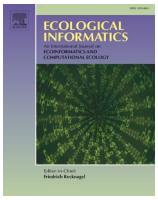
Ecological Informatics 69 (2022) 101683

Contents lists available at ScienceDirect

Ecological Informatics

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



Dealing with non-equilibrium bias and survey effort in presence-only invasive Species Distribution Models (iSDM); predicting the range of muntjac deer in Britain and Ireland

Marianne S. Freeman *, Jaimie T.A. Dick, Neil Reid

Institute of Global Food Security (IGFS), Queen's University Belfast, Belfast BT9 5DL, UK

ARTICLE INFO

Keywords:
Disequilibrium
Geographic Information System (GIS)
Maxent
Muntiacus reevesi
Random background
Spatial filtering
Targeted background
Weighted background

ABSTRACT

Invasive species managers utilise species records to inform management. These data can also be used in Species Distribution Models (SDM) to predict future spread or potential invasion of new areas. However, issues with non-equilibrium (also called disequilibrium) can cause difficulties in modelling invasive species that have not fully colonised their potential distribution and, in addition, sampling bias can result from a lack of information on survey effort, a particular issue for presence only modelling techniques. Geographical confounds are unavoidable when building iSDMs but there are methods that allow prediction to be optimised. We used maximum entropy (Maxent) to model suitable habitat for invasive Reeve's muntjac deer (*Muntiacus reevesi*) throughout Great Britain and Ireland comparing several methods that aimed to address invasive Species Distribution Modelling (iSDM) bias including spatial filtering, weighted background points and targeted background points built at varying spatial extents. Model evaluation metrics suggested that the model, which explicitly failed to account for non-equilibrium at the full extent of Great Britain and Ireland using random background points, predicted the species' current invasive range best. This highlighted that negative environmental relationships are likely to represent uncolonised areas rather than habitat selection and thus, low predicted suitability of uncolonised areas was misleading. Of the models that dealt with non-equilibrium conceptually best, by restricting the training extent to their current invasive range or core range, and utilised targeted background points accounting for survey effort (cells with other deer species recorded as present yet with no records for muntjac) as the best model evaluation metric, yielded relatively poor predictive performance. This implied limited habitat selectivity or avoidance within the colonised range which, when spatially extrapolated, suggested virtually all regions in Great Britain and Ireland may be vulnerable to future muntjac invasion.

1. Introduction

Knowledge and understanding of current distribution, spread and likely ultimate distribution of alien invasive species are prerequisite to their successful management (Hulme, 2009). Current distributions and species spread can be estimated using empirical survey data including species records, often sightings, and the rate of change in their range. Estimating an invasive species' likely ultimate distribution in an as yet uninvaded area is more challenging. Species Distribution Models (SDMs) are a widely used analytical technique in estimating the extent of suitable environmental niche envelopes for a given species; correlating species occurrence data with environmental variation (e.g., Guisan and Zimmermann, 2000). SDMs like all statistical analyses have inherent assumptions that must be met if the output results and spatial projections are to be robust, yet these are often ignored or broken when SDMs are used for alien invasive species (Elith et al., 2010).

SDMs utilise species record data that can be either: 1) presence-only that includes species occurrence data only (McDonald, 2013; McDonald et al., 2013) or ii) presence/absence that includes both species occurrence (presence) and data for sites surveyed but at which the species was not detected and thus assumed absent (Pearson, 2007). Species occurrence data are increasingly available for download from easily accessible online biological record centre databases, thus the use of presence-only data for SDM input has become common (Warton and Aarts, 2013). Moreover, the availability of easily downloaded and used freeware, for example, the machine learning maximum entropy platform, Maxent

* Corresponding author.
E-mail address: mfreeman02@qub.ac.uk (M.S. Freeman).

https://doi.org/10.1016/j.ecoinf.2022.101683
Received 1 July 2021; Received in revised form 15 May 2022; Accepted 15 May 2022
Available online 19 May 2022

(Elith et al., 2011; Phillips and Dudík, 2008), has made SDM available to the masses (there were 5720 papers that included the word 'Maxent' of which 413 had the word in their title on the Web of Science Core Collection during May 2022). Maxent typically uses randomly generated 'background points' distributed throughout the spatial extent of the analysis when defining the range of environmental variation relative to the environmental envelope suitable for species presence (Phillips et al., 2009). However, the complexity of multiple processes in invasion ecology presents significant difficulties when attempting to use SDM for predicting the extent of suitable environmental conditions for non-native species currently at non-equilibrium (also referred to as disequilibrium). Alien invasive species still in the process of invasion, have an actively expanding range edge boundary (wave front) such that their presence within their current range reflects their invasion history, rate of spread and habitat selection. However, areas beyond their current range cannot be characterised within any SDM as (pseudo) absence, by the use of background points, as such regions do not represent species absence or habitat avoidance but merely that the species has yet to invade such areas. Invasive Species Distribution Models (iSDMs) are often built to assume that the input data cover the full extent of the area over which spatial extrapolation is to take place (i.e., prediction into novel environments) yet frequently ignore the problem of species non-equilibrium (Gallien et al., 2012; Kelly et al., 2014).

Another concern with presence-only data is spatial autocorrelation resulting from sampling bias, as the lack of true absence data results in no information on the distribution or density of sampling effort (Dormann et al., 2007; Merow et al., 2013; Raes and Ter Steege, 2007). Spatial autocorrelation can result in over-representation of more accessible habitats or extensively sampled areas (Kramer-Schadt et al., 2013) and violates the assumption of independence leading to Type I errors or false positives (Dormann et al., 2007). Almost all species records suffer from potential sampling bias due to the methodology used in surveys or due to the focal species being either elusive, rare, common (thus under-represented in public sighting surveys), invading or understudied (Kramer-Schadt et al., 2013). Using SDMs for highly elusive invading species early in their colonisation process, therefore, presents an even greater source of potential bias due to recent inoculation, lag periods and limited spread i.e., habitat selectivity (Siesa et al., 2011). Thus, there is a danger that any iSDM is ultimately modelling sampling effort constrained by the species wave front rather than actual habitat selectivity and the species' ultimate range at full colonisation. Inaccurate model outputs due to ignored survey bias and error may influence invasive species management plans, erroneously.

Sampling bias using presence-only data can be addressed using three different methods: i) spatial filtering of the input data (Engler et al., 2004), ii) weighting background points (Elith et al., 2010) and iii) targeting background points (Phillips et al., 2009). Spatial filtering of input data (reducing either the spatial resolution i.e., cell or pixel size or reducing the density of species records in densely recorded clusters) helps to control potential spatial autocorrelation bias prior to model building. Kramer-Schadt et al. (2013) found that reducing the resolution of species records, to one record per 10km^2, increased the predictive power of their model. Taking one record within a cell or pixel at a lower spatial resolution than the original species record data is often used to filter spatial bias but also reduces fine-scale environmental variation obscuring local ecological relationships in the process. A different method is to thin out clustered records based on their geographical density, whereby the probability of a record removal is proportional to the density of occurrence records in the area of a kernel density grid (Verbruggen et al., 2013). Weighting background point selection allows for the manipulation of the model building process itself; altering the selection of background points to match the bias in species records. Some software programmes, including Maxent, allow the inclusion of a bias file to offset the effects of clustering by increasing the selection of background points from high-density locations (Elith et al., 2010). Alternatively, linear modelling approaches can be weighted to incorporate dispersal probabilities (Sullivan et al., 2012). Targeting background points is conceptually similar but allows the modeller to direct the selection of background points to specifically defined cells in a manner that reduces the probability of Type II errors i.e., false negatives. For example, using species records for similar, related taxa, from which to draw background points assumes that a surveyor or observer was present at a location from which they would have reported the focal taxa had they encountered it, by virtue of having recorded and reported a sister taxon. Thus, such locations are more closely allied to being true absences for the focal taxa than randomly selected background points and, therefore, presence-only models can include some attempt to account for bias due to survey or observer effort (e.g., Anderson and Gonzalez, 2011).

Here, we use a factorial combination of a range of techniques to explicitly test the comparative impact of a) spatial filtering input data, b) weighting background points and c) targeting background points when attempting to develop an iSDM for a highly elusive, invasive species whose range is expanding and is thus at non-equilibrium; Reeves' muntjac deer (*M. reevesi*) throughout Great Britain and Ireland. This species originated from China and Taiwan and a known population was introduced to Bedfordshire by the release of 11 individuals in 1901 (Chapman et al., 1994) with as few as 4–5 founding females (Freeman et al., 2016). Muntjac have since expanded to colonise most of England except the north-west, western region of Wales and most of Scotland. More recently, muntjac have become established in Ireland since around *ca.* 2008 (Dick et al., 2010) with multiple sightings but a highly restricted distribution i.e., they have yet to spread significantly (National Biodiversity Data Centre, 2021). Muntjac have been well documented since their introduction in both Great Britain and Ireland and as such are an ideal subject by which to test iSDM techniques.

2. Methods

2.1. Species record data

All species records for Reeves' muntjac deer (*M. reevesi*) throughout Great Britain and Ireland were collated from a wide range of data recording centres (Table S1). Only records with a spatial resolution of <1km^2 were retained for analysis and any duplicated records (those exact matches of recorder, date and location that appeared in different databases) were removed. In Great Britain, there were a total of 9905 records covering 4244 x 1km^2 squares whilst in Ireland there were 46 records in 17 x 1km^2 squares (Fig. 1a). Additionally, a total of 4970 species records were also collated for non-muntjac deer species, namely, red deer (*Cervus elaphus*), fallow deer (*Dama dama*), sika deer (*Cervus nippon*), roe deer (*Capreolus capreolus*), and Chinese water deer (*Hydropotes inermis*) throughout Great Britain and 691 records from Ireland.

2.2. Environmental parameters

All environmental data were summarised at a raster cell (pixel) resolution of 1 km matching the spatial resolution of the species records. A total of nine variables were chosen as potential predictors of muntjac distribution (Fig. S1). Habitat coverage data were obtained from the CORINE Land Cover map (EEA, 2007) and extracted using ArcGIS 10.5 (ESRI, California, USA), namely: the percentage (%) cover of: arable, broad-leaved woodland, coniferous plantations, scrub, grassland, and urban areas. These specific habitats were chosen as likely ecological determinants of muntjac presence due to their importance in previous ecological studies of the species (Chapman et al., 1994) and to deer in Britain and Ireland more generally. Bioclimatic factors, sourced from Worldclim (Hijmans et al., 2005) were also included, specifically, mean annual temperature (Bio1) and mean annual precipitation (Bio12). Altitude was explicitly excluded as animals have no direct sensory perception of elevation above sea level (i.e., air pressure) but instead perceive its correlates of temperature and precipitation. Nevertheless,

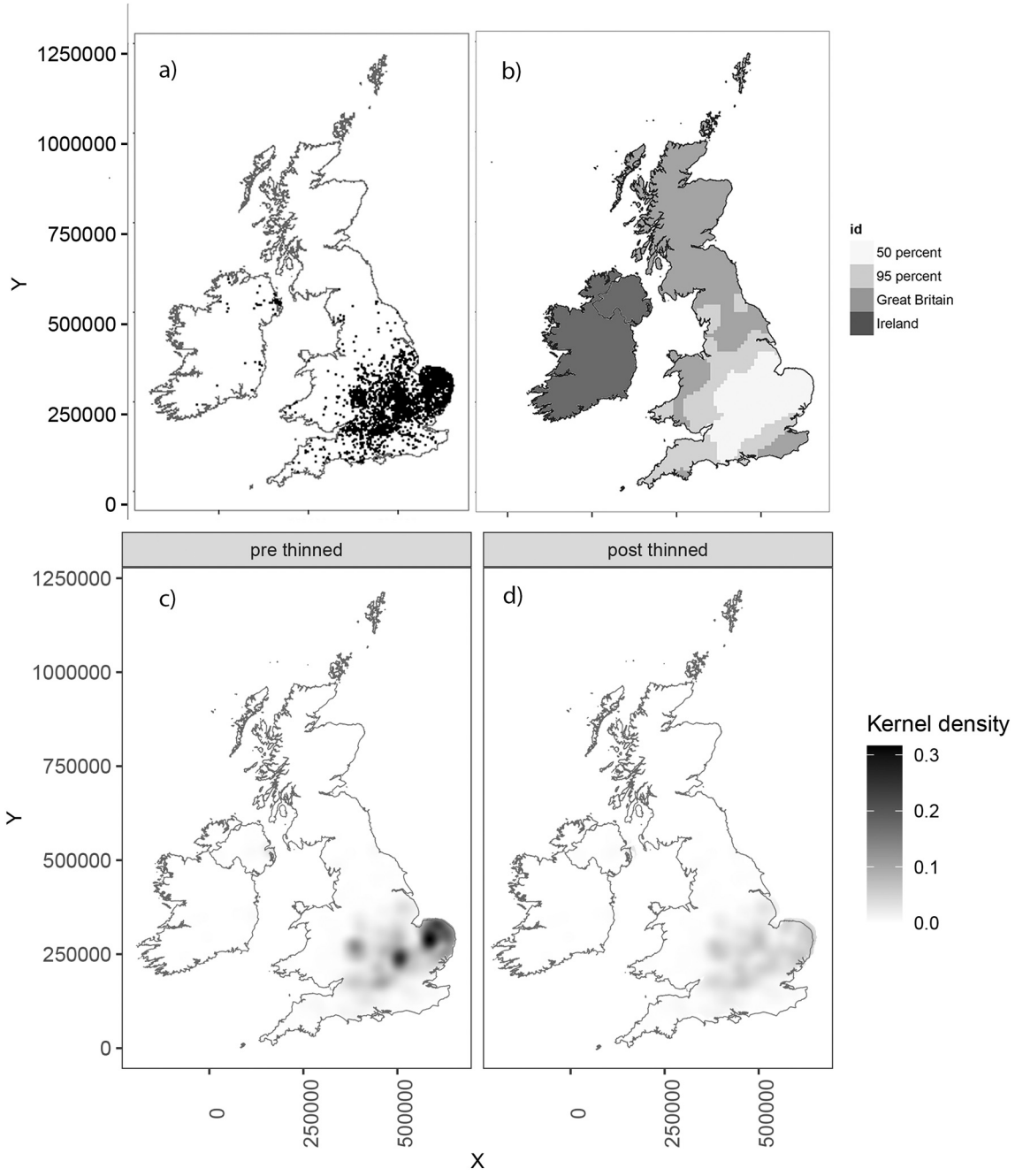


Fig. 1. a) Distribution of records Reeve's muntjac (*Muntiacus reevesi*) in Great Britain and Ireland, b) different extents used to select background points, c) kernel density of muntjac occurrences before thinning and d) after thinning.

topography was captured as average slope (in degrees) per 1km^2, extracted from a Digital Elevation Model (EEA, 2007). The datasets were checked for multicollinearity using pairwise Pearson correlations within each spatial extent tested as model inputs i.e., i) throughout Great Britain and Ireland, ii) Great Britain only, iii) within the 95% kernel range of muntjac records in Great Britain (their invasive range only) and iv) within the 50% kernel range in Great Britain (their invasive core range only; Fig. S2). Bivariates with significant correlations and a coefficient of $r_p > 0.75$ were deemed collinear.

2.3. Invasive Species Distribution Models (iSDMs)

iSDMs were built using the software *R* v. 3.6.3 (Team, 2020) and the package *dismo* (Hijmans et al., 2017) linking with Maxent v. 3.4.1 (Phillips et al., 2017).

To examine the effect of input data, four spatial extents were used with training (species record) datasets that matched each extent (Fig. 1b); i) muntjac records from throughout Great Britain and Ireland (i.e. 100% of records from Great Britain plus records from Ireland) with background points selected from throughout Great Britain and Ireland explicitly ignoring the problem of non-equilibrium, ii) 100% of records from throughout Great Britain only with background points selected from throughout Great Britain only; again ignoring non-equilibrium but focusing the model on the landmass with the best quality data i.e. largest number of records, iii) records from the 95% kernel range in Great Britain with background points selected from *within* this range i.e. restricted to the extent of the colonised area making some effort to account for non-equilibrium, iv) records from the 50% kernel range (i.e.

the species' core range) in Great Britain with background points selected from *within* the core range, again accounting for non-equilibrium but more completely than in the 95% kernel model. In the latter case, it could be assumed that muntjac have fully colonised their core range and thus any cells without species presence records (especially those with some form of surveyor or observer effort i.e., other deer species having been recorded) may represent habitat avoidance thus informing models more meaningfully before spatial extrapolation into areas beyond the species invasion wave front. In all cases, 10,000 background points were used and each model was projected (i.e., spatially extrapolated) to the entire extent of Great Britain and Ireland to examine variation in the predicted extent of suitable habitat beyond the species' colonised area. In each case, models were built using a training set comprising 75% of the species records randomly selected and evaluated using a test set comprising the remaining 25% of species records (Test dataset #1) with a further independent test comprising the 17 verified locations of muntjac records from Ireland only (Test dataset #2). The latter was to test the utility of a model built in one region with good data (Great Britain) in predicting the appearance of muntjac early in the establishment and colonisation phase in another region with poor data (Ireland). Environmental response curves were generated using a combination of linear, quadratic and product features.

Each of the iSDMs at the four different spatial extents were further subject to five treatments to test the impact of potential biases and error: a) Random backgrounds where 100% of training records were used, and background points were selected at random from the full extent of the model. b) Spatial filtering input data to reduce problems of spatial autocorrelation, where all species records were thinned using the software OccurrenceThinner (Verbruggen et al., 2013), in a manner to filter out a greater proportion of sightings where the density of records was highest based on their kernel density (Fig. 1c; with densities per cell rescaled to an index from 0 to 1). Species records were selected for deletion randomly from those cells with a kernel density index between 0.5 and 1.0 giving an increased chance of removing records from the highest density areas. Over 10 cross-validated model runs, an average of 2139 data points were retained per run (Fig. 1d). Random background points were selected for the model. c) Targeting background points i.e., use of non-random points (Phillips et al., 2009). In another attempt to account for likely sampling effort, background points were extracted from cells within which at least one record of another deer species was recorded i.e., someone visited those cells and was predisposed to reporting deer, if seen, but happened not to report muntjac (assumed absent). d) Spatial filtering (of presence records only) with targeted background points to assess their impact when combined. e) Weighting background points based on an estimated proxy for likely survey effort (Elith et al., 2010). A bias file was created to weight records on whether survey effort (sufficient to assume lack of presence is absence) was known for each cell. Some data recording centers provided a measure of sampling effort e.g., some species records were sourced from surveys that recorded absence in addition to presence-only incidental records from the public. All records were combined to create a bias grid (using the previous upweighted area combined with a weighted distribution of occurrences), calculated using the Gaussian function:

$$exp\frac{\left(-[d]^2\right)}{2s^2} \quad (1)$$

where d = nearest neighbour distance to each survey point and s = standard deviation based on the dispersal distance of muntjac taken as 1 km per year (Chapman et al., 1994) over a maximum of 10 years per survey (to reflect the variation in survey rate from the historically collected data), thus a value of 10,000 was chosen (see Elith et al., 2010).

Various thresholds can be chosen by which to render continuous predicted probability surfaces into binary maps indicative of likely presence or absence (suitable vs unsuitable habitat). The literature suggests convergence on the maximum test sensitivity plus specificity (TSS_{max}) as the single most appropriate threshold (Guisan et al., 2017). However, in this case we used two different test data sets and thus two different thresholds would have been generated making each non-comparable to the other. Thus, in this case we used the 10th percentile training presence (Elith et al., 2010) which was the same across both test datasets allowing model evaluation metrics to be directly compared.

2.4. Model evaluation

There are many varying opinions in the literature as to the most suitable SDM evaluation statistic to use. The most commonly used is the Area Under the Curve (AUC value) of the Receiver Operating Characteristic (ROC) curve as this is threshold and scale invariant (Merow et al., 2013). However, AUC values are influenced by the extent of model prediction (Smith, 2013). If the extent of the model is large and the species has a restricted distribution within that extent, then AUC values will be artificially inflated (as would be expected for an invasive species at non-equilibrium). Thus, we used a corrected AUC (cAUC) as suggested by Hijmans (2012). Alternative metrics of model fit are not without their own issues (Allouche et al., 2006) such as sensitivity (proportion of presences which are correctly predicted, or True Positive Rate (TPR)), specificity (proportion of absences which are correctly predicted) or True Skill Statistic (TSS), a prevalence independent model metric calculated using sensitivity and specificity. False negatives (omissions) and false positives (commission) can lead to errors, arising from species not being at equilibrium and can affect such metrics. On the other hand, Kappa (k) utilises input species occurrences and background points that have been adjusted for a random proportion of correct predictions, thus is as objective a measure of prediction accuracy as any metric (Monserud and Leemans, 1992). The Kappa statistic involves the use of commission *and* omission errors (Manel et al., 2001), and although it does take into account prevalence like the True Skill Statistic (Allouche et al., 2006), the thresholds are widely accepted and so useful in model evaluation (Altman, 1990; Landis and Koch, 1977; Monserud and Leemans, 1992). Due to all the issues associated with the various evaluation statistics and no consensus on the best measure, all model statistics (AUC, TPR, TSS, Omission, and Kappa values) are presented here, to allow for comparison.

An AUC value $\geq$0.9 was considered an excellent model fit, a value between 0.7 and 0.9 was deemed good and $\leq$ 0.5 no better than random (Hosmer et al., 2013). TPR should be closer to 1 reflecting high sensitivity. TSS lies between 0 and 1 with >0.9 considered perfect, 0.85–0.9 excellent, 0.7–0.85 very good, 0.5–0.7 good, 0.4–0.5 fair and $\leq$ 0.4 representing a poor fit (Landis and Koch, 1977). Omission values should be closer to 0 to reflect low false negative scores (Peterson, 2006). A Kappa score, based on the 10th percentile training presence logistic threshold, of 1 indicated models fitted perfectly, 0.85–0.99 excellent, 0.70–0.85 very good, 0.55–0.70 good, 0.40–0.55 fair 0.20–0.40 poor 0.05–0.20 very poor and < 0.05 indicated no agreement. Negative values indicated an especially poor model fit (Landis and Koch, 1977).

2.5. Multivariate Environmental Similarity Surface (MESS)

In order to understand the extent to which models are transferable over the range of environmental conditions they are fitted to; a Multivariate Environmental Similarity Surface (MESS) analysis was undertaken using the '*Dismo*' package in R using muntjac presence locations. A score of <0 indicates dissimilar conditions to the training sample.

3. Results

Multicollinearity between environmental variables, tested independently at each of the four spatial extents modelled, was not deemed a substantial problem with only one pair of significant bivariates at one spatial extent (Fig. S2). Mean annual temperature (Bio_1) and annual

precipitation (Bio_12) were negatively correlated only at the full extent of Great Britain and Ireland ($r_p = 0.7$) but as this was not >0.75 both were retained. This also ensured all models, at all extents, had the same environmental parameters facilitating direct comparability; a necessity when comparing modelling techniques. In addition, the Multivariate Environmental Similarity Surface (MESS) map (Fig. 2) suggested that environmental combinations of only the highest altitudes (the Scottish Highlands and Islands, Cambrian mountains of Wales, Pennines of northern England and the uplands of Donegal and Kerry in Ireland) and the urban centre of London were not adequately captured by training models.

Across all spatial extents and all model treatments, mean annual precipitation (Bio12) had the highest median permutation importance at 66.5% (Fig. 3). All other variables had substantially lower importance values, however, broad-leaved woodland, mean annual temperature (Bio1) and arable were the next most important variables (contributing 3.9%, 3.6% and 3.5% respectively). The muntjac iSDMs built at the full spatial extent of Great Britain and Ireland with random background points had the highest training AUC and test set cAUC, TPR and TSS values and lowest Omission rates of any model (Table 1). Using kappa as the most conceptually optimal evaluation metric and the test set (randomly withheld 25% of species records) to be the most rigorous test of each model, then the single best model was at the full extent of Great Britain and Ireland using 100% of species records with weighted background points (weighted, based on an estimated proxy for likely survey effort) with a fair $k = 0.509$. This value was notably larger than any other kappa value from the other models but was confirmed as accurate after double checking model inputs and structure. The best kappa value

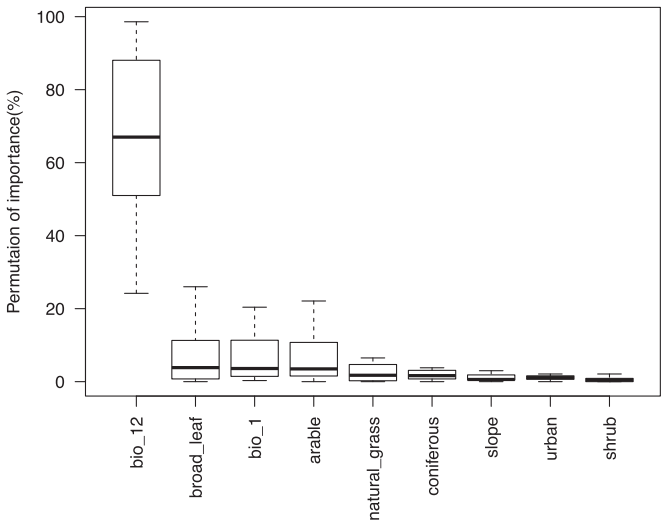


Fig. 3. Permutation importance of environmental parameters used to build 20 Maxent models using different spatial extents and bias treatments. Bold line represents the median, boundaries are at the first and last quartile and whiskers are at 1.5 inter-quartile range.

for models built using the 95% and 50% kernel core ranges (those models that were conceptually least vulnerable to false negatives of background points beyond their invasive wave front boundary) was for the targeted 50% core range weighted by likely survey effort by using

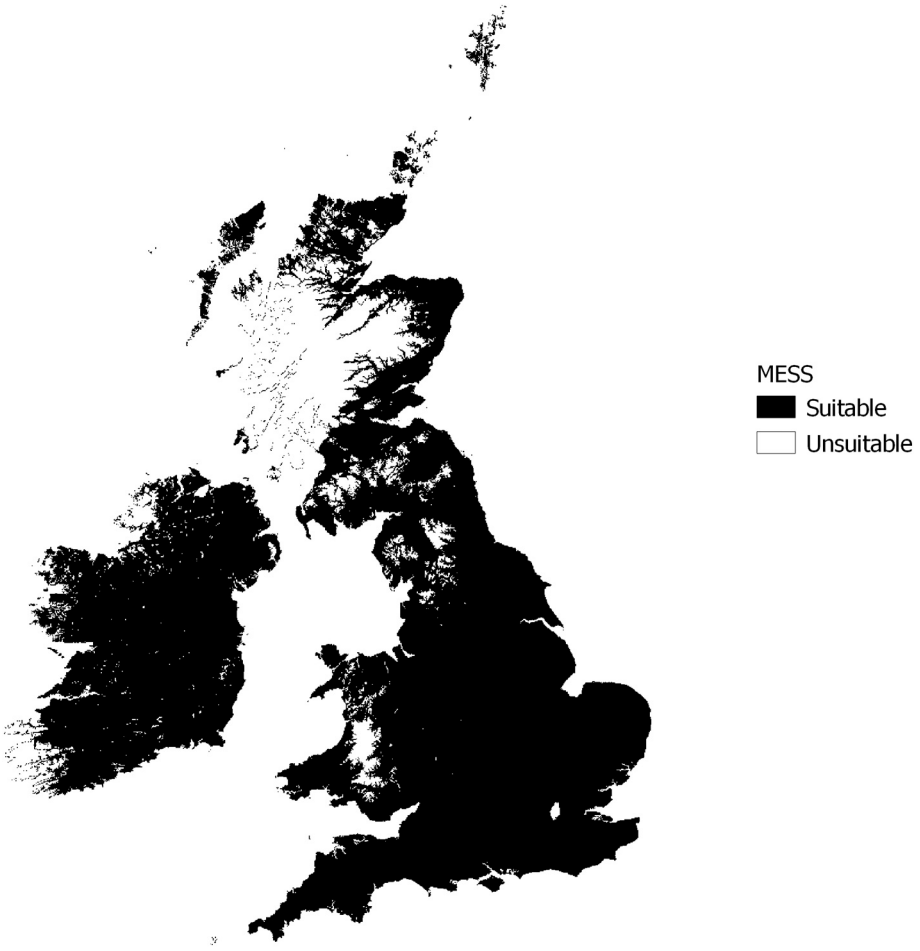


Fig. 2. Multivariate Environmental Similarity Surface (MESS) analysis with white areas indicating regions whose bioclimatic habitat envelope were poorly captured by model training potentially limiting the value of model spatial extrapolation.

Table 1
iSDM evaluation metrics for a) the training sets representing 75% of presence data selected at random, b) Test dataset #1 representing 25% of presence data selected at random and c) Test dataset #2 representing presence data from Ireland only for four spatial extents with five model treatments with i) random background points, ii) filtered (pre-thinned) input records adjusted for presence record density using OccurrenceThinner, iii) targeted selection of background points drawn from only those 1 km squares with a deer record, iv) filtered (pre-thinned) *and* targeted background point selection and v) weighted background points based on an estimated proxy for likely survey effort derived from a bias grid, throughout Great Britain and Ireland (GB & Ireland), Great Britain or restricted to the area currently colonised by muntjac (95% kernel range) or its core range (50% kernel range). The threshold used was the 10th percentile training presence. For each metric the best model is highlighted in bold where higher was better for AUC, TPR, TSS and Kappa and lower better for Omission.

Model evaluation metrics	i) Random background				ii)Filtered				iii) Targeted				iv) Filtered *and* Targeted				v) Weighted			
	GB & Ireland Isles	Great Britain	Muntjac range		GB & Ireland Isles	Great Britain	Muntjac range		GB & Ireland Isles	Great Britain	Muntjac range		GB & Ireland Isles	Great Britain	Muntjac range		GB & Ireland Isles	Great Britain	Muntjac range	
	100%	100%	95%	50%	100%	100%	95%	50%	100%	100%	95%	50%	100%	100%	95%	50%	100%	100%	95%	50%
TSS_{max}	0.204	0.251	0.403	0.52	0.264	0.331	0.481	0.592	0.55	0.556	0.489	0.508	0.579	0.579	0.51	0.532	0.487	0.466	0.503	0.582
a) Training set (75% random)																				
AUC	**0.875**	0.847	0.777	0.785	0.849	0.824	0.727	0.708	0.783	0.780	0.806	0.787	0.692	0.705	0.722	0.707	0.791	0.790	0.744	0.743
Omission	**0.024**	0.027	0.053	0.181	0.042	0.041	0.122	0.330	0.167	0.164	0.112	0.099	0.223	0.222	0.127	0.093	0.101	0.097	0.113	0.213
TPR	**0.977**	0.974	0.949	0.819	0.958	0.959	0.878	0.670	0.833	0.836	0.888	0.901	0.777	0.778	0.873	0.907	0.899	0.903	0.887	0.787
k	**0.635**	0.565	0.407	0.485	0.459	0.393	0.242	0.275	0.553	0.547	0.607	0.591	0.365	0.382	0.433	0.388	0.518	0.386	0.033	0.409
TSS	**0.750**	0.692	0.552	0.571	0.698	0.647	0.454	0.415	0.566	0.560	0.611	0.574	0.384	0.409	0.444	0.414	0.582	0.579	0.489	0.485
b) Test set #1 (25% random)																				
cAUC	**0.780**	0.753	0.662	0.691	0.757	0.729	0.584	0.599	0.691	0.689	**0.712**	0.688	0.607	0.622	0.658	0.613	0.687	0.701	0.621	0.663
Omission	**0.014**	0.014	0.083	0.172	0.028	0.033	0.211	0.350	0.152	0.147	0.102	0.099	0.193	0.189	0.058	0.082	0.112	0.084	0.108	0.173
TPR	**0.986**	0.986	0.917	0.828	0.972	0.967	0.789	0.650	0.848	0.853	0.898	0.901	0.807	0.811	0.942	0.918	0.888	0.916	0.892	0.827
k	0.210	0.166	0.094	0.144	0.104	0.080	0.211	0.052	0.226	0.237	0.287	**0.316**	0.078	0.080	0.105	0.108	**0.509**	0.131	0.072	0.119
TSS	**0.758**	0.704	0.522	0.580	0.713	0.655	0.365	0.396	0.581	0.576	0.621	0.574	0.413	0.442	0.514	0.425	0.571	0.599	0.441	0.525
c) Test set #2 (Irish records only)																				
cAUC	0.687	0.701	0.621	0.663	0.515	0.693	0.599	0.525	0.531	0.557	0.613	0.528	0.535	0.539	0.539	0.565	0.598	0.586	0.615	0.648
Omission	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.923
TPR	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.077
k	−0.003	0.000	−0.003	−0.003	−0.003	−0.003	−0.003	−0.003	−0.005	−0.005	−0.006	−0.011	−0.005	−0.005	−0.006	−0.011	−0.003	−0.003	−0.003	−0.002
TSS	−0.228	−0.585	−0.394	−0.248	−0.260	−0.312	−0.424	−0.255	−0.267	−0.277	−0.276	−0.328	−0.394	−0.369	−0.428	−0.494	−0.317	−0.317	−0.451	−0.225

record density of other deer as a surrogate of potential observer bias, but model performance was poor ($k = 0.316$). Equally, the best cAUC value for models within these restricted spatial extents was for the targeted 95% range with a good level of model performance (cAUC = 0.712). Models evaluated using test dataset #2 (Irish records only), regardless of their extent or conditions, failed to predict the appearance of the first muntjac records in Ireland with high omission rates (>0.923), and negative kappa values suggesting notably poor fit. Models with restricted ranges (95 or 50%) and specified background points (either targeted or weighted) offered the best visualisation (Fig. 4).

4. Discussion

Despite sampling bias being widely acknowledged as a concern in invasive Species Distribution Models (iSDM) and a large number of studies attempting to address the problem, no conclusion has been drawn as to the best method to deal with it (Kramer-Schadt et al., 2013; Verbruggen et al., 2013). Here, three very different options to reduce sampling bias were examined i) weighted background points, ii) filtered (pre-thinned) records and iii) targeted background points, fitted to various spatial extents. AUC values were, in the main, higher when adopting more expansive input extents than restricted extents. However, it is known that these can be misleading; reflecting artificial inflation as

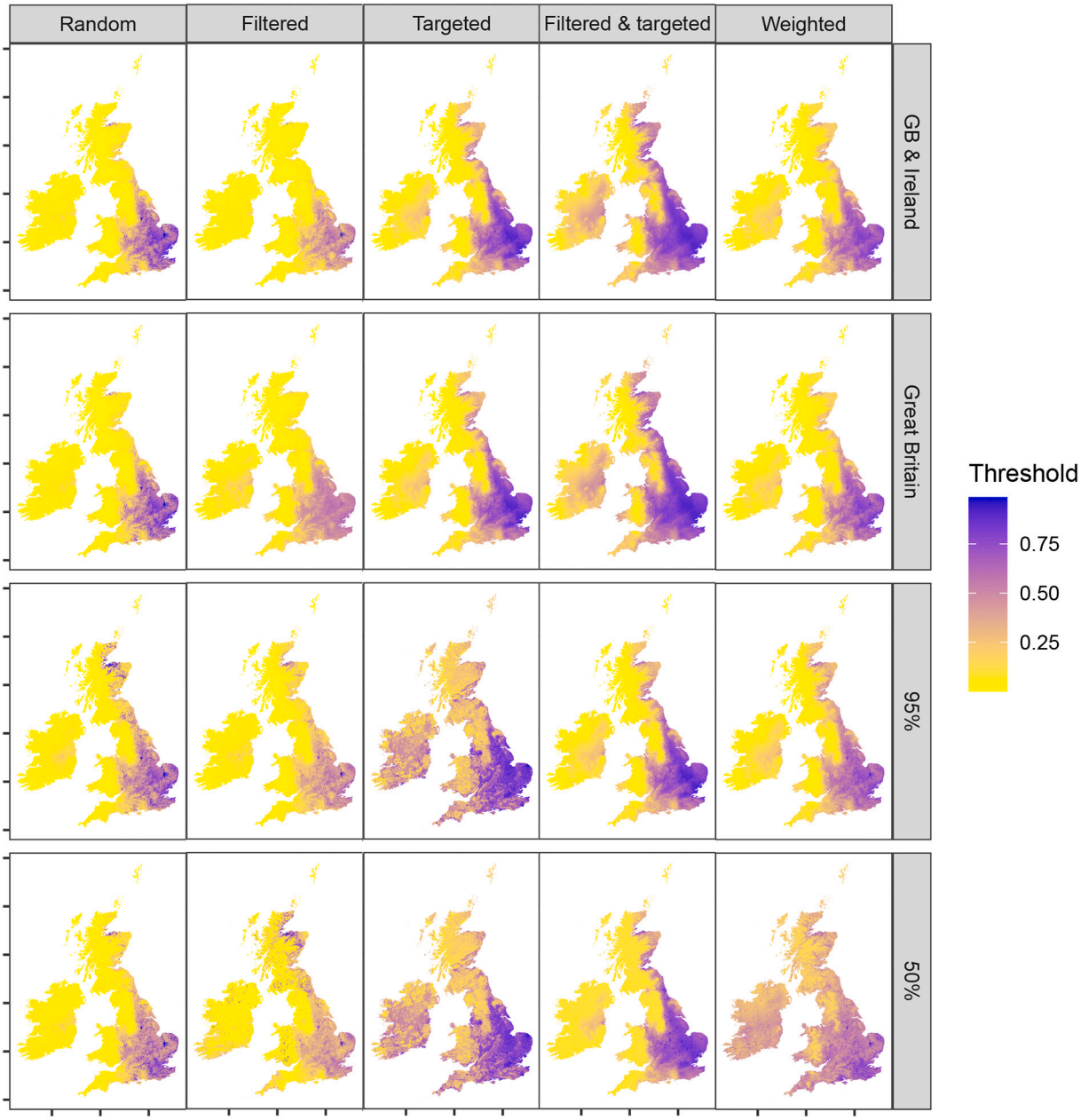


Fig. 4. MaxEnt spatial predictions of landscape suitability across different modelling extents (rows); Great Britain and Ireland, Great Britain only, the species' invasive range (95% kernel of species records) and core invasive range (50% kernel) and data bias treatments (columns); random background points, pre-filtered points, targeted background points, pre-filtered data with targeted background points and weighted background points. The random and weighted models covering Great Britain and Ireland predicted the species current invasive range best (but are conceptually problematic drawing background points from beyond the species invasion wave front) while the targeted models spatially restricted to the 95% and 50% kernel species ranges and extrapolated to the rest of Great Britain and Ireland (the models that conceptually dealt with non-equilibrium best) suggested few regions are invulnerable to future invasion by range expansion with the possible exception of the highest elevations (though these were notably poorly captured by model training; see Fig. 2).

background points are drawn from areas beyond the area colonised by muntjac and thus do not represent real non-selection or avoidance of habitats but a statistical artefact created by non-uniform coverage of habitats leading to apparently strong, but erroneous, species responses (Allouche et al., 2006; Merow et al., 2013; Smith, 2013).

Despite high AUC scores for background points taken from the whole of Great Britain and Ireland, the most biologically suitable extent to take background points for a range shifting species such as muntjac is the reachable area they have had the opportunity to encounter so far (Elith et al., 2010). Thus, the restricted kernel range of their current distribution is the most suitable option to avoid discounting environmental variable gradients that have not yet been encountered. The single best model using the restricted area colonised by muntjac, was targeted with selection of background points drawn from 1 km squares with deer records (those with known survey effort) yielding a spatial extrapolation that suggested most of Great Britain and Ireland is vulnerable to invasion by range expansion with the exception of the highest elevations. However, the Multivariate Environmental Similarity Surface (MESS) analysis suggested that model training failed to adequately capture the bioclimatic-habitat envelope of the highest elevations (as these are missing from the species present invasive range) and thus the lower predicted suitability of higher elevations may be an unavoidable artefact of the input data and model construction and may not represent lower actual suitability in the event of their range expansion. Whether muntjac colonise the highlands of Great Britain and Ireland remains to be seen. The lower performance of the 50% core range model may suggest that within their fully colonised range few areas that are available for colonisation remain unoccupied and that no major area or habitat is invulnerable to invasion. Indeed, Ward et al. (2021) demonstrate that the pattern of range expansion observed currently is consistent with continuous expansion and in-filling. Naturally such models fail to discriminate clear bioclimatic or habitat associations and as such are equally poor in predicting the appearance of records in Ireland, which was likely driven by the locations of introduction (for example, proximity to captive herds and potential escapees as well as movement by human agency) more than colonisation of optimal habitat.

Importance permutations values suggested the only variable that consistently and strongly effected models was annual precipitation with muntjac exhibiting a negative association with rainfall. Muntjac, in their native range in Southeast China, are used to high rainfall similar in range to that of much of Scotland, where rainfall is highest in Great Britain and Ireland. However, this area is currently uncolonized, with the drier south-west of England having lowest rainfall, yet the highest muntjac record density, by virtue that this was the location of their introduction from Bedfordshire. There is a confounding geographical effect where muntjac have yet to colonise the full gradient of precipitation in Great Britain which is unavoidable in any iSDM. The species may be at a climatic disequilibrium and so this failure to colonise has either resulted from strong associations with non-climatic factors that correlate or a dispersal lag. Whilst, overall, the models have poor performance they are, nevertheless, useful. If muntjac had a clear and specific bioclimatic-habitat envelope this would have been captured by models yielding good predictive fit. As this was not the case, we conclude that muntjac do not exhibit a strong pattern in their distribution within their invasive range and within the environmental conditions in those areas that are as yet uncolonized but lie within the range of conditions in their native range. This suggests that nowhere in Great Britain or Ireland will be imperious to invasion. Thus, it is reasonable to assume that muntjac will continue their south-east to north-west colonisation of Great Britain and, where introductions have occurred in Ireland, with subsequent inaction, it is reasonable to expect muntjac to colonise the whole of the island of Ireland in time spreading from known locations in the east in a westerly direction.

Invasive deer distributions are difficult to predict due to increasing species spread that violates Species Distribution Model assumptions (Elith et al., 2010) and in many cases human-mediated dispersal, over large ranges, is involved (Dolman and Wäber, 2008). Difficulties in modelling human dispersed species have previously been encountered (Richardson et al., 2011; Rödder, 2009; Wilson et al., 2009). Chivers and Leung (2012) demonstrated the difference in species distribution predictive abilities based on the chosen human-mediated vector and suggest a modelling framework that incorporates human behaviour to better predict invasion, while Croft et al. (2019) included a spatial factor to account for the anthropogenic effect on presence and absence. The specific mode of continuous invasion of muntjac still remains unknown with the possibility of accidental escapes from captive locations or intentional release from members of the public (Chapman et al., 1994), and a better understanding of this propagule network structure could improve any predictive model.

SDM has previously been implemented in deer distribution research; from invasive distribution modelling (Croft et al., 2019; Gormley et al., 2011) to habitat suitability (Chapman et al., 1994) and interspecific interactions (Acevedo et al., 2010). Using presence/absence data, Chapman et al. (1994) found that muntjac, selected for arable land and avoided marginal upland land classes, however, they were still unable to determine a pattern in their dispersal. When comparing test set cAUC values, the targeted 95% model performed better than the 0.7 benchmark stated by Croft et al. (2019). No previous models have produced high resolution maps ($<10km^2$ resolution) for Great Britain and Ireland. With relatively small home ranges, 1 km was a more biologically relevant resolution with which to construct a model for muntjac, with higher spatial scales likely more informative.

5. Conclusion

We suggest that invasive Species Distribution Models (iSDMs) restricted to already invaded regions with some attempt to account for survey effort are conceptually best (likely defining habitat selectivity most accurately) yet in reality performed relatively poorly suggesting little habitat avoidance within colonised areas; the corollary of which is that most regions into which the model was extrapolated appear vulnerable to future invasion by continued range expansion. Future work would benefit from surveys of species presence and absence to define habitat selection, both within the colonised invasive range and at their expanding wave front. In addition, as muntjac have a notably small home range extent, modelling at higher spatial resolutions may further improve model discrimination. An understanding of species dispersal dynamics and population network structure may help determine range expansion processes more reliably. In any case, the evidence suggests nowhere in Great Britain and Ireland is invulnerable to future muntjac invasion by range expansion and thus, further negative impacts on native ecosystems are to be expected.

Declaration of Competing Interest

None.

Acknowledgments

This study was part of a PhD funded by the Department of Employment and Learning (DEL), the funding mechanism now transferred to the Department for the Economy (DfE), Northern Ireland. The authors thank the National Biodiversity Network (NBN) and local recording centres (see Table S1 for full list) for access to species records, and the two anonymous reviewers for their comments, that have helped improve the paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2022.101683.

References

Acevedo, P., Ward, A.I., Real, R., Smith, G.C., 2010. Assessing biogeographical relationships of ecologically related species using favourability functions: a case study on British deer. Divers. Distrib. 16, 515–528. https://doi.org/10.1111/j.1472-4642.2010.00662.x.

Allouche, O., Tsoar, A., Kadmon, R., 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). J. Appl. Ecol. 43, 1223–1232. https://doi.org/10.1111/j.1365-2664.2006.01214.x.

Altman, D., 1990. Practical Statistics for Medical Research. CRC Press.

Anderson, R.P., Gonzalez, I., 2011. Species-specific tuning increases robustness to sampling bias in models of species distributions: an implementation with Maxent. Ecol. Model. 222, 2796–2811. https://doi.org/10.1016/j.ecolmodel.2011.04.011.

Chapman, N., Harris, S., Stanford, A., 1994. Reeves' Muntjac Muntiacus reevesi in Britain: their history, spread, habitat selection, and the role of human intervention in accelerating their dispersal. Mammal Rev. 24, 113–160. https://doi.org/10.1111/j.1365-2907.1994.tb00139.x.

Chivers, C., Leung, B., 2012. Predicting invasions: alternative models of human-mediated dispersal and interactions between dispersal network structure and Allee effects. J. Appl. Ecol. 49, 1113–1123. https://doi.org/10.1111/j.1365-2664.2012.02183.x.

Croft, S., Ward, A.I., Aegerter, J.N., Smith, G.C., 2019. Modeling current and potential distributions of mammal species using presence-only data: A case study on British deer. Ecol. Evo 9 (15), 8724–8735.

Dick, J.T.A., Freeman, M., Provan, J., Reid, N., 2010. First record of free-living Reeves' munjtac deer (*Muntiacus reevesi* (Ogilby 1839)) in Northern Ireland. Irish Nat. J. 31, 152.

Dolman, P.M., Wäber, K., 2008. Ecosystem and competition impacts of introduced deer. Wildl. Res. 35, 202. https://doi.org/10.1071/WR07114.

Dormann, C.F., McPherson, J.M., Araújo, M.B., Bivand, R., Bolliger, J., Carl, G., Davies, G., Hirzel, A., Jetz, W., Kissling, W. Daniel, Kühn, I., Ohlemüller, R., Peres-Neto, P.R., Reineking, Björn, Schröder, B., Schurr, F.M., Wilson, R., 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography (Cop.). 30, 609–628. https://doi.org/10.1111/j.2007.0906-7590.05171.x.

Elith, J., Kearney, M., Phillips, S., 2010. The art of modelling range-shifting species. Methods Ecol. Evol. 1, 330–342. https://doi.org/10.1111/j.2041-210X.2010.00036.x.

Elith, J., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E., Yates, C.J., 2011. A statistical explanation of Max Ent for ecologists. Divers. Distrib. 17, 43–57. https://doi.org/10.1111/j.1472-4642.2010.00725.x.

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, 263–274. https://doi.org/10.1111/j.0021-8901.2004.00881.x.

© European Union, 2007. Copernicus Land Monitoring Service European Environment Agency. (EEA).

Freeman, M.S., Beatty, G.E., Dick, J.T.A., Reid, N., Provan, J., 2016. The paradox of invasion: Reeves' muntjac deer invade the British Isles from a limited number of founding females. J. Zool. 298 https://doi.org/10.1111/jzo.12283.

Gallien, L., Douzet, R., Pratte, S., Zimmermann, N.E., Thuiller, W., 2012. Invasive species distribution models - how violating the equilibrium assumption can create new insights. Glob. Ecol. Biogeogr. 21, 1126–1136. https://doi.org/10.1111/j.1466-8238.2012.00768.x.

Gormley, A.M., Forsyth, D.M., Griffioen, P., Lindeman, M., Ramsey, D.S.L., Scroggie, M. P., Woodford, L., 2011. Using presence-only and presence-absence data to estimate the current and potential distributions of established invasive species. J. Appl. Ecol. 48, 25–34. https://doi.org/10.1111/j.1365-2664.2010.01911.x.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135, 147–186. https://doi.org/10.1016/S0304-3800(00)00354-9.

Guisan, A., Thuiller, W., Zimmermann, N.E., 2017. Habitat Suitability and Distribution Models, Habitat Suitability and Distribution Models, Cambridge University Press. https://doi.org/10.1017/9781139028271.

Hijmans, R.J., 2012. Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. Ecology 93, 679–688.

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. https://doi.org/10.1002/joc.1276.

Hijmans, R.J., Phillips, S., Leathwick, J., Elith, J., Hijmans, M.R.J., 2017. Package 'dismo'. Circles 9 (1), 1–68.

Hosmer, D.W., Lemeshow, S., Sturdivant, R.X., 2013. Applied Logistic Regression, Applied Logistic Regression, Third edition. Wiley. https://doi.org/10.1002/9781118548387. Wiley Series in Probability and Statistics.

Hulme, P.E., 2009. Trade, transport and trouble: managing invasive species pathways in an era of globalization. J. Appl. Ecol. 46, 10–18. https://doi.org/10.1111/j.1365-2664.2008.01600.x.

Kelly, R., Lundy, M.G., Mineur, F., Harrod, C., Maggs, C.A., Humphries, N.E., Sims, D.W., Reid, N., 2014. Historical data reveal power-law dispersal patterns of invasive aquatic species. Ecography (Cop.). 37, 581–590. https://doi.org/10.1111/j.1600-0587.2013.00296.x.

Kramer-Schadt, S., Niedballa, J., Pilgrim, J.D., Schröder, B., Lindenborn, J., Reinfelder, V., Stillfried, M., Heckmann, I., Scharf, A.K., Augeri, D.M., Cheyne, S.M., Hearn, A.J., Ross, J., Macdonald, D.W., Mathai, J., Eaton, J., Marshall, A.J., Semiadi, G., Rustam, R., Bernard, H., Alfred, R., Samejima, H., Duckworth, J.W., Breitenmoser-Wuersten, C., Belant, J.L., Hofer, H., Wilting, A., 2013. The importance of correcting for sampling bias in Max Ent species distribution models. Divers. Distrib. 19, 1366–1379. https://doi.org/10.1111/ddi.12096.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159. https://doi.org/10.2307/2529310.

Manel, S., Williams, H.C., Ormerod, S.J., 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. J. Appl. Ecol. 38, 921–931. https://doi.org/10.1046/j.1365-2664.2001.00647.x.

McDonald, T.L., 2013. The point process use-availability or presence-only likelihood and comments on analysis. J. Anim. Ecol. 82, 1174–1182. https://doi.org/10.1111/1365-2656.12132.

McDonald, L., Manly, B., Huettmann, F., Thogmartin, W., 2013. Location-only and use-availability data: analysis methods converge. J. Anim. Ecol. 82, 1120–1124. https://doi.org/10.1111/1365-2656.12145.

Merow, C., Smith, M.J., Silander, J.A., 2013. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. Ecography (Cop.). 36, 1058–1069. https://doi.org/10.1111/j.1600-0587.2013.07872.x.

Monserud, R.A., Leemans, R., 1992. Comparing global vegetation maps with the Kappa statistic. Ecol. Model. 62, 275–293. https://doi.org/10.1016/0304-3800(92)90003-W.

National Biodiversity Data Centre, 2021. Ireland, Chinese Muntjac (*Muntiacus reevesi*). accessed 03 May 2021. https://maps.biodiversityireland.ie/Species/119475.

Pearson, R., 2007. Species' distribution modeling for conservation educators and practitioners. Lessons Conserv. 3, 54–89.

Peterson, A.T., 2006. Uses and requirements of ecological niche models and related distributional models. Biodivers. Inform. 3, 59–72. https://doi.org/10.17161/bi.v3i0.29.

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

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, 181–197. https://doi.org/10.1890/07-2153.1.

Phillips, S.J., Anderson, R.P., Dudík, M., Schapire, R.E., Blair, M.E., 2017. Opening the black box: an open-source release of Maxent. Ecography (Cop.). 40, 887–893. https://doi.org/10.1111/ecog.03049.

Raes, N., Ter Steege, H., 2007. A null-model for significance testing of presence-only species distribution models. Ecography (Cop.). https://doi.org/10.1111/j.2007.0906-7590.05041.x.

Richardson, D.M., Carruthers, J., Hui, C., Impson, F.A.C., Miller, J.T., Robertson, M.P., Rouget, M., Le Roux, J.J., Wilson, J.R.U., 2011. Human-mediated introductions of Australian acacias - a global experiment in biogeography. Divers. Distrib. 17, 771–787. https://doi.org/10.1111/j.1472-4642.2011.00824.x.

Rödder, D., 2009. Human Footprint, facilitated jump dispersal, and the potential distribution of the invasive Eleutherodactylus johnstonei Barbour 1914 (Anura Eleutherodactylidae). Tropical Zool. 22, 205–217.

Siesa, M.E., Manenti, R., Padoa-Schioppa, E., de Bernardi, F., Ficetola, G.F., 2011. Spatial autocorrelation and the analysis of invasion processes from distribution data: a study with the crayfish Procambarus clarkii. Biol. Invasions 13, 2147–2160. https://doi.org/10.1007/s10530-011-0032-9.

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

Sullivan, M.J.P., Davies, R.G., Reino, L., Franco, A.M.A., 2012. Using dispersal information to model the species-environment relationship of spreading non-native species. Methods Ecol. Evol. 3, 870–879. https://doi.org/10.1111/j.2041-210X.2012.00219.x.

Team, R.C.-R.C., 2020. R: A Language and Environment for Statistical Computing.

Verbruggen, H., Tyberghein, L., Belton, G.S., Mineur, F., Jueterbock, A., Hoarau, G., Gurgel, C.F.D., De Clerck, O., 2013. Improving transferability of introduced species' distribution models: new tools to forecast the spread of a highly invasive seaweed. PLoS One 8, e68337. https://doi.org/10.1371/journal.pone.0068337.

Ward, A.I., Richardson, S., Mergeay, J., 2021. Reeves' muntjac populations continue to grow and spread across Great Britain and are invading continental Europe. Eur. J. Wildl. Res. 67, 34. https://doi.org/10.1007/s10344-021-01478-2.

Warton, D., Aarts, G., 2013. Advancing our thinking in presence-only and used-available analysis. J. Anim. Ecol. 82, 1125–1134. https://doi.org/10.1111/1365-2656.12071.

Wilson, J.R.U., Dormontt, E.E., Prentis, P.J., Lowe, A.J., Richardson, D.M., 2009. Something in the way you move: dispersal pathways affect invasion success. Trends Ecol. Evol. https://doi.org/10.1016/j.tree.2008.10.007.