# Combining multiple data sources in species distribution models while accounting for spatial dependence and overfitting with combined penalised likelihood maximisation

Ian W. Renner[1]*, Julie Louvrier[2], and Olivier Gimenez[3]

[1] School of Mathematical and Physical Sciences, The University of Newcastle, Australia

[2] Department of Ecological Dynamics, Leibniz Institute for Zoo and Wildlife Research, Department of Evolutionary Ecology, Alfred-Kowalke-Str. 17, D-10315 Berlin, Germany

[3] CEFE, CNRS, Univ Montpellier, Univ Paul Valéry Montpellier 3, EPHE, IRD, Montpellier, France

Word count: 6,060

*ian.renner@newcastle.edu.au

## Summary

1. The increase in availability of species data sets means that approaches to species distribution modelling that incorporate multiple data sets are in greater demand. Recent methodological developments in this area have led to combined likelihood approaches, in which a log-likelihood comprised of the sum of the log-likelihood components of each data source is maximised. Often, these approaches make use of at least one presence-only data set and use the log-likelihood of an inhomogeneous Poisson point process model in the combined likelihood construction. While these advancements have been shown to improve predictive performance, they do not currently address challenges in presence-only modelling such as checking and correcting for violations of the independence assumption of a Poisson point process model or more general challenges in species distribution modelling such as overfitting.

2. In this paper, we present an extension of the combined likelihood framework which accommodates alternative presence-only likelihoods in the presence of spatial dependence as well as lasso-type penalties to account for potential overfitting. We compare the proposed approach combined penalised likelihood approach to the standard combined likelihood approach via simulation and apply the method to modelling the distribution of the Eurasian lynx in the Jura Mountains in eastern France.

3. The simulations show that the proposed combined penalised likelihood approach outperforms the standard approach when spatial dependence is present in the data. The lynx analysis shows that the predicted maps vary significantly with the different implementations of the proposed approach.

4. This work highlights the benefits of careful consideration of the presence-only components of the combined likelihood formulation, and allows greater flexibility and ability to accommodate real datasets.

Keywords: area-interaction models; diagnostic tools; lasso; occupancy models; point process models; presence-only data

# 1 Introduction

Species distribution models (SDMs), in which the distributions of species are modelled as a function of environmental predictors, rely on information about where a species has been observed (Guisan *et al.*, 2017). Different SDM methods have been developed over the past few decades to accommodate the different protocols by which this species information is collected. For example, logistic regression and its extensions are often used when species detections and non-detections are recorded at a set of systematically designed locations (known as "presence-absence" data), while point process models (PPMs, see Renner *et al.* (2015) for an overview) have emerged as a unifying framework for fitting SDMs informed by "presence-only" data, in which only information about species presence locations are available. Statistically, these methods are often fitted by maximising a corresponding likelihood expression, and the parameter estimates which maximise the likelihood may be used to produce maps of relative habitat suitability, reported as a habitat suitability index (Hirzel *et al.*, 2002), probability of species presence (Phillips *et al.*, 2006), or intensity of locations per unit area (Warton & Shepherd, 2010) depending on the method.

Increasingly, species data are available from multiple sources and types. Many papers have advocated for fitting models to a combination of the available data types, illustrating benefits in model performance (Miller *et al.*, 2019). Dorazio (2014) illustrated that adding a small amount of systematically-collected presence-absence data to available presence-only data significantly improves predictive performance. Fithian *et al.* (2015) showed that fitting a combined presence-only and presence-absence model to multiple species leverages the information of more abundant species to improve predictive performance for less prevalent species and allows sampling bias inherent in presence-only data to be estimated and corrected. These models are fitted by maximising a combined log-likelihood expression which is the sum of the log-likelihoods of the presence-only and presence-absence components:

$$\ell(\boldsymbol{\alpha}, \boldsymbol{\beta}; \mathbf{s}_{\mathrm{PO}}, \mathbf{y}_{\mathrm{PA}}) = \ell_{\mathrm{PO}}(\boldsymbol{\alpha}_{\mathrm{PO}}, \boldsymbol{\beta}; \mathbf{s}_{\mathrm{PO}}) + \ell_{\mathrm{PA}}(\boldsymbol{\alpha}_{\mathrm{PA}}, \boldsymbol{\beta}; \mathbf{y}_{\mathrm{PA}}).$$

Here, $\mathbf{s}_{\mathrm{PO}}$ contains the locations of a presence-only data source, while $\mathbf{y}_{\mathrm{PA}}$ contains a vector of presence-absence detections and non-detections at a set of pre-selected sites. The biases unique to the presence-only and presence-absence data sets are parameterised by $\boldsymbol{\alpha}_{\mathrm{PO}}$ and $\boldsymbol{\alpha}_{\mathrm{PA}}$, respectively, and collectively contained in the vector $\boldsymbol{\alpha}$. The key advancement of the combined likelihood approach is that the environmental response, parameterised by

$\boldsymbol{\beta}$, is informed by both the presence-only and presence-absence data.

Such an approach implicitly assumes that the data sets are statistically independent, which allows for the combined log-likelihood to be expressed as a sum of the single-source log-likelihoods.

Other combinations may be done in similar fashion. For example, Koshkina *et al.* (2017) considered a combination of presence-only data with site-occupancy data, and Pacifici *et al.* (2017) developed a multivariate conditional autoregressive model to account for spatial autocorrelation in occurrence and detection error.

While these papers clearly advance the practice of fitting SDMs in important ways, they do not address some common challenges that arise in real datasets. For example, they all consider an inhomogeneous Poisson point process model (IPPPM) for the presence-only data in the combination. In many real data sets, however, the implicit assumption that the point locations are independently distributed conditional on the environment is not met. Residual clustering or repulsion of the point locations not accounted for with an IPPPM due to the observation process, unconsidered environmental covariates, or biological factors would hence render the IPPPM inappropriate. Furthermore, none of the current literature in combined likelihood approaches includes ways to account for possible overfitting that results from including too many covariates in the model.

However, advances in SDM literature provide solutions to these common problems. Diagnostic tools such as the inhomogeneous $K$ function (Baddeley & Turner, 2000) and its simulation envelope (Diggle, 2003) can be used to determine departures from the independence assumption, and a wide number of alternative PPMs which account for spatial dependence may be included in the likelihood combination instead. Furthermore, penalised regression techniques such as the lasso penalty (Tibshirani, 1996) and its extension the adaptive lasso (Zou, 2006) may be used as a way to perform variable selection. Lasso regularisation has been shown to boost predictive performance of SDMs and has been applied to IPPPMs (Renner & Warton, 2013) and occupancy models (Hutchinson *et al.*, 2015).

In this paper, we present a penalised combined likelihood model in a way that it is more suitable for real data sets. In particular, we accommodate alternative forms of presence-only models to account for spatial dependence and affix a penalty on model complexity to address overfitting. In Section 2, we present the penalised combined likelihood formulation. In Section 3, we illustrate via simulations the improvements that this formulation provides and apply the proposed formulation to analyse the distribution of the Eurasian lynx (*lynx lynx*) in the Jura Mountains in eastern France. Finally, we present a discussion and further avenues for research in this area in Section 4.

# 2 Materials and Methods

## 2.1 Combined Penalised Likelihood Formulation

We define the weighted, combined penalised log-likelihood as follows

$$\ell(\boldsymbol{\alpha}, \boldsymbol{\beta}; \mathbf{y}) = \sum_{i=1}^{D} \ell_i(\boldsymbol{\alpha}_i, \boldsymbol{\beta}; \mathbf{y}_i) - p(\boldsymbol{\alpha}, \boldsymbol{\beta}). \quad \text{(eqn 1)}$$

Here, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_D)^\top$ is a $q$-dimensional vector that collects coefficients for the variables $\mathbf{Z}$ used to model bias for each of the $D$ components individually. The environmental response is measured by a set of variables $\mathbf{X}$ and is parametrised by $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, which is collectively informed by all $D$ components. The species data for all $D$ components is collected in a set $\mathbf{y}$, with each individual data source $\mathbf{y}_i$ determining the form of the component likelihood $\ell_i(\boldsymbol{\alpha}_i, \boldsymbol{\beta}; \mathbf{y}_i)$.

While many possibilities for the likelihood terms $\ell_i(\boldsymbol{\alpha}_i, \boldsymbol{\beta}; \mathbf{y}_i)$ are possible, we will focus on likelihood expressions for a PPM and for an occupancy model. For an IPPPM, we typically model the intensity of points $\mu(s)$ over a given study region $\mathcal{A}$ as a log-linear function of environmental variables $\mathbf{X}$ and bias terms $\mathbf{Z}$ and derive estimates $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}_{\text{PO}}$ of the associated parameters by maximising a log-likelihood expression given by (Cressie, 1992):

$$\ell_{\text{PO}}(\boldsymbol{\alpha}_{\text{PO}}, \boldsymbol{\beta}; \mathbf{s}_{\text{PO}}) = \sum_{s \in \mathbf{s}_{\text{PO}}} \ln \mu(s) - \int_{s \in \mathcal{A}} \mu(s) ds. \quad \text{(eqn 2)}$$

In the simple occupancy model we consider, each site $i$ is visited $J_i$ times. We collect the history of detections and non-detections for all $N$ sites in a matrix $\mathbf{y}_{\text{occ}}$. Assuming that the probability that site $i$ is occupied is given by $\psi_i$ and that the occupancy of the sites remains constant throughout the history of visits. We further assume the probability of detecting the species if present is $p_i$. Under these assumptions, we can then model the probability of observing $y_i$ detections at site $i$ as

$$P(Y_i = y_i) = \underbrace{\psi_i \binom{J_i}{y_i} p_i^{y_i} (1 - p_i)^{J_i - y_i}}_{\text{species present}} + \underbrace{I(y_i = 0)(1 - \psi_i),}_{\text{species absent}}$$

where $I(\cdot)$ is the indicator function.

We can relate the occupancy $\psi_i$ of site $i$ to an inhomogeneous Poisson intensity $\mu_i$ of the species distribution at over site $i$ as in Koshkina *et al.* (2017):

$$\psi_i = 1 - e^{-\mu_i \times A_i},$$

where $A_i$ is the area of site $i$.

As with the IPPPM, we can then model intensity as a log-linear function of environmental variables $\mathbf{X}$ and model detection probability $p_i$ as a function of some detection covariates $\mathbf{Z}$, such as the logit or complementary log-log function. We can then compute estimates $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}_{\text{occ}}$ of the associated model parameters by maximising the log-likelihood expression given by:

$$\ell_{\text{occ}}(\boldsymbol{\alpha}_{\text{occ}}, \boldsymbol{\beta}; \mathbf{y}_{\text{occ}}) = \ln \prod_{i=1}^{N} P(Y_i = y_i)$$

The term $p(\boldsymbol{\alpha}, \boldsymbol{\beta})$ in eqn 1 is a penalty on model complexity applied to both the environmental parameters $\boldsymbol{\beta}$ and the bias parameters $\boldsymbol{\alpha}$ to shrink these parameters toward zero in order to boost predictive performance. Here, we consider both the traditional lasso penalty (Tibshirani, 1996) and the adaptive lasso penalty (Zou, 2006). For the traditional lasso penalty,

$$p(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \lambda \left( \sum_{j=1}^{p} |\beta_j| + \sum_{k=1}^{q} |\alpha_k| \right),$$

where $\lambda$ is the size of the tuning parameter. For the adaptive lasso penalty,

$$p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma) = \lambda \left( \sum_{j=1}^{p} w_j |\beta_j| + w_{p+k} \sum_{k=1}^{q} |\alpha_k| \right),$$

where $\mathbf{w} = (w_1, \ldots, w_{p+q})^\top$ are weights for the adaptive lasso, typically of the form:

$$w_i = \begin{cases} \left| \hat{\beta}_i^{(\text{unp})} \right|^{-\gamma} & 1 \leq i \leq p \\ \left| \hat{\alpha}_{i-p}^{(\text{unp})} \right|^{-\gamma} & p+1 \leq i \leq p+q, \end{cases}$$

for $\gamma > 0$. Here, $\hat{\beta}_i^{(\text{unp})}$ is the unpenalised coefficient estimate corresponding to the $i^{\text{th}}$ environmental variable $\mathbf{x}_i$ and $\hat{\alpha}_i^{(\text{unp})}$ is the unpenalised coefficient estimate corresponding to the $i^{\text{th}}$ bias variable $\mathbf{z}_i$. The shape of the weights is determined by the parameter $\gamma$.

The data-driven choice of the adaptive weights $\mathbf{w}$ ensures that more important covariates (*i.e.* those with coefficient estimates further away from 0) will be penalised less. This construction also enables the adaptive lasso to achieve so-called oracle properties (Zou, 2006), which means that asymptotically, the correct subset of coefficients will be chosen and the procedure has optimal estimation rate.

We can use eqn 1 to represent the simpler framework introduced by Dorazio (2014) and Fithian *et al.* (2015) by setting $p(\boldsymbol{\alpha}, \boldsymbol{\beta}) = 0$. We further extend this framework by considering alternative choices for those component likelihoods $\ell_i(\boldsymbol{\alpha}_i, \boldsymbol{\beta}; \mathbf{y}_i)$ informed by presence-only data. Rather than consider only inhomogeneous Poisson point process models, we consider area-interaction models (Widom & Rowlinson, 1970; Baddeley & van Lieshout, 1995) when diagnostic analysis of these data sources identifies spatial dependence among the presence-only locations. Area-interaction models account for spatial dependence through a vector of computed point interactions, which measure the proportion of overlap among circles of a nominal radius around the observed points. They can account for both clustering and repulsion of points – the model parameter $\eta$ characterises the nature of the spatial dependence, with values of $\eta$ less than 1 signalling point repulsion and values of $\eta$ greater than 1 signalling point clustering.

Because the likelihood expression of an area-interaction model is intractable, it is typically fitted via maximum pseudolikelihood (Besag, 1977):

$$\ell_{\mathrm{AI}}(\boldsymbol{\alpha}_{\mathrm{PO}}, \boldsymbol{\beta}, \eta; \mathbf{s}_{\mathrm{PO}}) = \sum_{s \in \mathbf{s}_{\mathrm{PO}}} \ln \mu(s; \mathbf{s}_{\mathrm{PO}}) - \int_{s \in \mathcal{A}} \mu(s; \mathbf{s}_{\mathrm{PO}}) ds.$$

This log-pseudolikelihood expression appears the same as eqn 2, with the exception that the intensity $\mu(s)$ is replaced by conditional intensity $\mu(s; \mathbf{s}_{\mathrm{PO}})$ (Papangelou, 1974), reflecting the fact that for the area-interaction model, intensity at a location $s$ is conditional on the other points in the pattern $\mathbf{s}_{\mathrm{PO}}$.

## 2.2 Implementation in R

To fit models with the combined penalised log-likelihood in eqn 1, we have developed a set of functions in `R` inspired by the `optim` function and `ppmlasso` package (Renner & Warton, 2013). The main function `comb_lasso` takes an input a list of species data, associated environmental data, and formulae for the environmental trend and bias trends for each component, along with details such as type of presence-only likelihoods to use, the type of penalty, the number of models to fit, and the tuning parameter criterion.

The function applies the coordinate descent algorithm of Osborne *et al.* (2000). This requires the derivatives of the component likelihoods (also known as "score equations") to be computed. Analytical score equations are supplied directly to the `optim` function, which serves as the machinery of the optimisation. A tutorial illustrating use of this code for the simulations as performed in Section 3.1 as well as some functions written to plot intensity maps and features of the lasso penalisation is provided in the supplementary material.

# 3 Results

## 3.1 Simulations

To investigate the benefits of the proposed penalised combined likelihood formulation, we used the `rpoispp` function in `spatstat` (Baddeley & Turner, 2005) to generate a large inhomogeneous Poisson pattern $\mathbf{s}_{\text{true}}$ of roughly 10,000 points on a 30 $\times$ 30-unit square window from an intensity pattern defined by linear and quadratic terms of four generated variables (hence eight meaningful covariates $\mathbf{x}_1, \ldots, \mathbf{x}_8$).

From this pattern, we generated two biased presence-only subsamples $\mathbf{s}_1$ and $\mathbf{s}_2$. The first presence-only subsample $\mathbf{s}_1$ was biased by $z_1$, the distance to a simulated road network, and the other $\mathbf{s}_2$ by $z_2$, the distance to a simulated categorical covariate. We varied the size of the subsamples such that each pattern had either 50 or 200 points. We also varied the strength of the clustering of the presence-only subsamples by setting the coefficient of the interaction term $\beta_{\text{interact}}$. Here, the patterns either exhibit no clustering ($\beta_{\text{interact}} = 0$), moderate clustering ($\beta_{\text{interact}} = 0.5$) or strong clustering ($\beta_{\text{interact}} = 1$). To sample the points in $\mathbf{s}_1$, we proceed as follows:

1. Initialise the set of sampled points $\mathbf{s}_1 = \emptyset$ and the point interactions to be a vector of 0s

2. Compute the biased conditional intensity at every point in $\mathbf{s}_{\text{true}}$ using $\mathbf{x}_1, \ldots, \mathbf{x}_8$, the bias covariate $z_1$, and the current vector of point interactions

3. Set the conditional intensity for any point already selected in $\mathbf{s}_1$ to 0 to ensure these points are not resampled

4. Randomly select a point from $\mathbf{s}_{\text{true}}$ with sampling probabilities proportional to the conditional intensities and add the selected point to $\mathbf{s}_1$

5. Update the vector of point interactions for all points in $\mathbf{s}_{\text{true}}$ using the `evalInteraction` function in `spatstat`

6. Repeat steps 2-5 until we have sampled the desired number of points

We sample $\mathbf{s}_2$ in a similar manner, using $z_2$ instead of $z_1$ to create the bias.

Because the true pattern $\mathbf{s}_{\text{true}}$ is Poisson, this simulation setup emulates a scenario in which the clustering of the observed point patterns is an artefact of the observation process – this can happen if, for example, records are publicly available and enthusiasts for the species report further observations near the publicly available locations (Johnston *et al.*, 2019).

We also generated a history $\mathbf{y}_{\text{occ}}$ of detections and non-detections from 5 visits to each of 100 sites centred along a regular grid in the 30 $\times$ 30-unit observation window to emulate a data set for which we could consider occupancy modelling. The species was considered present at a site if the closest point in the pattern $\mathbf{s}_{\text{true}}$ was within a distance of 0.25 units of the centre of the site. The history of detections and non-detections at each site where the species was considered present was randomly generated according to detection probabilities defined by the inverse of the cloglog function evaluated at a generated detection covariate $\mathbf{z}_3$.

Finally, we generated 8 dummy covariates $\mathbf{d}_1, \ldots, \mathbf{d}_8$ to include in fitted models that were meaningless in describing the true species distribution. We did this to reflect the fact that in real applications, we may not know which among a suite of candidate variables truly determine the species distribution. We ensured that the maximum absolute correlation among all pairs of variables was smaller than 0.5.

After generating the species data, we fit a number of models, using as input environmental covariates the 8 meaningful covariates $\mathbf{x}_1, \ldots, \mathbf{x}_8$ as well as 8 dummy covariates $\mathbf{d}_1, \ldots, \mathbf{d}_8$ and using as bias covariates $\mathbf{z}_1$, $\mathbf{z}_2$, and $\mathbf{z}_3$. For each of seven different combinations of input data and presence-only likelihood, we fit a model without any penalty, with a lasso penalty, and with an adaptive lasso penalty. For the models fitted with either a lasso or an adaptive lasso penalty, we fit regularisation paths of 1000 models, increasing the penalty from 0 to the smallest penalty $\lambda_{\max}$ that would shrink all coefficients to 0, thus covering the entire scope of possible model sizes. The model which minimised BIC was chosen among the 1000 fitted models. We considered as species data using a combination of all three of $\mathbf{s}_1$, $\mathbf{s}_2$, and $\mathbf{y}_{\text{occ}}$ as well as $\mathbf{s}_1$, $\mathbf{s}_2$, and $\mathbf{y}_{\text{occ}}$ individually. For the models with presence-only species inputs, we also varied whether these were modelled using an IPPPM or an area-interaction model. This led to a total of 21 models being fitted, summarised in Table 1.

| Model | Species Data | Presence-only likelihood | Penalty |
|---|---|---|---|
| 1 | $\mathbf{s}_1$, $\mathbf{s}_2$, and $\mathbf{y}_{\text{occ}}$ | IPPPM | None |
| 2 | $\mathbf{s}_1$, $\mathbf{s}_2$, and $\mathbf{y}_{\text{occ}}$ | IPPPM | Lasso |
| 3 | $\mathbf{s}_1$, $\mathbf{s}_2$, and $\mathbf{y}_{\text{occ}}$ | IPPPM | Adaptive Lasso |
| 4 | $\mathbf{s}_1$, $\mathbf{s}_2$, and $\mathbf{y}_{\text{occ}}$ | Area-interaction | None |
| 5 | $\mathbf{s}_1$, $\mathbf{s}_2$, and $\mathbf{y}_{\text{occ}}$ | Area-interaction | Lasso |
| 6 | $\mathbf{s}_1$, $\mathbf{s}_2$, and $\mathbf{y}_{\text{occ}}$ | Area-interaction | Adaptive Lasso |
| 7 | $\mathbf{s}_1$ only | IPPPM | None |
| 8 | $\mathbf{s}_1$ only | IPPPM | Lasso |
| 9 | $\mathbf{s}_1$ only | IPPPM | Adaptive Lasso |
| 10 | $\mathbf{s}_1$ only | Area-interaction | None |
| 11 | $\mathbf{s}_1$ only | Area-interaction | Lasso |
| 12 | $\mathbf{s}_1$ only | Area-interaction | Adaptive Lasso |
| 13 | $\mathbf{s}_2$ only | IPPPM | None |
| 14 | $\mathbf{s}_2$ only | IPPPM | Lasso |
| 15 | $\mathbf{s}_2$ only | IPPPM | Adaptive Lasso |
| 16 | $\mathbf{s}_2$ only | Area-interaction | None |
| 17 | $\mathbf{s}_2$ only | Area-interaction | Lasso |
| 18 | $\mathbf{s}_2$ only | Area-interaction | Adaptive Lasso |
| 19 | $\mathbf{y}_{\text{occ}}$ only | – | None |
| 20 | $\mathbf{y}_{\text{occ}}$ only | – | Lasso |
| 21 | $\mathbf{y}_{\text{occ}}$ only | – | Adaptive Lasso |

Table 1: Models fitted in each simulation. Models 1-6 were fitted using the proposed penalised combination framework, whereas models 7-21 were fitted using only one source of data. The models also varied based on the likelihood expression for any presence-only components and the type of penalty used, if any.

To evaluate performance, we compared the integrated mean squared error of the true intensity surface with rescaled fitted intensity surfaces of the 21 models. The fitted intensity surfaces were rescaled to have the same mean intensity as the true intensity surface to ensure that fair comparisons are made as models using different species data sources will have varying intercepts to reflect the estimated abundance of the points.

We performed 1,000 simulations of the data sets for each of the six combinations of presence-only data set size and clustering strength and the resultant model fits on 512GB nodes powered by 3.0 GHz Intel Xeon Gold (E5-6154) processor from the University of Newcastle's High Performance Computing cluster. The 6,000 simulation tasks took approximately 12,000 hours.

Figure 1 shows boxplots of the calculated integrated mean squared errors from the simulations. From these results, we can draw the following conclusions. First, the combined likelihood approaches tend to outperform the models that rely on a single source of data, confirming previous results by Dorazio (2014) and Fithian *et al.* (2015). This effect is mitigated somewhat in the presence of clustering for the combined model which uses the incorrect IPPPM likelihood, however.

Second, penalisation via the lasso or adaptive lasso improves model performance, and this benefit is stronger with smaller data sets. This can be seen by comparing the same models with different sample sizes (e.g. Models 7-9 when $N = 50$ to Models 7-9 when $N = 200$ regardless of the clustering), as well as noting that the benefits of penalisation are less pronounced for Models 1-3 and 4-6 which use all three data sets. This is an expected conclusion given the danger of overfitting is greater with fewer observations. Models penalised with the adaptive lasso tend to outperform models penalised with the lasso for the presence-only data sets. Although the benefits of penalisation are negligible with large data sets, fitting models with a penalty does not hurt the performance.

Finally, the performance benefits of selecting the correct presence-only area-interaction likelihood in the presence of clustering tend to be greater with larger data sets. In the first column of Figure 1, there is no clustering, such that models that use the IPPPM likelihood for the presence-only components are appropriate, whereas in the second and third columns, the models which use the area-interaction likelihood for the presence-only components are appropriate. From the middle and right columns, we see that differences in performance with the correct area-interaction likelihood for the presence-only components are more pronounced as the presence-only sample size increases and as the strength of the clustering becomes more pronounced. Curiously, the left column shows that the benefits of the correct IPPPM likelihood are less pronounced with greater presence-only sample sizes for Models 7-18 which only use a presence-only data source, and the combined models appear to perform better with an area-interaction term when $N = 200$. We suspect that the area-interaction term is not harming the performance because it is partially explained by the effect of the bias used to generate the presence-only patterns – that is, biasing the observations to be near roads or near level 1 of the categorical covariate induces some clustering.

In summary, it appears that the proposed combined penalised likelihood framework provides the best performance.

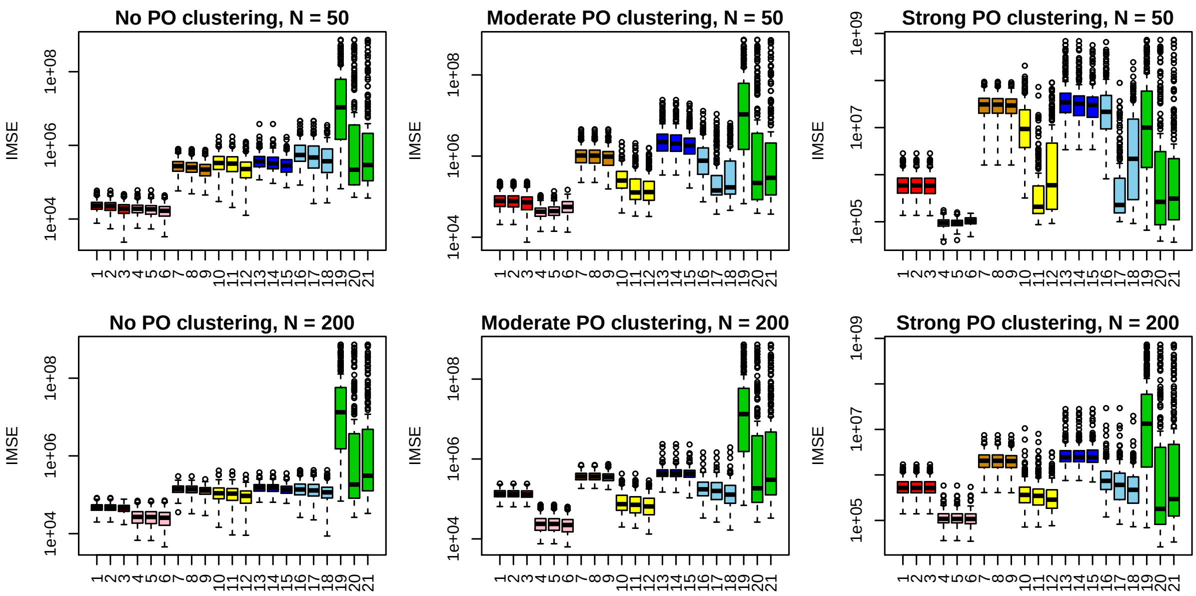


Figure 1: Boxpots of integrated mean squared error for the 21 models described in Table 1.

## 3.2 Analysis of Eurasian lynx distribution in the Jura Mountains

We now demonstrate the use of the combined penalised likelihood approach to analyse the distribution of the Eurasian lynx in the Jura Mountains in eastern France.

Lynx went extinct in France in at the end of the 19th century due to habitat degradation, human persecution and decrease in prey availability (Vandel & Stahl, 2005). The species was reintroduced in Switzerland in the 1970s (Breitenmoser *et al.*, 1998), then re-colonised France through the Jura mountains in the 1980s (Vandel & Stahl, 2005). The species is listed as endangered under the 2017 IUCN Red list and is of conservation concern in France due to habitat fragmentation, poaching and collisions with vehicles. The Jura holds the bulk of the French lynx population.

We have three sources of lynx data in the Jura Mountains: a presence-only data set consisting of 440 opportunistic sightings in the wild from 2009-2011 (denoted $\mathbf{s}_{\mathrm{w}}$), another presence-only data set consisting of 240 reported interferences of lynx with domestic livestock in 2009-2011 (denoted $\mathbf{s}_{\mathrm{d}}$), and pictures of lynx taken from cameras set up in

73 locations $\mathbf{s}_{\mathrm{c}}$ in the Jura Mountains in 2012. Lynx presence-only data were made of presence signs sampled all year long thanks to a network of professional and non-professional observers. Every observer is trained during a 3-day teaching course led by the French National Game and Wildlife Agency (ONCFS) to document signs of the species' presence (Duchamp *et al.*, 2012). Presence signs went through a standardised control process to prevent misidentification (Duchamp *et al.*, 2012). The camera data has daily reportings of the lynx across a total of 77 days. Due to this, we can consider the picture history of lynx at the camera locations in an occupancy modelling framework (Blanc *et al.*, 2014). In particular, we split the 77-day period into seven 11-day periods, such that the site history $\mathbf{y}_{\mathrm{c}}$ comprises seven detections and non-detections at each site in $\mathbf{s}_{\mathrm{c}}$ over each 11-day period.

Figure 2 shows the locations of the sightings in both presence-only data sets as well as the locations of the cameras. Both presence-only data sources appear to have different distributions, reflecting different sampling biases. There are more wild sightings in the northeast of the Jura Mountains, and more domestic interferences toward the southwest. Additionally, there appear to be some tight clusters within both data sets, with several records very close to each other.

To model the lynx distribution, we consider altitude, percentage of forest cover, distance to the nearest water source, and human population density as environmental variables. We model sampling bias in the wild records $\mathbf{s}_{\mathrm{w}}$ with distance to the nearest main road and distance to the nearest train line, and sampling bias in the domestic records $\mathbf{s}_{\mathrm{d}}$ with distance to the nearest farm and percentage of agricultural land. Finally, we model detection probability for the camera data with distance to the nearest urban area. We established this set of potential candidate environmental and detection variables based on previously studied species habitat preferences and detectability (Bouyer *et al.*, 2015). The Corine Land Cover land use repository from 2012 (Büttner *et al.*, 2014) supplies a map of land coverage including urban areas, water areas, forest areas, farm areas, and agricultural areas that was used to generate the percentage of forest areas and agricultural areas over 1 km $\times$ 1 km cells as well as distances to the nearest urban area, water source, and farm. Altitude was averaged over 1km $\times$ 1km cells from data available in the `raster` package in R, while human population density was averaged over 1 km $\times$ 1km cells taken from version 4 of the Gridded Population of the World data repository (Center for International Earth Science Information Network (CIESIN) – Columbia University, 2016). Distances from the nearest main road and railway were computed from shapefiles from Version 151

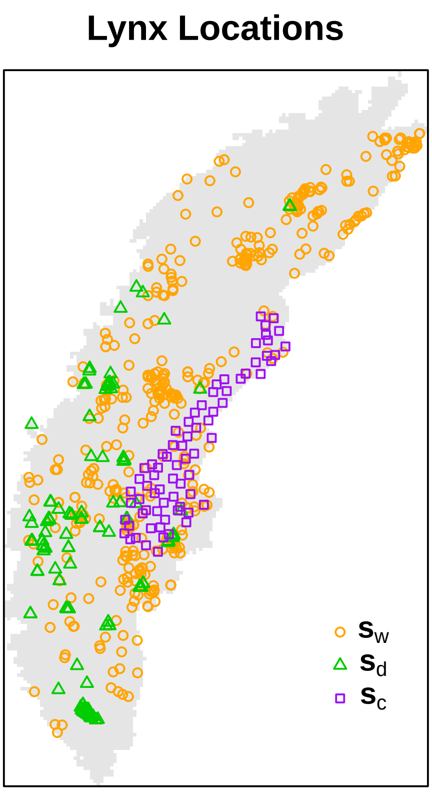


Figure 2: Locations of the lynx data in the Jura Mountains, including 440 observations in the wild $\mathbf{s}_{\mathrm{w}}$, 220 reports of domestic interference $\mathbf{s}_{\mathrm{d}}$, and 73 camera traps $\mathbf{s}_{\mathrm{c}}$.

of the ROUTE 500 database, accessible at http://professionnels.ign.fr/route500.

We fitted initial separate IPPPMs to the wild records $\mathbf{s}_{\mathrm{w}}$ and the domestic records $\mathbf{s}_{\mathrm{d}}$ using linear, quadratic, and interaction terms for the four environmental covariates, and linear terms for the bias covariates. From these models, we are able to assess whether the assumption of independence inherent to the IPPPMs is appropriate with simluation envelopes of the inhomogeneous $K$-function in `spatstat`, as shown in Figure 3. Both of the envelopes for the IPPPMs fitted to the wild model (left panel) and the domestic model (middle panel) demonstrate additional clustering as the observed inhomogeneous $K$-function values plotted in red fall above the simulation envelopes for small radii. This suggests that fitting an IPPPM is inappropriate for these data sets. The right panel shows a simulation envelope comparing clustering across the two data sources, where the intensities are estimated from area-interaction models, and as the observed values of the $K$-function fall within the envelope boundaries, this suggests that there is no clustering across the two data sets. This, in turn, suggests that the observed clustering within the wild and domestic data sets may be more likely attributable to the observation process than to some biological reality that induces clustering or a missed environmental covariate.

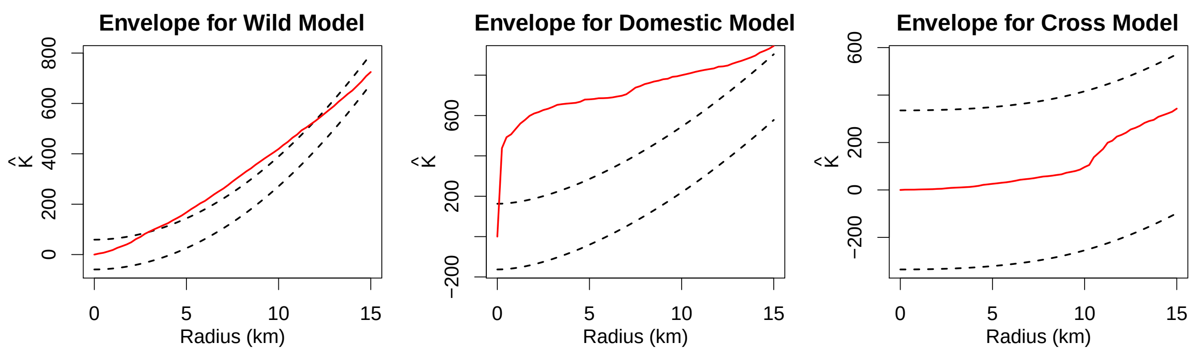


Figure 3: 95% simulation envelopes of the inhomogeneous $K$-function for the fitted IPPPM of the wild records (left), the fitted IPPPM of the domestic records (middle), and across the two fitted IPPPMs (right).

Consequently, we fit combined likelihood models using both the standard, unpenalised approach (analogous to Model 1 in Table 1) and the combined penalised likelihood formulation eqn 1 with a lasso penalty and area-interaction models for the presence-only data sources (analogous to Model 5 in Table 1). The radii chosen to capture the residual spatial patterning in the wild and domestic models are 2km and 5km, as chosen by the `profilepl` function in `spatstat`.

Figure 4 shows the bias-corrected fitted intensities from these two models. For the combined model which uses IPPPMs (left panel), the fitted intensity is corrected for the bias terms modelled for the presence-only components using the method of Warton *et al.* (2013). For the combined penalised model which uses area-interaction models (right panel), the fitted intensity is corrected for these same bias terms as well as the fitted point interactions – that is, we treat the interaction parameter $\eta$ as belonging to the set of bias parameters $\boldsymbol{\alpha}$. The fitted models show strikingly different patterns, with the model which uses area-interaction components highlighting much more of the Jura Mountains as preferred habitat of lynx than the model which uses IPPPMs. We do not have access to additional data with which to validate the performance of these models such as GPS data as in Gould *et al.* (2019), but the results of Section 3.1 suggest that the model which uses area-interaction components is likely to better reflect the true distribution of lynx.

The combined penalised model with the area-interaction components found the optimal lasso penalty was 0, resulting in a model which included all 18 covariates and both interaction terms.

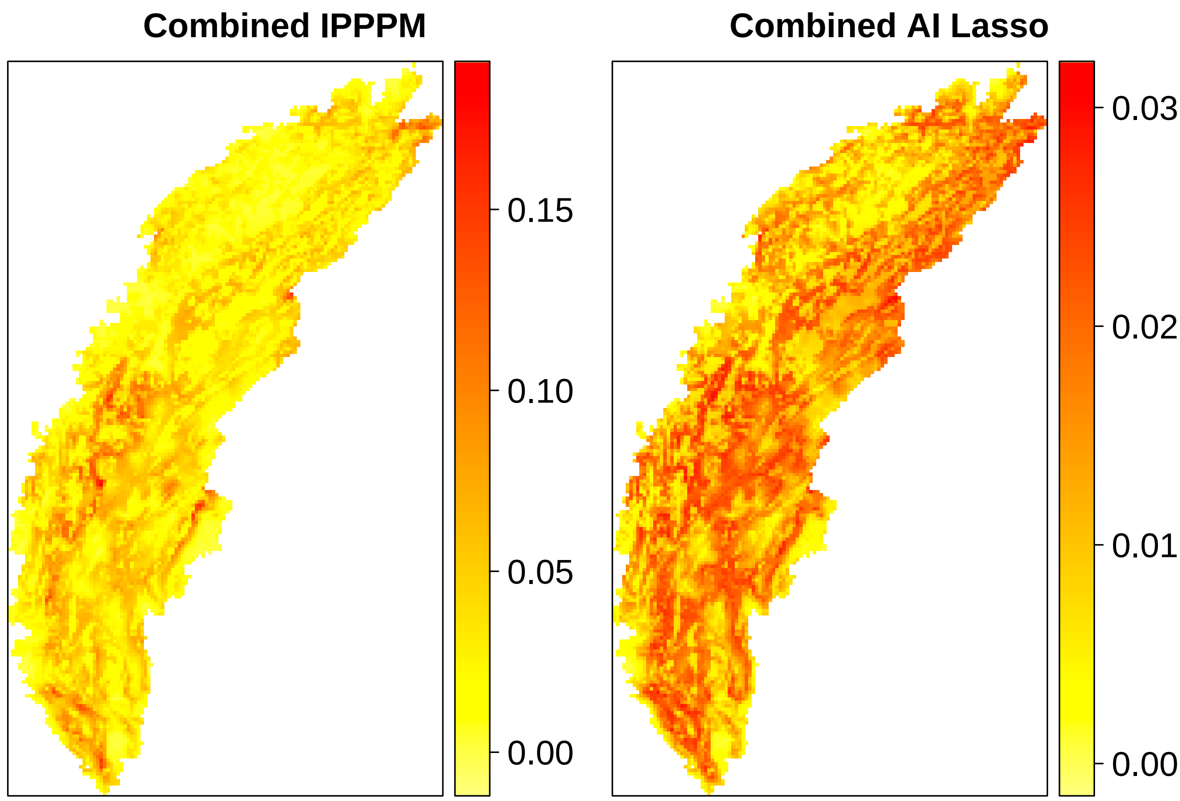


Figure 4: Fitted intensities using the combined likelihood formulation. Left: the model is fitted without any penalty and using inhomogeneous Poisson point process models for the presence-only data sources. Right: the model is fitted with a lasso penalty and using area-interaction models for the presence-only sources.

# 4 Discussion

The proposed combined penalised likelihood framework addresses some common problems that arise in real datasets. The flexibility to incorporate an area-interaction likelihood when there is spatial dependence in the presence-only data set and affix a penalty on model complexity enables improvements in predictive performance, as shown in Section 3.1.

## 4.1 Possible extensions

Despite these improvements, further advances are possible. Other penalty structures could be incorporated into the same framework. While the lasso and adaptive lasso showcased here show clear benefits in simulations, other penalised likelihood variants such as SCAD (Fan & Li, 2001) could lead to superior performance in some situations, and alternative methods to BIC of choosing the size of the penalty such as the Extended

Bayesian Information Criterion ("ERIC", Hui *et al.*, 2015) could likewise be used.

While we make use of the area-interaction likelihood in this paper, there is a large family of Gibbs PPMs (Cressie, 1992) which accommodate different sorts of spatial dependence that could be used. Our choice of the area-interaction model as the alternative is motivated by the fact that it accommodates interactions of all orders instead of just pairwise interactions and that it can be used to model both clustering and repulsion of points.

In both the simulations in Section 3.1 and the lynx data analysis in Section 3.2, we made the rather limiting assumption of a closed population and that sites are either always occupied or always unoccupied. Nonetheless, occupancy models which take into account changing site dynamics could be used (MacKenzie *et al.*, 2003). Similarly, we have ignored the temporal aspect of the lynx distribution in this paper, but there is a wide suite of tools to fit spatio-temporal models in order to capture distribution dynamics for both the aforementioned occupancy modelling component as well as presence-only components (Cressie & Wikle, 2015).

Further improvements could be made by incorporating source weights in situations in which the data sources vary in quality. Indeed, presence-only data sources may be more prone to errors in coordinate locations as well as correct species identification, as they often include records by amateur enthusiasts. The combined penalised likelihood framework could easily be extended to include weights for the various data sources by adding a vector of source weights $\mathbf{w} = (w_1, \ldots, w_D)^\top$ to the formulation in eqn 1:

$$\ell(\boldsymbol{\alpha}, \boldsymbol{\beta}; \mathbf{y}) = \sum_{i=1}^{D} w_i \ell_i(\boldsymbol{\alpha}_i, \boldsymbol{\beta}; \mathbf{y}_i) - p(\boldsymbol{\alpha}, \boldsymbol{\beta}). \quad \text{(eqn 3)}$$

## 4.2 Accounting for dependence within and among data sources

One possible strategy to incorporate such weights in eqn 3 could be to compare performance of single source models on independent data and upweight the contribution of data sources that are shown to have good performance.

In the lynx data analysis in Section 3.2, we diagnosed spatial dependence within each of the presence-only data sources but found no spatial dependence across data sources. Tools such as the inhomoegenous $K$-envelope provide great insight into the underlying individual spatial processes that are observed. However, such diagnostic tools are not currently available for the combined likelihood models, and research in this area would be valuable as these models grow in popularity.

Another approach to constructing SDMs from multiple data sources could be to introduce a common latent spatial term $\xi(s)$, such as a Gaussian random field, which would account for spatial dependence among points in all of the data sources. The resulting likelihood expression would be:

$$\ell(\boldsymbol{\alpha}, \boldsymbol{\beta}; \mathbf{y}) = \sum_{i=1}^{D} \ell_i(\boldsymbol{\alpha}_i, \boldsymbol{\beta}; \mathbf{y}_i) + \xi(\mathbf{y}) - p(\boldsymbol{\alpha}, \boldsymbol{\beta}), \qquad \text{(eqn 4)}$$

where $\xi(\mathbf{y}) \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma})$. Models of this type are typically fitted in a Bayesian framework. We could reduce the dimension of $\xi$ through methods like fixed rank kriging or induce sparsity in $\boldsymbol{\Sigma}$ through lasso-type penalties such that the likelihood in eqn 4 could be fitted with software such as Template Model Builder (TMB, Kristensen *et al.*, 2016).

## 4.3 Conclusion and Perspectives

The development of statistical methods is often motivated by new challenges raised by novel types of data sets. While the current literature on combined likelihood approaches represents a significant recent advancement, advances in other areas can be lost if not carried over with such methodological developments. This paper attempts to build a bridge between this exciting new arena for species distribution modelling and the rich suite of tools available for species distribution modelling, particularly that for presence-only data. Our hope is that other such bridges continue to be built in this spirit.

# 5 Acknowledgements

We thank the staff from the French National Game and Wildlife Agency, the Forest National Agency and the Departmental Federation of Hunters of Jura department, who collected the photographs during the camera-trapping session. We also thank all the volunteers that are members of the Réseau Loup-Lynx that collect every year precious presence signs of lynx all over its distribution area. We thank the University of Newcastle through the Early to Mid-Career Researcher Visiting Fellowship and the Université Paul-Valéry Montpellier through the Professeurs en mobilité universitaire fund for funding visits for I.R. which helped facilitate the collaborations that resulted in this work.

# 6 Authors' Contributions

I.R. and O.G. conceived the concept of the paper. I.R. developed the code to fit the models. J.L. sourced the species coordinates and covariates for the lynx analysis. I.R. and O.G. wrote the manuscript. I.R. and J.L. developed the tutorial in the supplementary information. All authors were involved in editing drafts of the manuscript.

# References

Baddeley, A. & Turner, R. (2000) Practical maximum pseudolikelihood for spatial point patterns: (with discussion). *Australian & New Zealand Journal of Statistics*, **42**, 283–322.

Baddeley, A. & Turner, R. (2005) Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software*, **12**, 1–42.

Baddeley, A.J. & van Lieshout, M.N.M. (1995) Area-interaction point processes. *Annals of the Institute of Statistical Mathematics*, **47**, 601–619.

Besag, J. (1977) Some methods of statistical analysis for spatial data. *Bulletin of the International Statistical Institute*, **47**, 77–92.

Blanc, L., Marboutin, E., Gatti, S., Zimmermann, F. & Gimenez, O. (2014) Improving abundance estimation by combining capture–recapture and occupancy data: example with a large carnivore. *Journal of Applied Ecology*, **51**, 1733–1739.

Bouyer, Y., San Martin, G., Poncin, P., Beudels-Jamar, R.C., Odden, J. & Linnell, J.D. (2015) Eurasian lynx habitat selection in human-modified landscape in norway: Effects of different human habitat modifications and behavioral states. *Biological Conservation*, **191**, 291–299.

Breitenmoser, U., Breitenmoser-Würsten, C. & Capt, S. (1998) Re-introduction and present status of the lynx (lynx lynx) in switzerland. *Hystrix, the Italian Journal of Mammalogy*, **10**.

Büttner, G., Soukup, T. & Kosztra, B. (2014) Clc2012 addendum to clc2006 technical guidelines. *Final Draft, Copenhagen (EEA)*.

Center for International Earth Science Information Network (CIESIN) – Columbia University (2016) Gridded population of the world, version 4 (gpwv4): population density.

Cressie, N. (1992) *Statistics for spatial data*, volume 4. Wiley Online Library.

Cressie, N. & Wikle, C.K. (2015) *Statistics for spatio-temporal data*. John Wiley & Sons.

Diggle, P.J. (2003) *Statistical analysis of spatial point patterns*. Edward Arnold.

Dorazio, R.M. (2014) Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Global Ecology and Biogeography*, **23**, 1472–1484.

Duchamp, C., Boyer, J., Briaudet, P.E., Leonard, Y., Moris, P., Bataille, A., Dahier, T., Delacour, G., Millisher, G., Miquel, C. *et al.* (2012) A dual frame survey to assess time- and space-related changes of the colonizing wolf population in france. *Hystrix-Italian Journal of Mammalogy*, **23**, 14–28.

Fan, J. & Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, **96**, 1348–1360.

Fithian, W., Elith, J., Hastie, T. & Keith, D.A. (2015) Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, **6**, 424–438.

Gould, M.J., Gould, W.R., Cain III, J.W. & Roemer, G.W. (2019) Validating the performance of occupancy models for estimating habitat use and predicting the distribution of highly-mobile species: A case study using the american black bear. *Biological Conservation*, **234**, 28–36.

Guisan, A., Thuiller, W. & Zimmermann, N.E. (2017) *Habitat suitability and distribution models: with applications in R*. Cambridge University Press.

Hirzel, A.H., Hausser, J., Chessel, D. & Perrin, N. (2002) Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? *Ecology*, **83**, 2027–2036.

Hui, F.K., Warton, D.I. & Foster, S.D. (2015) Tuning parameter selection for the adaptive lasso using ERIC. *Journal of the American Statistical Association*, **110**, 262–269.

Hutchinson, R.A., Valente, J.J., Emerson, S.C., Betts, M.G. & Dietterich, T.G. (2015) Penalized likelihood methods improve parameter estimates in occupancy models. *Methods in Ecology and Evolution*, **6**, 949–959.

Johnston, A., Hochachka, W., Strimas-Mackey, M., Gutierrez, V.R., Robinson, O., Miller, E., Auer, T., Kelling, S. & Fink, D. (2019) Best practices for making reliable inferences from citizen science data: case study using ebird to estimate species distributions. *bioRxiv*, p. 574392.

Koshkina, V., Wang, Y., Gordon, A., Dorazio, R.M., White, M. & Stone, L. (2017) Integrated species distribution models: combining presence-background data and site-occupany data with imperfect detection. *Methods in Ecology and Evolution*, **8**, 420–430.

Kristensen, K., Nielsen, A., Berg, C.W., Skaug, H. & Bell, B.M. (2016) TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, **70**, 1–21.

MacKenzie, D.I., Nichols, J.D., Hines, J.E., Knutson, M.G. & Franklin, A.B. (2003) Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. *Ecology*, **84**, 2200–2207.

Miller, D.A., Pacifici, K., Sanderlin, J.S. & Reich, B.J. (2019) The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution*, **10**, 22–37.

Osborne, M.R., Presnell, B. & Turlach, B.A. (2000) On the lasso and its dual. *Journal of Computational and Graphical Statistics*, **9**, 319–337.

Pacifici, K., Reich, B.J., Miller, D.A., Gardner, B., Stauffer, G., Singh, S., McKerrow, A. & Collazo, J.A. (2017) Integrating multiple data sources in species distribution modeling: A framework for data fusion. *Ecology*, **98**, 840–850.

Papangelou, F. (1974) The conditional intensity of general point processes and an application to line processes. *Probability Theory and Related Fields*, **28**, 207–226.

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

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

Renner, I.W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S.J., Popovic, G. & Warton, D.I. (2015) Point process models for presence-only analysis. *Methods in Ecology and Evolution*, **6**, 366–379.

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

Vandel, J.M. & Stahl, P. (2005) Distribution trend of the eurasian lynx lynx lynx populations in france. *Mammalia mamm*, **69**, 145–158.

Warton, D.I., Renner, I.W. & Ramp, D. (2013) Model-based control of observer bias for the analysis of presence-only data in ecology. *PloS one*, **8**, e79168.

Warton, D.I. & Shepherd, L.C. (2010) Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. *Annals of Applied Statistics*, **4**, 1383–1402.

Widom, B. & Rowlinson, J.S. (1970) New model for the study of liquid–vapor phase transitions. *The Journal of Chemical Physics*, **52**, 1670–1684.

Zou, H. (2006) The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418–1429.