, 2012, for details), other 28% used profile techniques and a 12% used target group background (note that some of the articles analysed used more than one type of technique, and therefore percentages do not sum up to 100%). Percentages under 10% correspond to the novel approaches analysed in this article. In spite of its wide application, the random sampling method rises the risk of introducing false absences into the model from locations that are suitable for the species, leading to underestimates of its fundamental niche and potential distribution (Anderson and Raza, 2010). This occurs naturally due to biotic interactions and dispersal limitations that do not allow the species to inhabit, and also very often as a result of sampling biases in the data collections. Faced with this problem, it is common practice to set a buffer distance from known presence localities in order to minimize the false negative rate (e.g. Mateo et al., 2010; Bedia et al., 2013). More elaborated approaches employ a presence-only algorithm as a preliminary step to move pseudo-absences away in the environmental space (see e.g. Zaniewski et al., 2002; Engler et al., 2004; Barbet-Massin et al., 2012; Liu et al., 2013) or apply a geographically weighted exclusion, which keeps pseudo-absences out from presences using distance maps (Hirzel et al., 2001; Barbet-Massin et al., 2012; Norris et al., 2011; Hengl et al., 2009). These strategies are intended to reduce the background data to those areas where false absences are less likely to occur, while the target group background method has been posited as a solution to remove some of the bias in presence-data collections, using the presence localities of other species as biased background data (Phillips et al., 2009).

Another critical matter regarding pseudo-absence data is the extent from which background is sampled. In fact, the available data in the background are usually much larger than the data characterized by presence localities (Anderson and Raza, 2010). A constrained distribution of pseudo-absences around presence locations can lead to misleading models, while unconstrained sampling can artificially inflate test statistics, as well as the weight of less informative response variables (Van der Wal and Shoo, 2009). As a result, the three-step method has been recently proposed as an adequate approach to overcome these limitations, envisaged to define the extent and the environmental range of the background from which pseudo-absences are sampled (Senay et al., 2013, see Section 2.4 for details). From an ecological perspective, the uncertainty associated to the presence of a biological entity is a combined effect of separate factors (biotic, abiotic and movement factors), that in turn depend on the environment of a specific site. In this context, the three-step method pursues the estimation of the fundamental distribution (regions of favourable abiotic factors) by the introduction of pseudo-absences within the niche space corresponding to areas of non-presence (outside the realized niche) and where movement factors are likely favourable (accessible geographic areas) but not so the abiotic factors (Peterson et al., 2011). On the opposite, random sampling would produce predictions closer to a realized distribution, since it only excludes the presence locations for pseudo-absence data generation.

The aims of this study are: (i) to analyse the effect of the method used for pseudo-absence data generation on resulting SDMs, and (ii) to provide a modelling framework implementing the state-of-the-art techniques yielding optimal results. In particular, we compare five pseudo-absence data generation methods, ranging from the classical random sampling of the whole region and the target group method, to more sophisticated three-step techniques, combining environmental profiling and spatial restrictions on the sampling domain. We also propose a new criterion for background extent selection based on the theoretical properties of model performance as a function of distance to presence locations. We consider three modelling techniques commonly used in SDM applications and 11 phylogenetic groups of *Quercus* sp. identified in Europe (*Quercus* sp Europe database, Petit et al., 2002b). In addition, we provide an integrated modelling framework based on the open-source R language (R Core Team, 2014), implementing the methods tested in this study (Supplementary Material).

2. Methods and materials

2.1. Species data

The term "species" is a taxonomic designation, and may not necessarily refer to an ecologically homogeneous group of organisms when different ecotypes occur within the study area (Oney et al., 2013). Experimental evidence suggests that conventional SDM is not able to properly capture the climatic response of species by treating them as homogeneous units (Beierkuhnlein et al., 2011). With this regard, Hernández et al. (2006) suggested that research in environmental niche modelling should focus on broad distributional subunits based on distinct genetic linages. For instance, Gonzàlez et al. (2011) demonstrated that omission error is reduced when "biologically meaningful" data (in reference to genetically distinct populations of the same species) are modelled. Hence, in this study we consider genetically differenced groups of *Quercus* sp. in Europe. Each group corresponds to a different chloroplast haplotype, determined by PCR analysis on more than 2600 populations of Oaks in Europe (see Petit et al., 2002a,b,c). We considered 11 out of the total 42 Oak haplotypes identified, attending to the minimum population size needed to build the models ($n > 30$) while attending to the best possible representation of all European *Quercus* linages (Petit et al., 2002b, Table 1).

The study area was divided in 11 parts (in correspondence to each haplotype distribution) by defining a bounding box around the presence points (Fig. 1).

Table 1
Haplotypes considered ordered by decreasing sample size (n), and the lineages they belong to, according to the *Quercus* sp Europe database (Petit et al., 2002b). Only one linage (F) out of five was not included in the analyses due to insufficient sample size of all its haplotypes.

Haplotype	Linage	n
H7	A	734
H10	B	651
H1	C	490
H12	B	466
H11	B	283
H5	A	250
H17	E	67
H4	A	53
H6	A	41
H15	E	36
H27	D	31

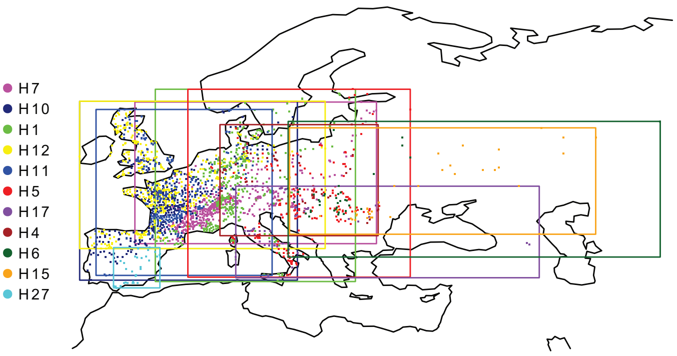

Fig. 1. Phylogenetic distribution of *Quercus* sp in Europe. Oak groups in decreasing sample size order are: H7 (*n* = 734), H10 (*n* = 651), H1 (*n* = 490), H12 (*n* = 466), H11 (*n* = 283), H5 (*n* = 250), H17 (*n* = 67), H4 (*n* = 53), H6 (*n* = 41), H15 (*n* = 36) and H27 (*n* = 31).

2.2. Climate data

We used the bioclimatic variables of the WorldClim dataset (Hijmans et al., 2005) at 10 km resolution as explanatory variables to build the SDMs. The chosen resolution is adequate to the aims of this study, given the "false precision" provided by the downscaled WorldClim climate surfaces of 1 km, as highlighted in previous niche modelling studies (Bedia et al., 2013). After a pairwise cross-correlation analysis of the bioclimatic variables (following Bedia et al., 2013), we retained a subset of uncorrelated predictors (bio02, bio03, bio08, bio13, bio14 and bio15) rescaled in the range [0, 1].

2.3. SDM development and assessment

SDMs were built using three different popular techniques, namely maximum entropy (MAXENT, Phillips et al., 2006), generalized linear models (GLMs, Guisan and Zimmermann, 2000) and multivariate adaptive regression splines (MARS, Friedman, 1991). Constrained by data availability, we resorted to cross-validation techniques (Steyerberg et al., 2010) to replace truly independent data for model validation, as it is commonplace in ecological studies (e.g. Manel et al., 1999). In particular, we used a 10-fold cross validation approach, given that it is equally efficient in the error estimation as other techniques computationally more demanding like for instance leave-one-out cross validation (Kohavi, 1995).

We used the area under the ROC curve (AUC) as the most widely used metric for model performance assessment. The ROC curve describes the predictive ability of the system under the whole range of probability thresholds, thus representing a global measure of model performance, that is quantitatively assessed by the area it encloses. Thus, high AUC values (closer to 1) indicate good model discrimination, although this is not necessarily coupled to a high numerical accuracy of the predictions (Bedia et al., 2011). With this regard, *calibration plots* (also known as *reliability diagrams*) can be used in order to provide additional information regarding the level of agreement between predicted and observed probabilities of occurrence. This information is displayed in the form of a plot such that the better the agreement, the closer the line is to the diagonal for the whole range of probability values (see e.g. Bedia et al., 2011; Vaughan and Ormerod, 2005, for a wider explanation in the context of SDM assessment).

2.4. Pseudo-absence data generation

A larger proportion of pseudo-absences against presences can affect model performance positively or negatively, introducing biases in model inter-comparisons, for which prevalence should be kept constant at an intermediate level (McPherson et al., 2004; Liu et al., 2005). Thus, for all methods tested we kept the number of pseudo-absences equal to the number of presences in all cases (prevalence = 0.5, Hengl et al., 2009; Mateo et al., 2010; Hanspach et al., 2011; Senay et al., 2013). Additionally, a exclusion buffer of 10 km around the occurrence points was set in order to avoid cells containing both presence and pseudo-absence data (Chefaoui and Lobo, 2008). All steps involved in pseudo-absence generation according to the different methods tested are indicated in the diagram of Fig. 2.

Random selection (RS). Pseudo-absences were sampled at random in the whole background, excepting the grid points within the exclusion buffer.

Random selection with environmental profiling (RSEP). The RSEP method is aimed at defining the environmental range of the background from which pseudo-absences are sampled. Environmentally unsuitable areas are defined using a presence-only profiling algorithm. To this aim, we run one-class support vector machines (OCSVM, Scholkopf and Smola, 2001) for each Oak group (see e.g. Drake et al., 2006; Bedia et al., 2011, for specific details on the use of support vector machines in SDM studies). OCSVM has been indicated as the most adequate algorithm for this purpose as it can handle high dimensional data and complex non-linear relationships between predictors (Senay et al., 2013).

Three-step selection (TS). The TS method adds two more steps to the RSEP method to define the environmental range, and also the extent of the background from which pseudo-absences are sampled (Fig. 2). Thus, the first step is the definition of the environmentally unsuitable areas as is done in the RSEP method.

In the second step, alternative SDMs are built using random pseudo-absences generated for different spatial extents within the unsuitability background zones defined in the first step. In order to consider all possible extents, we set different maximum *distance thresholds* to each presence location, considering a sequence from 20 km (twice the exclusion buffer) to the length of half diagonal of the bounding box (the maximum possible distance between any pair of points within the area (Fig. 1)), each 10 km (the grid resolution).

The third step consists in selecting the optimum background extent and the corresponding fitted model from all possible pseudo-absence configurations generated in Step 2. Senay et al. (2013) limited the background data using a variable importance change criterion based on principal component analysis to reduce the dimensionality of the environmental space. In our case, we applied a model performance criterion, as variable importance may not always vary significantly for the whole range of distances tested. Thus, a threshold extent is chosen according to the best model performance, while minimizing the distance to presences. With this regard, Van der Wal and Shoo (2009) evaluated the relationship between the geographic extent from which pseudo-absences are taken and model performance, and found that AUC rapidly increased as background size expanded from 10 to 100 km while subsequent expansions resulted in only minor increases in AUC. We found a similar behaviour for all Oak groups, and concluded that the AUC *vs.* distance curve can be optimally fit to an asymptotic Michaelis–Menten type model of the form:

$$v(x) = \frac{Vm \times x}{Km + x}, \qquad (1)$$

where v and x represent the AUC and the background extent respectively, Vm (Fig. 3) is the asymptotic AUC value achieved by the system and the Michaelis constant Km is the extent at which the AUC is half of Vm. As a result, we propose a generalizable method to find the threshold extent for pseudo-absence sampling near the suitability boundary of the species, without penalizing model performance, which constitutes the major novelty in comparison with previous published methodologies. Thus, AUCs from

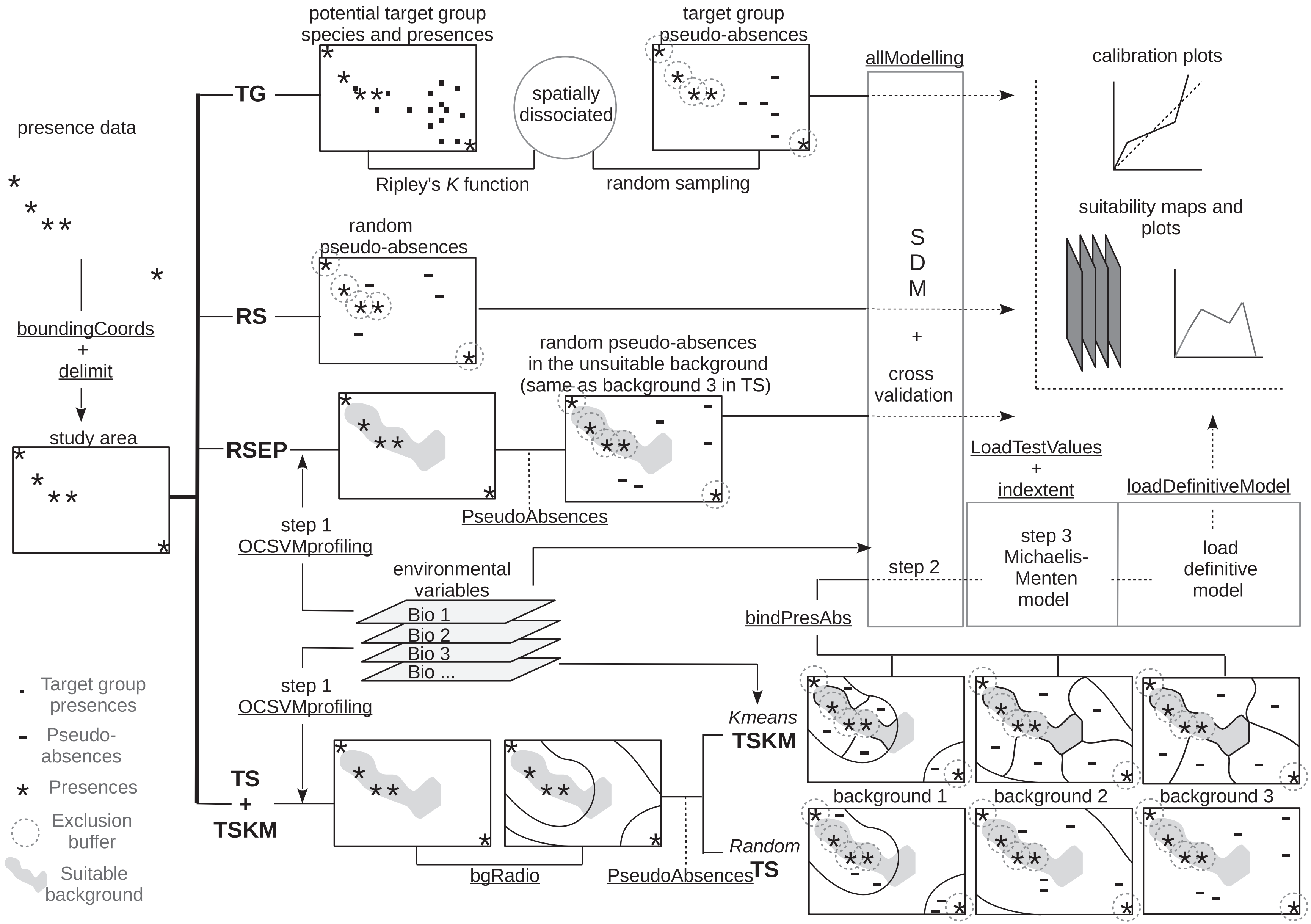

Fig. 2. Conceptual diagram of the methodology used in this study. Legend is shown in the bottom left corner. Underlined words refer to the names of the R functions performing each step in the developed framework (see Supplementary Material).

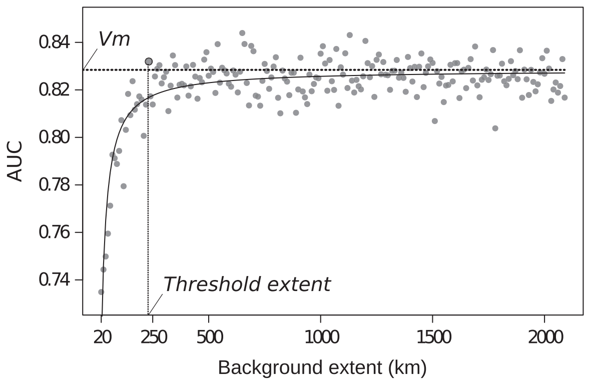

Fig. 3. Relation of the AUC to the background extent for group H7. The black curve corresponds to the fitted Michaelis–Menten model. V_m represents the maximum AUC achieved by the system. The highlighted point corresponds to the smallest background extent greater than V_m (i.e., the threshold extent). This relationship is similar to that described in Figure 2 in Van der Wal and Shoo (2009). All Oak groups in the study exhibited the same type of curve (see also the examples in the Supplementary Material).

Table 2
Threshold distances to presences (kilometres) defining the background extents from which pseudo-absences are sampled. Each data in the column d_{max} correspond to the length of the half diagonal of the bounding box that encloses the study area (Fig. 1), i.e.: the maximum possible distance between a pair of points within the study area.

	d_{TS}	d_{TSKM}	d_{max}
H7	230	290	2090
H10	500	670	2100
H1	580	800	2070
H12	620	620	2130
H11	390	560	1800
H5	190	240	2170
H17	690	830	2360
H4	150	380	1440
H6	1000	1050	2950
H15	360	80	2420
H27	30	70	450

the multimodel and the different background extents tested are fitted to the curve of Eq. (1) to extract the theoretical asymptotic AUC value (*Vm*). Then, the minimum threshold extent *x* at which $AUC_x > Vm$ is chosen (Fig. 3), and the corresponding fitted SDM is retained to produce the suitability maps for the entire study area.

Three-step with k-means selection (TSKM). The difference of TSKM with regard to TS is that the pseudo-absences are taken from the spatial subunits defined by a clustering on the background extent in Step 2. Instead of using a random selection on the unsuitable areas after Step 1, a *k*-means clustering is applied on the environmental and geographical space (*k* being equal to the number of presence points) and the coordinate values of each cluster centroid are retained, thus obtaining a regular distribution of dissimilar points for the study area which constitutes a representative sample of the unsuitable environment (Senay et al., 2013). Step 3 is then done as in TS method. The resulting background extents for the TS and TSKM methods are listed in Table 2.

Target group selection (TG). In order to select a target group for each phylogenetic Oak group we searched for presence records of species not belonging to the *Fagaceae* family in the database of The Global Biodiversity Information Facility (GBIF, http://data.gbif.org). To ensure a sufficiently high number of presence points, we focused on species with a widespread distribution in Europe as target group candidates.

For each candidate and Oak group, we computed the cross type of the Ripley's *K* function (Dixon, 2006) to analyse the spatial behaviour of the point pattern. From the estimated Cross *K*-functions, those showing spatial dissociation of the TG candidate with regard to the Oak group were chosen (see Grantham, 2012, for wider explanation regarding point pattern analysis and Rypley's *K* function interpretation), resulting in the following target groups: *Ulex europaeus* for groups H3 and H11; *Picea glauca* for groups H1, H2, H4, H5, H6 and H8; *Pinus nigra* for groups H7 and H10; and *Pinus strobus* for group H9. TG locations were then randomly sampled to match the number of Oak localities in order to obtain balanced datasets for model training (see Section 2.4).

Table 3
Multimodel mean AUC values, according to the four pseudo-absence generation methods tested, for each of the Oak groups analysed. Values for TG method are underlined when they are the best of all methods. Values in bold are the maximum AUC values excluding the TG method.

	RS	RSEP	TS	TSKM	TG
H7	0.771	**0.834**	0.832	0.830	0.981
H10	0.772	0.854	0.851	**0.856**	0.970
H1	0.764	0.822	**0.823**	0.820	0.976
H12	0.781	0.839	**0.864**	0.852	0.971
H11	0.760	0.815	0.842	**0.846**	0.985
H5	0.786	**0.830**	0.829	0.828	0.977
H17	0.798	0.847	0.878	**0.897**	0.935
H4	0.720	**0.873**	0.835	0.824	0.962
H6	0.802	0.847	**0.862**	0.859	0.939
H15	**0.762**	0.668	0.748	0.707	0.941
H27	0.726	**0.843**	0.741	0.677	0.712

3. Results and discussion

3.1. TG method

TG attained the highest AUCs for almost all the phylogenetic groups (Table 3; Fig. 4), but in turn it yielded poorly calibrated models (Fig. 5), with a strong under-estimation of high probability values. We argue that these results are due to the spatially clustered distribution of targeted group presences used as pseudo-absences, leading to spatially autocorrelated background samples resulting in inflated AUC values (Gonzàlez et al., 2011), and also to an over-estimated suitability for a large proportion of non-sampled areas (Figs. 6 and 7), as compared to the other methods. Phillips et al. (2009) and Mateo et al. (2010) recommended the TG pseudo-absence as the best method for discrimination, resulting in models with the best predictive performance. We find the same result, with TG attaining the highest AUC values, although this comes at the cost of a poor model calibration, and therefore we do not recommend this technique if reliable suitability maps are to be obtained. This stresses the importance of well-distributed presence/absence data across the environmental and geographical space of the study area in order to obtain reliable models (Lobo and Tognelli, 2011).

3.2. RSEP, TS and TSKM methods

RSEP and three-step methods (TS and TSKM) attained similar results. As expected, we did not find any significant differences in their AUCs (Fig. 4; Table 3) since both TS and TSKM define a threshold extent based on the asymptotic AUC value *Vm* (Fig. 3), close to the expected value of the maximum distance threshold used by the RSEP method. With this regard, TS and TSKM methods are preferable than RSEP, since using the theoretical AUC value given by *Vm* ensures the selection of a good model, while RSEP method may result in a sub-optimal model if the last point in the *X*-axis lies significantly below the *Vm* value by chance (Fig. 3).

The suitability plots (Fig. 7) show a similar behaviour, clearly different from RS and TG. Thus, we conclude that the relevant step that affects SDM results is the environmental profiling of the background, which constitutes the common characteristic of

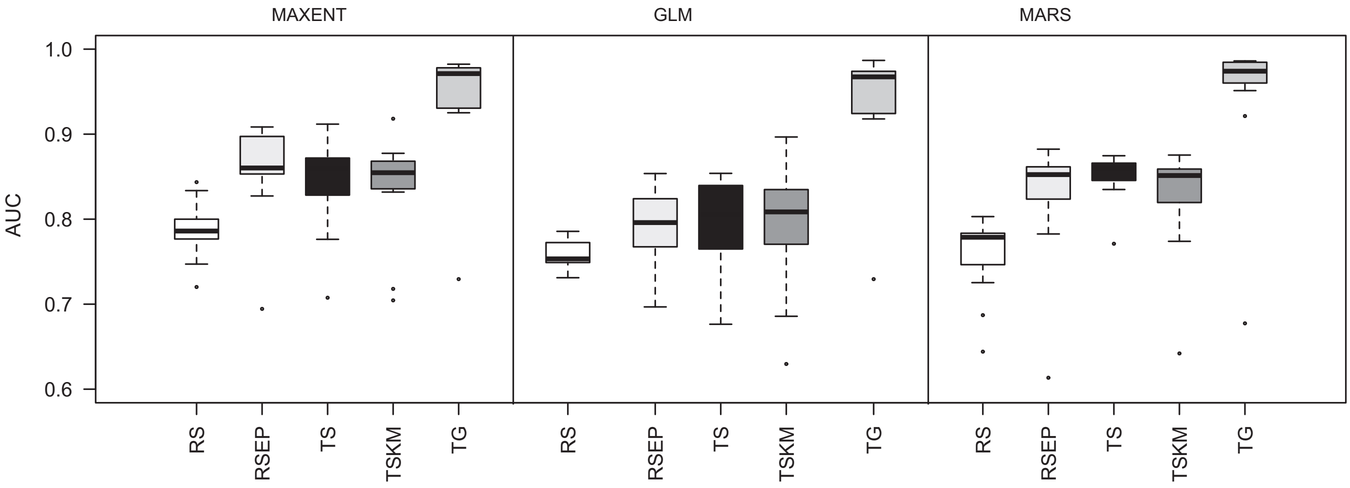

Fig. 4. AUC box-plots of the 11 oak groups modelled with the five pseudo-absence generation methods for each modelling technique. Oak groups were modelled with higher accuracy by MAXENT and MARS. The average AUC values improved for all modelling techniques when using a different method from RS.

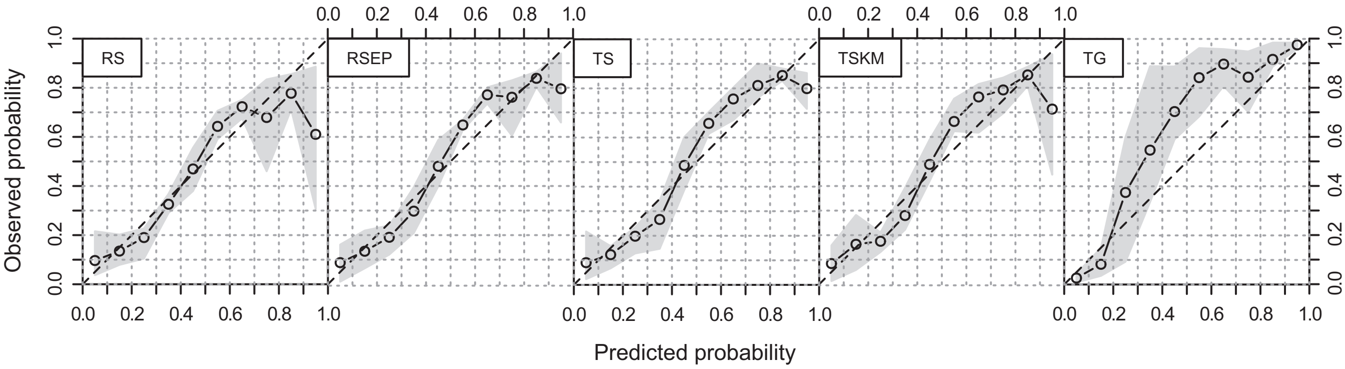

Fig. 5. Calibration plots of the multimodel predictions. Points connected by lines are the mean obtained from the different Oak groups and the grey area corresponds to the range between maximum and minimum values. Values below the diagonal indicate over-estimated probabilities and values above it under-estimated predictions. The smallest Oak groups H4 ($n = 53$), H6 ($n = 41$), H15 ($n = 36$) and H27 ($n = 31$), are excluded in the calibration plots, because their low sample size systematically yields poorly calibrated models that mask observable differences between methods.

the RSEP and three-step methods. As a result, RSEP was equally effective while entailing a more straightforward implementation. Analogously, since the background extent restriction does not impair final results, three-step methods are also recommendable as the effect of non-informative pseudo-absences from far regions could be significant in other case studies, especially when a wider study area is considered. In this sense, several authors argue that pseudo-absences from far regions should be avoided (Van der Wal and Shoo, 2009; Anderson and Raza, 2010). Moreover, Jiménez-Valverde et al. (2012) and Lobo et al. (2010) suggested that pseudo-absences should be located near the external boundary of the suitable environment to adequately represent the potential distribution of a species. At this respect, we consider that the three-step method proposed in this study satisfies this requirement while avoids misleading models with reduced AUCs. In addition, TS is generalizable and its implementation is straightforward using the R functions provided (Supplementary Material). Finally, since the TSKM method does not improve SDM results in relation to TS, the introduction of the *k*-means clustering in Step 2 of TSKM can be skipped in favour of a simple random selection within the background extent.

3.3. *RS method* vs. *RSEP, TS and TSKM methods*

The RS method produced well calibrated SDMs, excepting in the zones of higher environmental suitability, where the latter was over-estimated for all Oak groups (Fig. 5). This is due to the fact that many pseudo-absences are distributed around presences inside the potentially suitable environment, resulting in a lower rate of observed presences against absences in the zones predicted as most suitable, and is arguably one major disadvantage of the RS method with regard to methods applying environmental profiling as a previous step (RSEP, TS and TSKM). Furthermore, RS yielded the worst discrimination results, with the lowest AUC values for all algorithms tested (Fig. 4) and for most Oak groups (Table 3).

The use of a profiling technique as an intermediate step, characteristic of the three-step methods (TS and TSKM), has been criticized by some authors for producing artificially high probabilities of occurrence (Wisz and Guisan, 2009; Stokland et al., 2011) and wider predicted suitability areas. In ecological terms, the variability in the predicted probabilities is related to the ability of the SDMs to represent realized *vs.* potential species distributions, lying spatially wider predicted distributions closer to the fundamental niche of the target species (Chefaoui and Lobo, 2008). However, since the potential distribution of the species is uncertain, we see no reason to penalize the model based on the extent of the area predicted as suitable (see e.g. Jiménez-Valverde, 2012). Furthermore, our results indicate that the predicted potential areas are not significantly shrink/widened with the use of either profiling/RS techniques (they are though in case of TG method, Fig. 6). In fact, the most remarkable difference between both is a higher resolution of the profiling-based models as compared to RS for most Oak

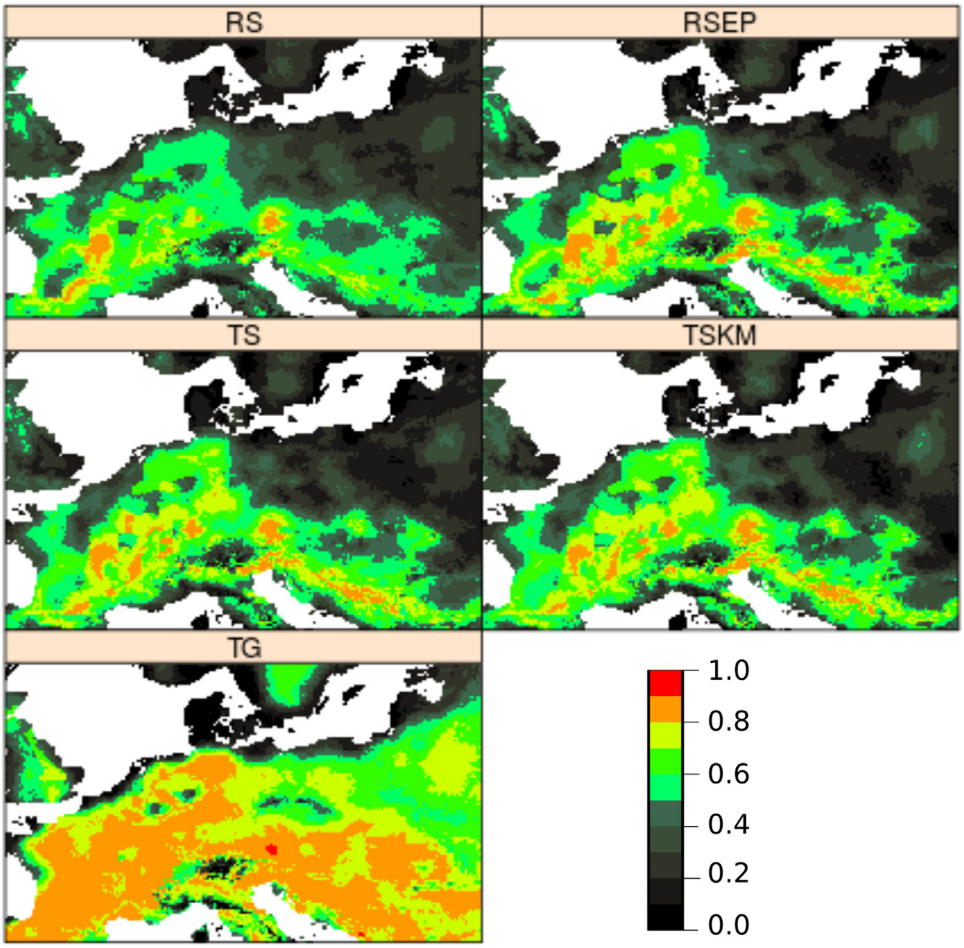

Fig. 6. Multimodel suitability maps according to the five pseudo-absence generation methods tested for Oak group H7. Maps for the rest Oak groups show the same pattern on the prediction change between methods as is shown in Fig. 7. Suitability is here expressed as a probability of occurrence given the environmental conditions, in the range [0, 1].

groups, as depicted by the suitability plots (Fig. 7). This means that ambiguous probabilities (around 0.5) are less likely to occur when RSEP or three-step methods are introduced, in favour of more informative predicted probabilities closer either to 1 or to 0, as opposed to the traditional RS approach. (see e.g. Bedia et al., 2011), for a more detailed explanation of model resolution in the context of SDMs). This is particularly important in order to reduce uncertainties when binary presence/absence maps are required for decision making and/or management plans.

Furthermore, the lack of records from suitable regions may simply derive from an inadequate sampling (Anderson, 2003; Hanspach et al., 2011). In fact, presence data are quite often environmentally biased (Bierman et al., 2010) resulting in presence data that do not represent the whole environmental range of the realized niche. In these cases, the RS method introduces false absences (within both the realized and fundamental niches) introducing a major source of uncertainty (Lobo et al., 2010) and resulting in over-constrained areas of high suitability (Fig. 7). In this sense, as long as RSEP, TS and TSKM methods sample pseudo-absences within a previously profiled unsuitable area, the risk of introducing false pseudo-absences is minimized, even in the case of relatively biased species collections. On the other hand, in case of error in the initial presence data (e.g. false positives), then profiling techniques may bear the risk of further reinforcing this bias rather than correcting it, although this particular situation should be further investigated.

3.4. Sensitivity of model performance to the pseudo-absence generation method

Our results show that the method of pseudo-absence generation strongly conditions output SDMs. Whilst the choice of the SDM algorithm is generally recognized as the principal factor of uncertainty in niche modelling studies (see e.g. Buisson et al., 2010; Fronzek et al., 2011), in this case study we demonstrate that pseudo-absence sampling design is even more important, leading to a larger variation of model AUC (Fig. 4; Table 3) than the modelling algorithms tested or the initial presence dataset choice, even though MAXENT and MARS performed better than GLMs (Fig. 4), indicating that algorithm selection is also an important factor (Phillips et al., 2009; Bedia et al., 2011; Senay et al., 2013). Our results also suggest that MARS performance was more sensitive to the pseudo-absence configuration than MAXENT (Fig. 4), although a more intensive testing beyond the scope of this study would be required to ascertain the sensitivity of different algorithms to the pseudo-absence generation scheme.

3.5. Sample size effect on results

As sample sizes are heterogeneous across Oak groups, this allowed us to indirectly evaluate the influence of the sample size in the performance. Caution has to be given to interpreting inflated AUC values due to small number of records (Wisz et al., 2008). For

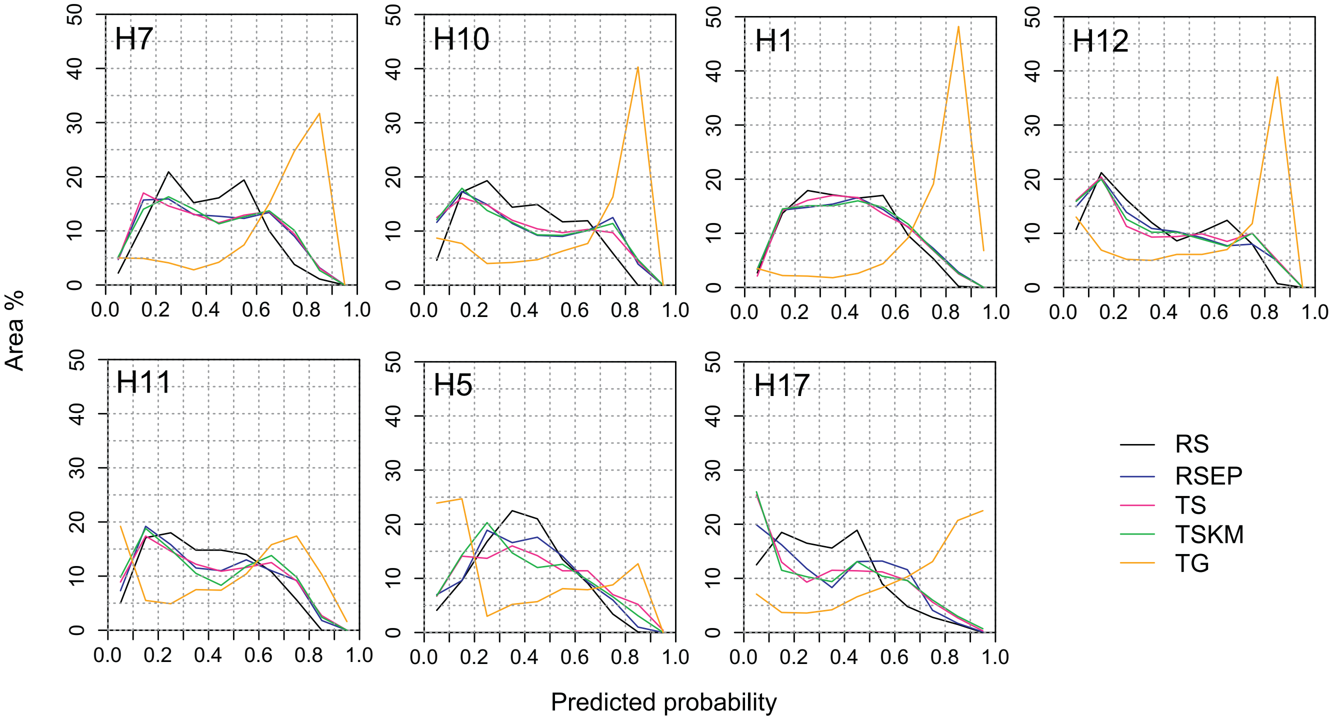

Fig. 7. Suitability plots. Percentage of area predicted into each interval of probability of occurrence for the Oak groups producing well calibrated models (see Fig. 5). These graphics give quantitative information on the suitability maps for a better interpretation of the results obtained. The first plot (H7) corresponds to the suitability maps shown in Fig. 6. Compared to RS, the RSEP, TS and TSKM methods produce incremented areas of high and low suitability and reduced mid suitable areas. The TG method predicts large areas of high suitability.

instance, Hanspach et al. (2011) excluded species with less than 50 records to allow reliable modelling. In this study, the calibration analysis shows that group H4 (53 presence records) and smaller groups (Table 1), did not produce reliable models for any of the pseudo-absence generation methods compared (not shown), even though AUC values were generally high (Table 3). In addition, the poor performance of the models for the smallest Oak groups (H15 and H27) is also reflected in the relationship of AUC and background extent, resulting in poor model fits in the TS and TSKM methods (Eq. (1)) and yielding small threshold extents and lower AUCs (Tables 2 and 3).

4. Conclusion

The method for pseudo-absence generation strongly affected output SDM performance regardless of the modelling algorithm chosen and for all the Oak groups tested. The classical random sampling method (RS) yielded the lowest overall performance, while the target group (TG) approach attained high AUC values at the cost of poorly calibrated models, resulting in unreliable suitability maps. Methods that include environmental profiling in a previous step (RSEP, TS and TSKM), clearly outperformed both RS and TG, yielding high AUC values and better calibrated predictions, resulting in the most reliable suitability maps with a higher resolution of the predicted probabilities. Thus, we suggest that further investigation on pseudo-absence data generation should focus in background data profiling. We recommend TS as the most adequate method, and also RSEP as a computationally simpler alternative. We also propose the AUC-driven method based on asymptotic curve fitting as an easily implementable and generalizable approach to obtain a suitable background extent threshold. RSEP, TS and TSKM methods are implemented in the open source R package mopa (*MOdelling Pseudo Absences*, https://github.com/SantanderMetGroup/mopa), described with worked examples in the Supplementary Material.

Acknowledgments

We are grateful to Rémy Petit and François Ehrenmann for providing the phylogenetic distribution of *Quercus*. We acknowledge the fruitful discussions arisen in the WG1 of the FPS COST Action FP1202 (MaP-FGR, “Strengthening conservation: a key issue for adaptation of marginal/peripheral populations of forest trees to climate change in Europe”). We also thank two anonymous referees for their thoughtful comments that greatly improved the manuscript. This work was supported by the EC-funded project ADAPTACLIMA II (INTERREG IVB SUDOE Program).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ecolmodel.2015.05.018

References

Anderson, R.P., 2003. Real vs. artefactual absences in species distributions: tests for oryzomys albigularis (rodentia: Muridae) in Venezuela. J. Biogeogr. 30, 591–605.

Anderson, R.P., Raza, A., 2010. The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: preliminary tests with montane rodents (genus Nephelomys) in Venezuela. J. Biogeogr. 37 (7), 1378–1393.

Araújo, M.B., Williams, P.H., 2000. Selecting areas for species persistence using occurrence data. Biol. Conserv. 96, 331–345.

Barbet-Massin, M., Jiguet, F., Albert, C.H., Thuiller, W., 2012. Selecting pseudo-absences for species distribution models: how, where and how many? Methods Ecol. Evol. 3 (2), 327–338.

Bedia, J., Busqué, J., Gutiérrez, J.M., 2011. Predicting plant species distribution across an alpine rangeland in Northern Spain: a comparison of probabilistic methods. Appl. Veg. Sci. 14, 415–432.

Bedia, J., Herrera, S., Gutiérrez, J.M., 2013. Dangers of using global bioclimatic datasets for ecological niche modeling. limitations for future climate projections. Glob. Planet. Change, 107.

Beierkuhnlein, C., Thiel, D., Jentsch, A., Willner, E., Kreyling, J., 2011. Ecotypes of European grass species respond differently to warming and extreme drought. J. Ecol. 99, 703–713.

Bierman, S.M., Butler, A., Marion, G., Kuehn, I., 2010. Bayesian image restoration models for combining expert knowledge on recording activity with species distribution data. Ecography 33 (3), 451–460.

Buisson, L., Thuiller, W., Casajus, N., Lek, S., Grenouillet, G., 2010. Uncertainty in ensemble forecasting of species distribution. Glob. Change Biol. 16, 1145–1157.

Chefaoui, R.M., Lobo, J.M., 2008. Assessing the effects of pseudo-absences on predictive distribution model performance. Ecol. Model. 210 (4), 478–486.

Dixon, P.M.,2006. Ripley's *k* function. In: Encyclopedia of Environmetrics, Ltd. John Wiley & Sons.

Domisch, S., Kuemmerlen, M., Jähnig, S., Haase, P., 2013. Choice of study area and predictors affect habitat suitability projections, but not the performance of species distribution models of stream biota. Ecol. Model. 257, 1–10.

Drake, J.M., Randin, C., Guisan, A., 2006, June. Modelling ecological niches with support vector machines. J. Appl. Ecol. 43, 424–432.

Elith, J., et al., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29, 129–151.

Engler, R., Guisan, A., Rechsteiner, L., 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. J. Appl. Ecol. 41 (2), 263–274.

Friedman, J.H., 1991. Multivariate adaptive regression splines. Ann. Stat. 19.

Fronzek, S., Carter, T., Luoto, M., 2011. Evaluating sources of uncertainty in modelling the impact of probabilistic climate change on sub-arctic Palsa mires. Nat. Hazards Earth Syst. Sci. 11, 2981–2995.

Gastón, A., García-Viñas, J., 2011. Modelling species distributions with penalised logistic regressions: a comparison with maximum entropy models. Ecol. Model. 222 (13), 2037–2041.

Gonzàlez, S., Soto-Centeno, J., Reed, D., 2011. Population distribution models: species distributions are better modeled using biologically relevant data partitions. BMC Ecol., 11.

Grantham, N., 2012. Analyzing Multiple Independent Spatial Point Processes. Senior Project. California Polytechnic State University.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135 (2), 147–186.

Hanspach, J., Kühn, I., Schweiger, O., Pompe, S., Klotz, S., 2011. Geographical patterns in prediction errors of species distribution models. Glob. Ecol. Biogeogr. 20 (5), 779–788.

Hengl, T., Sierdsema, H., Radović, A., Dilo, A., 2009. Spatial prediction of species' distributions from occurrence-only records: combining point pattern analysis, ENFA and regression-kriging. Ecol. Model. 220 (24), 3499–3511.

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

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

Hirzel, A.H., Helfer, V., Metral, F., 2001. Assessing habitat-suitability models with a virtual species. Ecol. Model. 145 (2), 111–121.

Jiménez-Valverde, A., Lobo, J.M., Hortal, J., 2012. Not as good as they seem: the importance of concepts in species distribution modelling. Divers. Distrib. 14, 885–890.

Jiménez-Valverde, A., 2012. Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. Glob. Ecol. Biogeogr. 21 (4), 498–507.

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the International Joint Conference on Artificial Intelligence, pp. 1137–1143.

Liu, C., Berry, P.M., Dawson, T.P., Pearson, R.G., 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography 28 (3), 385–393.

Liu, C., White, M., Newell, G., Griffioen, P., 2013. Species distribution modelling for conservation planning in Victoria, Australia. Ecol. Model. 249, 68–74.

Lobo, J.M., Jiménez-Valverde, A., Hortal, J., 2010. The uncertain nature of absences and their importance in species distribution modelling. Ecography 33 (1), 103–114.

Lobo, J.M., Tognelli, M.F., 2011. Exploring the effects of quantity and location of pseudo-absences and sampling biases on the performance of distribution models with limited point occurrence data. J. Nat. Conserv. 19 (1), 1–7.

Manel, S., Dias, J.M., Buckton, S.T., Ormerod, S.J., 1999. Alternative methods for predicting species distribution: an illustration with Himalayan river birds. J. Appl. Ecol. 36, 734–747.

Mateo, R.G., Croat, T.B., Felicísimo, A.M., Muñoz, J., 2010. Profile or group discriminative techniques? generating reliable species distribution models using pseudo-absences and target-group absences from natural history collections. Divers. Distrib. 16 (1), 84–94.

McPherson, J.M., Jetz, W., Rogers, D.J., 2004. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? J. Appl. Ecol. 41, 811–823.

Norris, D., Rocha-Mendes, F., Frosini de Barros Ferraz, S., Villani, J., Galetti, M., 2011. How to not inflate population estimates? Spatial density distribution of white-lipped peccaries in a continuous Atlantic forest. Anim. Conserv. 14 (5), 492–501.

Oney, B., Reineking, B., O'Neill, G., Kreyling, J., 2013. Intraspecific variation buffers projected climate change impacts on *Pinus contorta*. Ecol. Evol. 3 (2), 437–449.

Peterson, A.T., Soberón, J., Pearson, R.G., Robert, P., Anderson, R.P., Enrique Martínez-Meyer, E., Miguel Nakamura, M., 2011. Ecological Niches and Geographic Distributions (MPB-49). Princeton University Press, Princeton.

Petit, R.J., Brewer, S., Bordács, S., Burg, K., Cheddadi, R., Coart, E., Cottrell, J., Csaikl, U.M., van Dam, B., Deans, J.D., Espinel, S., Fineschi, S., Finkeldey, R., Glaz, I., Goicoechea, P.G., Jensen, J.S., König, A.O., Lowe, A.J., Madsen, S.F., Mátyás, G., Munro, R.C., Popescu, F., Slade, D., Tabbener, H., de Vries, S.G.M., Ziegenhagen, B., de Beaulieu, J.-L., Kremer, A., 2002a. Identification of refugia and post-glacial colonisation routes of European white oaks based on chloroplast DNA and fossil pollen evidence. For. Ecol. Manage. 156 (13), 49–74.

Petit, R.J., Csaikl, U.M., Bordács, S., Burg, K., Coart, E., Cottrell, J., van Dam, B., Deans, J.D., Dumolin-Lapégue, S., Fineschi, S., Finkeldey, R., Gillies, A., Glaz, I., Goicoechea, P.G., Jensen, J.S., König, A.O., Lowe, A.J., Madsen, S.F., Mátyás, G., Munro, R.C., Olalde, M., Pemonge, M.-H., Popescu, F., Slade, D., Tabbener, H., Taurchini, D., de Vries, S.G.M., Ziegenhagen, B., Kremer, A., 2002b. Chloroplast DNA variation in European white oaks: phylogeography and patterns of diversity based on data from over 2600 populations. For. Ecol. Manage. 156 (13), 5–26.

Petit, R.J., Latouche-Halle, C., Pemonge, M., Kremer, A., 2002c. Chloroplast DNA variation of oaks in France and the influence of forest fragmentation on genetic diversity. For. Ecol. Manage. 156 (1), 115–129.

Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecol. Model. 190 (34), 231–259.

Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J., Ferrier, S., 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecol. Appl. 19 (1), 181–197.

R Core Team, 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria http://www.R-project.org/

Scholkopf, B., Smola, A.J., 2001. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, USA.

Senay, S.D., Worner, S.P., Ikeda, T., 2013. Novel three-step pseudo-absence selection technique for improved species distribution modelling. PLOS ONE 8 (8), e71218.

Steyerberg, E.W., Vickers, A.J., Cook, N.R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M.J., Kattan, M.W., 2010. Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology (Cambridge, Mass.) 21 (1), 128–138.

Stokland, J.N., Halvorsen, R., Støa, B., 2011. Species distribution modelling – effect of design and sample size of pseudo-absence observations. Ecol. Model. 222 (11), 1800–1809.

Van der Wal, J., Shoo, L.P., 2009. Selecting pseudo-absence data for presence-only distribution modeling: how far should you stray from what you know? Ecol. Model. 4, 589–594.

Vaughan, I.P., Ormerod, S.J., 2005. The continuing challenges of testing species distribution models: testing distribution models. J. Appl. Ecol. 42, 720–730.

Wisz, M.S., Guisan, A., 2009. Do pseudo-absence selection strategies influence species distribution models and their predictions? An information-theoretic approach based on simulated data. BMC Ecol. 9 (1), 8.

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

Zaniewski, A.E., Lehmann, A., Overton, J.M., 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. Ecol. Model. 157 (2), 261–280.