Theoretical Ecology (2019) 12:73–88
https://doi.org/10.1007/s12080-018-0389-9

ORIGINAL PAPER

Understanding the connections between species distribution models for presence-background data

Yan Wang[1] · Lewi Stone[1,2]

Received: 10 January 2018 / Accepted: 26 July 2018 / Published online: 1 September 2018

Abstract
Models for accurately predicting species distributions have become essential tools for many ecological and conservation problems. For many species, presence-background (PB) data is the most commonly available type of spatial data. A number of important methods have been proposed to model PB data, and there have been debates on the connection between these seemingly disparate methods. The paper studies the close relationship between the LI (Lancaster and Imbens), LK (Lele and Keim), scaled binomial (SB), expectation-maximization (EM), partial likelihood based Lele method, MAXENT, and the point process models. We reveal that all these methods are the same in their ability to estimate the relative probability (or intensity) of presence from PB data, and the absolute probability of presence, when extra information of the species' prevalence is known. A new unified constrained LK (CLK) method is also proposed as a generalization of the better known existing approaches, with less theory involved and greater ease of implementation.

Keywords Likelihood · Link function · Point process model · Presence-background · Prevalence · Probability of presence · Species distribution model

Introduction

Ecologists employ species distribution models (SDMs) to assist in mapping the spatial distribution of a species over its geographic range, despite there being only limited observational data available. SDMs are typically used to study three different types of spatial data, which are referred to as presence-background, presence-absence, and occupancy-detection data (Guillera-Arroita et al. 2015). Presence-background (PB) data contains a list of "presences," or locations where individuals have been observed, but typically having no information about absences—sites where species have not been observed. PB data is often plentifully available from so-called "opportunistic surveys" and can be obtained from museum and herbarium collections, historical database records (Pearce and Boyce 2006), and is now becoming increasingly available via online repositories such as the Global Biodiversity Information Facility (GBIF; http://www.gbif.org). Given such data, one of the key goals of SDMs is to estimate the site-specific probability of presence in the study region. Since SDMs assume that covariates are ultimately responsible for determining species' spatial distributions, SDMs model how covariates affect the local probability of presence. To help estimate the site-specific presence probabilities, SDMs make use of extra background sites at which the information of the environmental covariates (e.g., temperature, altitude, etc.) are available.

In contrast, presence-absence (PA) data provides information on whether a species was detected or not at all sampling sites of the study area. This contrasts with PB data where the species absence status is unknown. Occupancy-detection data is similar to PA data, but requires data to be collected in repeat visits at each study site so that the detection process can be modeled. Both PA and occupancy-detection data are difficult to obtain as they require intensive and extensive survey efforts to get PA information at all surveyed sites. There have been many methods developed

✉ Yan Wang
yan.wang@rmit.edu.au

[1] Discipline of Mathematical Science, School of Science, RMIT University, Melbourne, VIC, Australia

[2] Biomathematics Unit, Department of Zoology, Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel

for modeling PA and occupancy-detection data, and Guillera-Arroita et al. (2015) provide a comprehensive literature review. In recent years, joint species distribution models (JSDMs) have emerged as an additional feasible method for explicitly incorporating environmental variables and biotic interactions simultaneously in modeling multiple species (e.g., Warton et al. (2015) and Ovaskainen et al. (2016)). These models have been used with both PA and occupancy-detection data. In this paper, we will focus on the methodologies in modeling PB data, which was used in 50% of the papers surveyed in Guillera-Arroita et al. (2015).

A number of methods exist for modeling species distributions based on PB data, including statistical regression methods (e.g., regression methods discussed in Phillips and Elith (2013) and generalized linear and additive models (Guisan et al. 2002)), machine learning methods (e.g., MAXENT (Phillips et al. 2006; Phillips and Dudík 2008), and boosted regression tree (Elith et al. 2008)) and spatial point process models (PPMs) (Warton and Shepherd 2010; Renner and Warton 2013). From a theoretical perspective, modeling PB data has major challenges some of which are insurmountable. Our paper is intended to explore these challenges from a statistical perspective. This paper deals with the key statistical species distribution models suitable for modeling PB data, including the regression-based methods of SC by Steinberg and Cardell (1992), LI by Lancaster and Imbens (1996), LK by Lele and Keim (2006) and Royle et al. (2012), expectation-maximization (EM) of Ward et al. (2009), scaled binomial loss model (SB) of Phillips and Elith (2011), and the partial likelihood-based Lele method (Lele 2009), as well as the PPMs and the widely applied MAXENT method. Other machine learning approaches do not in general adopt a conventional statistical approach (Elith et al. 2008), and therefore fall outside the scope of this paper.

More specifically, the regression methods in SDMs estimate the probability $p(y = 1|x)$ that a species of interest is present, $y = 1$ (versus absent, $y = 0$) at a particular site, conditional on environmental covariates x at that site. The probability $p(y = 1|x)$ is also referred to as the resource selection probability function (RSPF) (Keating and Cherry 2004; Lele 2009). A common practice is to assume a parametric structure for modeling $p(y = 1|x)$, for example, the widely used logit form, $\log \frac{p(y=1|x)}{1-p(y=1|x)} = \eta(x^T\beta)$. Here, $\eta(x)$ can be a linear or a nonlinear function of x, and a logit-linear specification is as follows:

$$\log \frac{p(y = 1|x)}{1 - p(y = 1|x)} = \beta_0 + \sum_i \beta_i x_i. \quad (1)$$

The goal of the SDM is to estimate all of the parameters β_i.

The methods discussed in the paper have been developed independently using different definitions and framework to model species distributions. The key goal of the paper is to show the equivalence between these seemingly disparate models. This is one of the major contributions of our manuscript. The connections are revealed initially by studying the close link between the LI and LK methods, which were among the first developed methods for analyzing PB data. It will be shown for the first time that the LK method is a numerical approximation of the LI method. Secondly, we examine the analogy between the PPM and the LK model, when the likelihood function of the PPM is approximated by its discrete counterpart. We also show the equivalence between the SB, EM, LI, and the Lele methods. These equivalences have not been noted previously in the literature. Along with other findings on relations in the field, such as those done by Baddeley et al. (2010), Warton and Shepherd (2010), Aarts et al. (2012), Fithian and Hastie (2013), and Renner and Warton (2013), we conclude that all these methods are essentially equivalent in their ability to estimate the relative probability of presence. Furthermore, we present a unified constrained LK (CLK) method, which bridges the gaps between these seemingly different approaches. Each of the methods discussed in the paper is shown to be a special case of the unified CLK method.

The relationship between LI and LK methods

Lancaster and Imbens (1996) proposed a contaminated case control study for representing PB data, in which the set of sites in the study area is divided into two subsets. Subset 1 consists of all those sites in the study area on which the species is present. Subset 0 comprises the whole set of sites in the study area, with no information made available regarding, which of these "background sites" the species is present or not. However, the relevant environmental covariates are known at all background sites.

LI defined a sequence of n Bernoulli trials with the probability h to choose between the presence (case) and background (contaminated controls) points. A binary indicator u was used to denote the stratum, with $u = 1$ if the observation was drawn from the presence, and $u = 0$ if it was drawn from the whole population. After the n Bernoulli trials, there are n_1 sites chosen with species presences, and n_0 background sites with unknown status. That is, we do not have knowledge as to whether any background point is a "presence" or an "absence." When analyzing PB data, the background points are usually taken as either a uniform sample or a regular grid with a large number of observations. The distribution of the environmental covariates $F(x)$ can be approximated by a discrete distribution with unknown probabilities α_l on $L + 1$ known points of support x_l (Lancaster and Imbens 1996). An empirical estimator of α_l

is the fraction of observations taking the value x_l in the background data, i.e., $\hat{\alpha}_l = n_l/n_0$.

From Bayes theorem, we can derive $p(x|y = 1) = \frac{p(y=1|x)f(x)}{\pi}$, where π is the proportion of sites with species' presence in the study region. Thus, $\pi = \int p(y = 1|x)dF(x)$, and $F(x)$ is the unknown probability distribution function for x. For PB data, π is generally unknown, since there is little or no information about the presence status of the background points.

It is worth mentioning that the density ratio approach, originally developed by machine learning researchers for covariate shift adaptation and outlier detection (Sugiyama et al. 2012), has a very similar framework as the above Bayes probability. The density ratio method has recently been adapted to ecological niche modeling (Drake and Richards 2017). In its estimation, the ratio $\frac{p(x|y=1)}{f(x)}$ is of particular interest, as this ratio is closely associated with the fundamental niche. In this paper, we are interested in whether the actual probability of presence $p(y = 1|x)$ can be estimated accurately from the PB data without information of π, one of the controversial arguments in the statistical modeling of the species distributions (Lele and Keim 2006; Ward et al. 2009; Royle et al. 2012; Phillips and Elith 2013; Hastie and Fithian 2013; Solymos and Lele 2016).

The joint distribution of stratum u and covariates x is: $g(x, u) = [p(x|y = 1)h]^u[f(x)(1 - h)]^{1-u}$ (Lancaster and Imbens 1996), which can be rewritten as $\left[\frac{p(y=1|x)f(x)h}{\pi}\right]^u [f(x)(1-h)]^{1-u}$. The full likelihood function for the contaminated sampling scheme based on the joint distribution of (x, u) is as follows:

$$\begin{aligned} L(\beta, h, \alpha, \pi) &= \prod_{i=1}^{n}\left[\frac{p(y_i=1|x_i,\beta)f(x_i)h}{\pi}\right]^{u_i}[f(x_i)(1-h)]^{1-u_i} \\ &= \prod_{i=1}^{n}\left[\frac{p(y_i=1|x_i,\beta)}{\pi}\right]^{u_i}\prod_{i=1}^{n}f(x_i)\prod_{i=1}^{n}[h^{u_i}(1-h)^{1-u_i}] \\ &= L_1(\beta,\pi) * L_2(\alpha) * L_3(h), \end{aligned} \tag{2}$$

where the total number of sample points is $n = n_0 + n_1$.

By splitting the full likelihood into three partial likelihoods in Eq. 2, the role of each likelihood becomes clear. The partial likelihood function $L_3(h)$, which is independent of other parts of the likelihood function, is used to estimate the unknown sampling proportion with a binomial type of estimator $\hat{h} = n_1/n$. Similarly, L_2 is relevant to the estimation of the probability distribution function of covariates $F(x)$. It is the partial likelihood $L_1(\beta, \pi)$ that contributes to the estimation of β, and the probability of presence $p(y = 1|x, \beta)$.

Let us take a further look at the partial likelihood of L_1. The population prevalence π involves an integral $\int p(y = 1|x)dF(x)$, which can be approximated by $\sum_{x_l} p(y = 1|x_l, \beta)\frac{n_l}{n_0}$ on $L + 1$ known points of support x_l, with $F(x)$ replaced by its empirical estimate of $\hat{\alpha}_l = n_l/n_0$. This approximation for π can be rewritten as $\frac{1}{n_0}\sum_{i=1}^{n_0} p(y = 1|x_i, \beta)$, when we shift the sample space from the environmental space $x(s)$ to the geographic feature s (Hastie and Fithian 2013). Upon this transformation for π, the partial likelihood $L_1(\beta, \pi)$ in the LI method becomes

$$L_1(\beta) = \prod_{i=1}^{n_1}\frac{p(y = 1|x_i, \beta)}{\frac{1}{n_0}\sum_{j=1}^{n_0} p(y = 1|x_j, \beta)}. \tag{3}$$

This approximation of the partial likelihood for β is exactly the likelihood of the popular LK method (Lele and Keim 2006; Royle et al. 2012). Through maximizing the likelihood, it is possible to estimate the best fitting parameters β required to determine the probability of presence. The LK method can be viewed as a numerical approximation of the LI method, where the accuracy of the approximation will improve, in a statistical sense, by increasing the size of the background samples. In the following Simulations section, we will show numerically the equivalence between the LI and LK estimates, when the number of background are large.

Can the true probability of presence $p(y = 1|x)$ be estimated from the LI or LK methods? There have been extensive discussions on this topic (Lele and Keim 2006; Ward et al. 2009; Royle et al. 2012; Phillips and Elith 2013; Hastie and Fithian 2013; Solymos and Lele 2016). In the Simulations section, we will demonstrate the need for extra information, such as the parametric structure of $p(y|x)$ or prior knowledge of species' prevalence, π, in order to estimate the absolute probability of presence. As both the LI and LK methods require no prior knowledge of π, their successful operation relies on the resource selection probability function (RSPF) conditions, which have been given by Lele and Keim (2006) and further discussed in Solymos and Lele (2016). Loosely speaking, the RSPF condition includes, for example, that the true (actual) function of $\log p(y = 1|x)$ is nonlinear, and not all covariates in the model are categorical. Note that the logit-linear link function in Eq. 1 automatically satisfies the first criterion since for this case $\log p(y = 1|x)$ is nonlinear.

The relationship between LK and PPM

The point process model (PPM) in spatial analysis has recently been proposed as a versatile approach for analyzing species presence-background data (Warton and Shepherd 2010; Chakraborty et al. 2011), because it treats space as continuous, which seems more realistic than discrete

space approaches. Poisson point process models, however, have been shown to be closely connected to other popular methods in ecology, such as MAXENT (Aarts et al. 2012; Fithian and Hastie 2013; Renner and Warton 2013), logistic regression (Baddeley et al. 2010; Warton and Shepherd 2010), and resource selection models (Aarts et al. 2012).

In this section, we will briefly demonstrate how the likelihood of the LK method is associated with the conditional likelihood of the PPM, which is equivalent to MAXENT. We note that Aarts et al. (2012) also observed the equivalence between the LK and the conditional PPM. However, they did not provide any formal details of how the equivalence can be reached through a numerical approximation of the PPM, as we show here.

In the PPM framework, PB data consists of a set of locations $s_1, s_2, \cdots, s_{n_1}$, where individuals of a species are observed in a region D. These locations are defined as a realization of a point process that is characterized by the intensity $\lambda(s)$, which varies spatially according to a parametric function of environmental features $x(s)$. The likelihood proposed for fitting an inhomogeneous Poisson process is as follows:

$$L(s_1, \cdots, s_{n_1}, n_1) = \exp(-\int_D \lambda(s)ds)(\prod_{i=1}^{n_1} \lambda(s_i))/n_1! \quad (4)$$

(Cressie and Wikle (2011) and Renner et al. (2015)). This likelihood function was derived as the product of the conditional likelihood,

$$L_c(s_1, \cdots, s_{n_1}|n_1) = \prod_{i=1}^{n_1} \frac{\lambda(s_i)}{\int_D \lambda(s)ds}, \quad (5)$$

and the marginal likelihood,

$$P(N(D) = n_1) = \exp(-\Lambda(D))\Lambda(D)^{n_1}/n_1! \quad (6)$$

(Møller and Waagepetersen (2003) and Dorazio (2014)), where $\Lambda(D) = \int_D \lambda(s)\text{ds}$ is the cumulative intensity over the study area of D.

The likelihood (5) involves an integral over a study area that cannot be computed exactly and must be approximated numerically. Berman and Turner (1992) developed a numerical quadrature method for estimating the integral by approximating it as a finite sum using any quadrature rule, i.e., $\int_D \lambda(s)\text{ds} \approx \sum_{i=1}^{n_0} \lambda(s_i)w_i$. In the simplest form, we assign equal weight to each quadrature point, for example $w_i = \frac{|D|}{n_0}$, by partitioning D into n_0 equal rectangular tiles and a single quadrature point selected from each tile (Baddeley et al. 2015). $|D|$ represents the total area of the region. With this simple quadrature scheme, the conditional likelihood for the PPM is approximated by the following:

$$L_c = \prod_{i=1}^{n_1} \frac{\lambda(s_i)}{\frac{|D|}{n_0}\sum_{j=1}^{n_0} \lambda(s_j)}. \quad (7)$$

The above discretized version of the conditional likelihood of the PPM is the same as the likelihood of MAXENT using the log-linear intensity function, $\log \lambda(s) = \beta' x(s)$ (Fithian and Hastie 2013; Renner and Warton 2013). We want to address that this approximation is also the analogy of the likelihood of the LK method in Eq. 3. It is worth noting that the background points used to approximate π in the LK method play exactly the same role as the quadrature points, which are used in the PPM for numerically evaluating the cumulative intensity $\int_D \lambda(s)\text{ds}$. The choice of different quadrature schemes in approximating the conditional PPM can lead to models very different from the LK method (a discussion of the various quadrature schemes can be found in Chapter 9 of Baddeley et al. (2015)).

The difference between the LK and the approximated version of the conditional PPM lies in the link functions being used for modeling $p(y|x, \beta)$ and $\lambda(s)$ respectively, often chosen by consideration of the range of values a probability and an intensity function can take. Nevertheless, we will show through numerical simulations that the choice between the different link functions, for example, the logit, log-linear, or the complementary log-log functions, makes little difference in estimating the ratio, $\frac{p(y=1|x_i,\beta)}{\pi}$, i.e., the relative probability (or intensity) of presence. In other words, when using either the LK/LI, MAXENT, or conditional PPM model in studying the PB data, all of them will yield the same relative probability (or intensity) of presence. It is also worth mentioning that although the PPM has been introduced as a natural framework for modeling PB data, its ability to produce the *relative* probability of presence (or relative intensity), which is free of grid or transect selection, is the same as other so-called discrete space models.

Can the true intensity of presence be estimated with PPM methods? From the discussions of Fithian and Hastie (2013) and Renner et al. (2015) and Dorazio (2012), we know that the PPM can only estimate the intensity of reported presence from its full likelihood function, instead of the intensity of true presence $\lambda(s)$. It is because an underlying equation $\Lambda(D) = n_1$ is derived from the full likelihood function, in addition to the conditional likelihood of the PPM (Fithian and Hastie 2013). This additional information of $\Lambda(D)$ is biased, as the cumulative intensity should equal the number of true presence over the study area, whereas n_1 was only observed by opportunity. One way to correct for this bias is to make use of an appropriate species presence number in the PPM for $\Lambda(D)$.

The relationship between EM, SB, Lele, and LI methods

Ward et al. (2009) proved that the probability of presence is not identifiable from PB data, if there was no information about the structure of the probability function. Under this circumstance, the knowledge of the population's prevalence is required to estimate the true probability of presence. They used the commonly used logit function to fit the PB data, and proposed the EM algorithm to estimate the parameters of the logistic regression. The EM algorithm was able to estimate the probability of presence accurately at any site, using the species' prevalence as an additional information.

Two other successful methods discussed in the literature, the SC (Steinberg and Cardell 1992) and SB (Phillips and Elith 2011) method, also require the true species' prevalence to obtain estimates of the site-specific probability of presence. Although the EM and the SB methods work on different likelihood functions, their estimates of the probability of presence are essentially the same. It is also the first time to show the equivalence between the likelihood function of the EM/SB method and the LI method (see details in Appendix A).

Lele (2009) proposed a new method, referred to as the Lele method in our paper, to improve the instability of the LK method. The Lele method is a combination of the partial likelihood and data cloning to obtain the maximum likelihood estimator of both β and π. In Appendix B, we show that the likelihood function of the Lele method is the same as that of the LI method, although these two seemingly different approaches were developed independently.

General connections between all methods

The methods discussed in this paper, i.e., LI, LK, Lele, EM, SC, SB, MAXENT, and the PPM, can be divided into three different camps, based on their underlying likelihood functions and type of extra information required. The LI, LK, and Lele methods are sorted into Camp 1, which can estimate the absolute probability of presence, provided that the probability function satisfies the RSPF conditions (Lele and Keim 2006; Solymos and Lele 2016) (as given at the end of "The relationship between LI and LK methods"). The method of EM, SB, and SC fall into Camp 2, and are also able to estimate the absolute site-specific probability of presence, using an additional input of the species' prevalence. The MAXENT, continuous space PPM method and its associated models, are categorized into Camp 3, according to their connection between one another (Warton and Shepherd 2010; Baddeley et al. 2010; Aarts et al. 2012; Fithian and Hastie 2013; Renner and Warton 2013).

We have discussed some of the pairwise relationships, such as between LI and LK, LK and PPM, SB and LI, and between Lele and LI, respectively. Is there a way to connect together all the methods discussed in the paper? We believe this is in fact possible and a summary of our findings is given in Fig. 1, where the relations inside the same camps and across different camps are first time presented.

The methods in Camp 1 and Camp 3 (e.g., the LK and the PPM) are shown to share a common conditional likelihood, which has the same structure as the partial likelihood $L_1(\beta, \pi)$ in Eq. 2. The methods in Camp 1 and 2 (e.g., the LI, Lele, EM, and SB methods) are constructed on the same likelihood function, i.e., $L(\beta, h, \alpha, \pi)$ in Eq. 2,

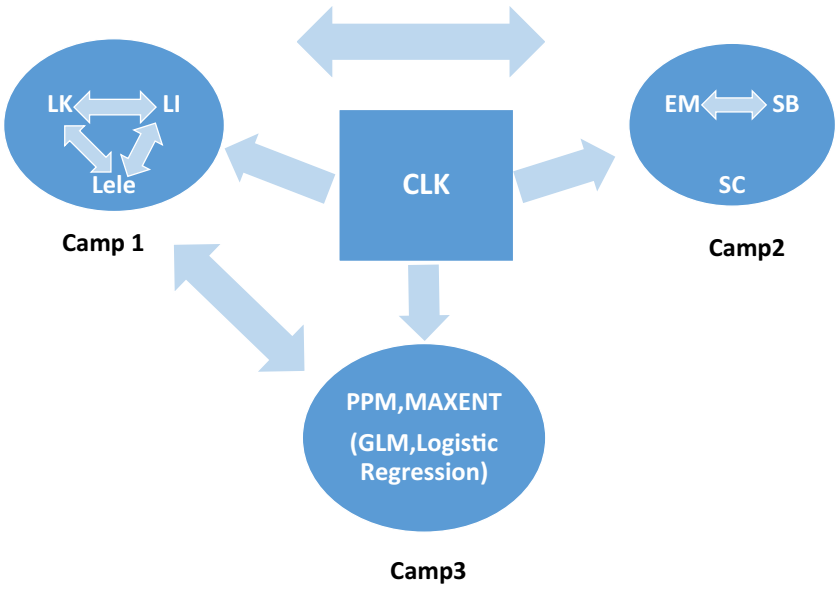

Fig. 1 The methods divide into three camps. Camp 1 includes the LI, LK, and Lele methods that can estimate the probability of presence, given the RSPF conditions are satisfied. Camp 2 includes the EM, SC, and SB methods that require the extra information of the species' population prevalence, in order to estimate the probability of presence. The MAXENT, PPM methods, and its associates are included in Camp 3, which in general estimate the relative probability of presence or the probability of reported presence

which can be further decomposed as the product of the likelihood $L_1(\beta, \pi)$ and other terms that do not involve both β and π. Therefore, all the methods in the three camps are actually built on the same partial/conditional likelihoods, i.e., $L_1(\beta, \pi)$. In other words, all these seemingly different SDM models are equivalent in their ability to estimate the *relative* probability of presence for modeling PB data, regardless of their different presentations.

The difference between Camp 1 and Camp 2 is that the methods in the latter require a pre-determined value of species' prevalence, π, while the LI and Lele methods in Camp 1 treats π as an unknown parameter. As for the LK method, in order for the LI and Lele methods to identify π, the RSPF conditions listed in Lele and Keim (2006) and Solymos and Lele (2016) need to be satisfied. However, this has led to controversy criticized by data scientists in particular (Ward et al. 2009; Phillips and Elith 2011; Hastie and Fithian 2013), because the true parametric functions are generally unknown in practice, and the functions used to fit these true functions can be of different structures. Under these circumstances, a revised version of the LK method is proposed in the next section, where the PB data is augmented with an additional datum on the species' prevalence π. This makes the LI/LK methods comparable to the EM, SB, and SC methods.

A unified Constrained LK (CLK) method

From previous studies, we have found that the LI, LK, MAXENT, and the conditional PPM share a similar likelihood function i.e., Eqs. 3 and 7, which alone (without extra information) can only provide the relative probability (intensity) of presence. In order to obtain the absolute probability of presence, an extra information of the species' prevalence π can be introduced as a constraint imposed on the optimization of this common likelihood function. In details, the CLK method maximizes the following (LK type of) likelihood function,

$$L_1(\beta) = \prod_{i=1}^{n_1} \frac{p(y=1|x_i, \beta)}{\frac{1}{n_0}\sum_{j=1}^{n_0} p(y=1|x_j, \beta)}, \tag{8}$$

with the constraint, i.e., $\frac{1}{n_0}\sum_{j=1}^{n_0} p(y=1|x_j, \beta) = \pi_0$, where π_0 is the population prevalence that is assumed to be known in advance. Note that this is very different from just maximizing the function of $\log L = \sum_{i=1}^{n_1} \log \frac{p_i}{\pi_0}$, since the constraint reduces the effective parameter space over which the maximization is performed. The statistical mechanism and efficiency underlying the CLK method is provided in Appendix C, where we have proved that the CLK is capable of estimating the true probability of presence, the same as the SB and SC methods.

The LI, LK, and the partial likelihood of the PPM (or MAXENT) would intrinsically have identification problems in solving their likelihood functions, if there is no prior knowledge of the species prevalence, and/or the structure of the function of the probability of presence. In other words, these methods in general would generate multiple solutions of the absolute probability of presence, i.e., the relative probabilities of presence. By introducing the constraint, the CLK method forces the estimates from these methods to converge to the unique solution, which is just one of the multiple solutions obtained from the LI, LK, and the MAXENT methods. The CLK method provides a unification of the seemingly disparate methods discussed so far (SB, SC, EM, LI, LK, Lele, PPM, and MAXENT). Each of these methods can be shown to be either equivalent to, or a special case of, the CLK method.

Firstly, LK, LI, and Lele are special cases of the CLK method, when the RSPF conditions (Lele and Keim 2006; Solymos and Lele 2016) are satisfied and no constraint is used. If the RSPF conditions are not satisfied, using the logit-linear or other functions to fit without constraint fails to estimate the probability of presence (Phillips and Elith 2013). The inclusion of the additional information of π in the CLK method fixes this problem, and enables the LI/LK methods to perform as well as SB, SC, or EM method.

Secondly, the CLK method has the same performance as the SB, SC, and EM methods, when the logit link function is employed. However, unlike these methods which were only derived for the logit function, the formulation of the CLK method is much simpler and can easily adapt to any type of link functions.

Next, the PPM can be reviewed as a special case of the CLK method, when the log-linear function is used for $p(y|x, \beta)$, and a constraint of $\frac{n_1}{|D|}$ is imposed on the denominator of Eq. 8. For a log-linear function, i.e., $\log p(y|x, \beta) = \beta_0 + \beta_1' x$, estimates of β_1 are the same for both the CLK method and the conditional PPM (equivalently the MAXENT), whereas the ratio of the two methods differ by a constant, i.e., the exponent of the difference between the two β_0's (Fithian and Hastie 2013). It is the constraint that provides the estimate of the intercept in the log-linear model. Similarly, MAXENT model is also a special case of the CLK method, using the logarithm function but without any constraint supplied.

Unlike all of these previous methods, the CLK does not specify any particular link function; instead, it can use any of the commonly used link functions, such as the logit, log, or the complementary log-log functions. We will show in the Simulations section that using different link functions actually have little difference on estimating both the relative and the absolute probabilities of presence. The proposed CLK method is easy to implement, and users can choose any general-purpose non-linear constraint optimization

package in their preferred programming language. We have implemented the CLK method in R, and used the constraint optimization package 'nloptr' (Ypma 2014).

Simulations

In this section, the performance of the proposed CLK method is evaluated through numerical simulations, using three commonly applied link functions, logit-linear (see Eq. 1), log-linear, and complementary log-log, denoted separately as CLK_logit, CLK_log, and CLK_clog. The CLK method can easily include other link functions. The large sample equivalence between the LI and LK methods is also demonstrated through these numerical experiments. As for other well performing methods, such as the SC, EM, and SB, their equivalence to each other and to the CLK_logit model have been proved in "The relationship between EM, SB, Lele, and LI methods" and "A unified Constrained LK (CLK) method". Furthermore, the numerical performance of these methods have already been assessed in Phillips and Elith (2013); these methods are therefore not included in our simulation study.

We consider eight species, with seven of them having the same probability functions of occurrence used in Phillips and Elith (2013) (see Table 1). The extra species considered in our paper has the exponential distribution. The probability of presence $p(y = 1|x)$ depends only on a single environmental covariate or explanatory variable x, and its value ranges uniformly between [0,1]. Multiple simulated datasets were constructed for each of the eight species. Five models were considered for fitting the data, i.e., LI, LK, CLK_logit, CLK_log, and CLK_clog. In our model fitting, no knowledge is assumed about the parametric structure of the true probability of presence, and we fit the data with the commonly used logit function for both the LI and LK methods. The log-linear function was also used to fit the data for the LI and LK methods (see Table 3).

We plot the logarithm of each probability function in Table 1 to verify the RSPF conditions listed in Lele and Keim (2006) and Solymos and Lele (2016). It is observed from Fig. 2 that only Logistic-2 and Gaussian distributions exhibit the required key condition, namely that $\log p(y = 1|x)$ is nonlinear, and appear to satisfy the RSPF conditions, from the eight species. The other species, which do not satisfy the RSPF conditions, cannot be expected to get reasonable parameter fits from the LI and LK methods.

In constructing simulations, for each species, 1,000 presence samples were drawn representing the locations of those observed individual, and 20,000 background samples were randomly drawn. Similarly, another 1,000 samples were also drawn as the validation data sets for model assessment. One hundred simulations were run, and both the LI and LK methods were used to fit each simulation. The three CLK models were only fitted and plotted for one of the 100 simulations respectively, as all the 100 fits were very similar to each other for each CLK model. The fits were compared both visually (Fig. 3) and using the validation root mean square (RMS) error (Fig. 4) as the assessment statistics, against the true probability of presence. We also calculated the AUC value to compare the performance of LI, LK, and CLK methods. For the "Quadratic" and "Gaussian" species, quadratic terms of x were added to fit the true probability. As the CLK method requires an estimate of the species' prevalence, we use the true prevalence as the estimate. Sensitivity analysis was also carried out by varying the true prevalence by $\pm$ 0.1, and the results are reported in Fig. 3 as well.

We note that the numerical results of the LI and LK methods reported in Phillips and Elith (2013) appear different to those reported here, because parameters of these two methods are not identifiable in some of the simulations. In our simulations, the identifiability was assessed by computing the reciprocal of the condition number, the ratio of the largest to the smallest eigenvalues of the Hessian matrix. A ratio very close to zero (not exactly zero using the Hessian matrix as the estimate) indicates an identifiability issue. We arbitrarily chose 0.001 as the threshold to assess the identifiability for each simulation. The summary statistics (means and standard errors of the estimates) were computed for the adjusted intercept $\hat{\beta}_0$ and slope $\hat{\beta}_1$, after removing unidentifiable simulations. In order to demonstrate the large sample equivalence between the LI and the LK methods, both methods were fitted with logit-linear (see Table 2), and log-linear functions (Table 3), and their summary statistics are shown in the Tables separately. Only the slope

Table 1 Probability of presence for eight simulated species

Simulated species	Probability of presence $P(y = 1\|x)$
Constant	0.3
Linear	$0.05 + 0.2x$
Exponential	$\exp(-4 + 4x)$
Quadratic	$0.5 - 1.333(x - 0.5)^2$
Gaussian	$0.75\exp[-(4x - 2)^2]$
Semi-logistic	$8/(1 + \exp[4 - 2x])$
Logistic-1	$1/(1 + \exp[4 - 2x])$
Logistic-2	$1/(1 + \exp[4 - 8x])$

x is the single environmental covariate that is uniformly distributed on [0,1]

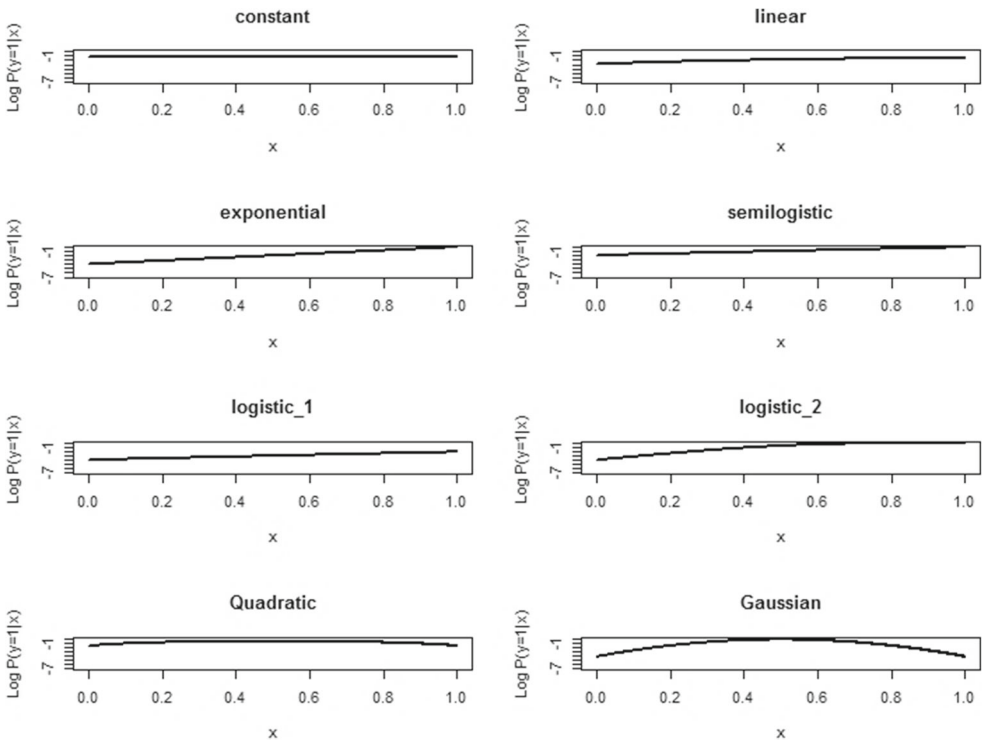

Fig. 2 The logarithm of each probability function in Table 1 is plotted, in order to verify the RSPF conditions (Lele and Keim, 2006; Solymos and Lele, 2016), i.e, log $p(y|x, \beta)$ being non-linear. Logistic-2, Quadratic and Gaussian distributions appear to satisfy the RSPF conditions

estimates are reported in Table 3, because the intercept of the log-linear model is not identifiable for the LI and LK methods.

We also plotted the ratio of $\frac{p(y=1|x,\hat{\beta})}{\hat{\pi}}$ for the three CLK methods, as well as the LK method fitted with log-linear and logit-linear functions respectively, shown in Fig. 6. The relative probabilities of the LK method fitted with the log and logit-linear link functions were plotted for each of the 100 simulations, while the relative probabilities for the CLK methods were only plotted once due to the high similarities among the 100 replications.

Simulations were also investigated in order to test and compare model performance with different size of presences. In this case, we chose to compare PB datasets with 100, 500, and 5,000 presences. For the 5,000 presence simulation, 50,000 background points have been used. Similarly, 100 simulations were run for each species, and the data was fit by the LI and LK methods (using logit-linear function), CLK_logit, CLK_log, and CLK_clog respectively. The validation RMS errors were calculated for these different approaches in Fig. 5, for different number of species presences.

The R code provided by Phillips and Elith (2013) facilitated our programming process. All model fitting was carried out in R version 3.2.2 (R Core Team, 2016). The R code for both the simulation and the CLK method can be requested from the correspondence author.

Results

Firstly, we see in Fig. 3 that when the true species probability of presence is logit-linear in the case of Logistic-2, the LI/LK methods fit the data well, because the logit-linear function satisfies the RSPF conditions (Lele and Keim 2006; Solymos and Lele 2016) as discussed at end of "The relationship between LI and LK methods". In most other cases, both LI/LK methods provide poor fits to different distributions. They show a widespread for their estimates in the plots, which gives an indication of the non-identifiability of LI/LK methods in estimating the probability of presence. This occurs because the probability functions for most simulated species do not satisfy the RSPF conditions, except for the Gaussian distribution (see Fig. 2 for details). However, even then, the performance of the LI/LK model was not good for fitting the Gaussian distribution. When the PB data is augmented with the species' prevalence, the CLK method closely approximates the true probability of presence, using the log-linear (yellow lines), logit (red lines), or the complementary log-log link functions (blue lines) (except for the species of Logistic-1). The CLK method consistently performs well in Fig. 5, where the number of species presences changes from small to large samples.

Upon examining Table 2, the coefficient estimates from both the LI and LK methods apparently are different

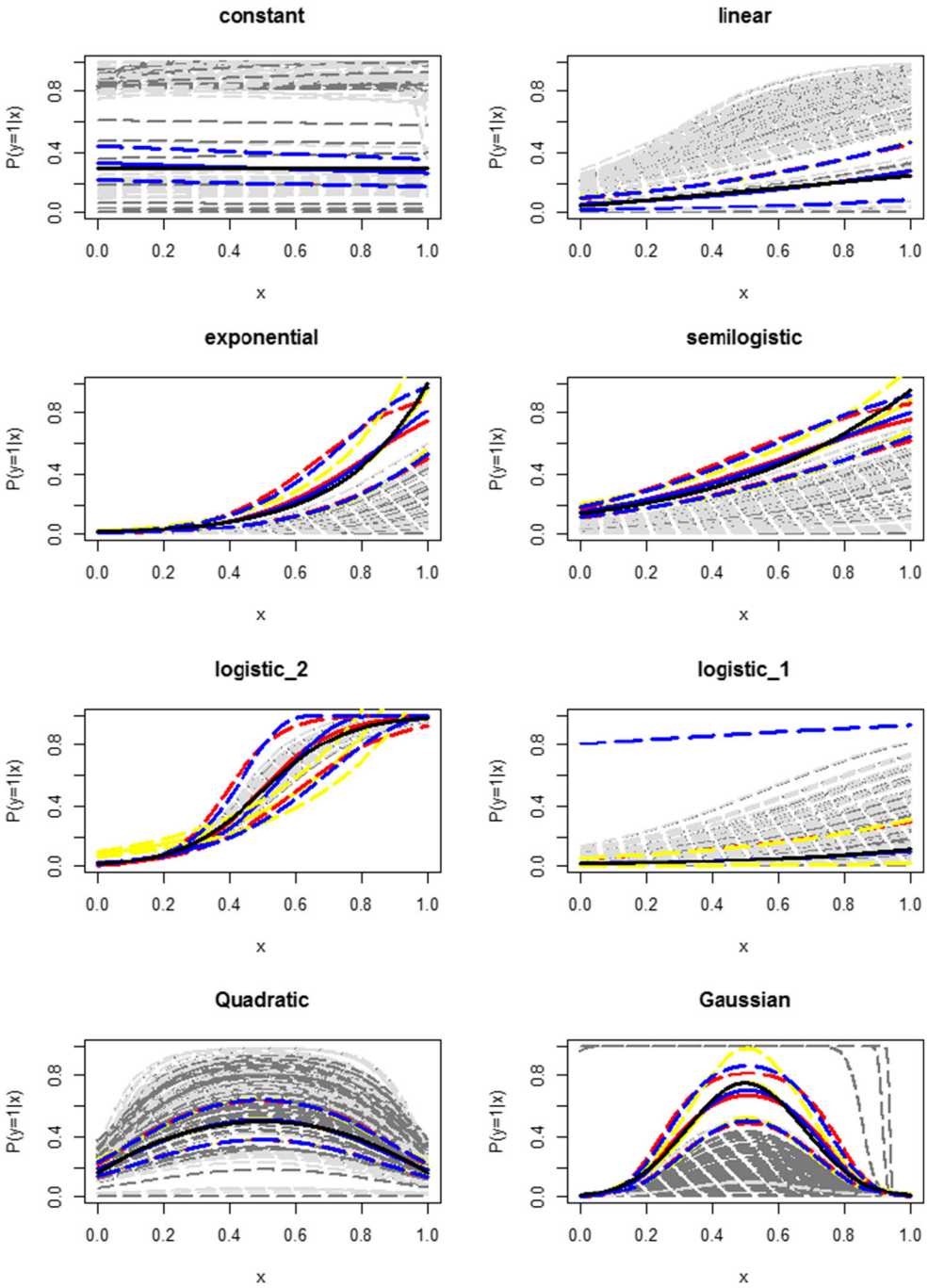

Fig. 3 LK and LI methods are fitted with logit-linear function for each species with a replication of 100 times (two types of gray-dotted lines). The true probability is noted with the solid black line. CLK method is fit with the logit-linear function (red line), log-linear function (yellow line), and the complementary log-log function (blue line), using the true prevalence as π_0. The red, yellow, and blue-dashed lines are the CLK estimates fitted with logit, log-linear, and complementary log-log functions, respectively, using the true prevalence $\pm$ 10% as π_0

(except for Logistic-2) from the CLK estimates, which all well approximate the true probabilities of presence. It also indicates the LI and LK are biased methods when making inference of the probabilities of species' presence. For species Logistic-2, the β's estimates of the LI/LK and CLK methods are very similar. This further confirms our statement in "A unified Constrained LK (CLK) method" that LK and LI are just special cases of the CLK method, when the RSPF condition is satisfied. Apart from the disparity in the estimates of β's between the LI/LK and the CLK method, the standard errors of LI/LK methods in general are higher than the CLK estimates. For some species, such as the Quadratic or the Gaussian distribution, both the intercept and the slope have significantly large standard errors that would lead to possible rejection of the influential covariate, if we were to use the LI and LK methods to make statistical inference.

Secondly, we can see a close resemblance between the two gray-dotted lines fitted with LI and LK methods using the logit function respectively (Fig. 3). This resemblance

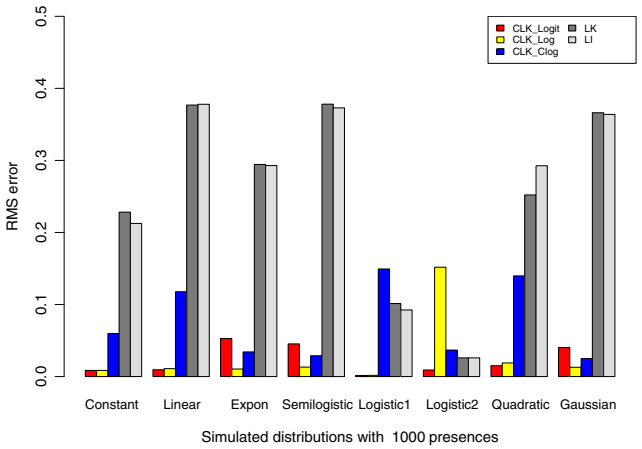

Fig. 4 Root mean square (RMS) error of the LI, LK both fitted with logit link functions (two gray columns) and the CLK method using the logit (CLK_Logit, red), log-linear (CLK_Log, yellow), and the complementary log-log (CLK_Clog, blue) functions, with the number of species presences of 1000

may also be seen when comparing the validation RMS errors of the two methods (see two gray column charts in Fig. 4 for the simulation of 1,000 presences). However, there still exists a small discrepancy between the two methods for some simulated species, for example, with the constant and quadratic distributions. This is because multiple solutions may be obtained, due to the identifiability issue inherent in the LI and LK methods in estimating the actual probabilities of presences. The discrepancy between the LI and LK methods caused by the identifiability issue is more obvious when the number of presence samples is small (e.g., 100 presences in Fig. 5). However, it become less obvious as both the number of presences and background points increase (e.g., 5,000 presences in Fig. 5). After removing all unidentified simulations in the 1,000 presence simulation, the estimates for the LI and LK methods (fitted with logit-linear function) are nearly identical to each other, with the mean and standard errors of the estimates provided in Table 2. The similar results of the LI and LK methods can also be seen in Table 3, where both methods were fitted with

Table 3 Mean of $\hat{\beta}_1$ for LI, LK, and the CLK, fitted with log-linear function (standard errors provided in parentheses)

	LK1-$\hat{\beta}1$	LI1-$\hat{\beta}1$	CLK_log-$\hat{\beta}_1$
Constant	0.015 (0.109)	0.015 (0.109))	0.015 (0.109))
Linear	1.367 (0.121)	1.370 (0.121)	1.367 (0.121)
Exponential	3.992 (0.183)	3.995(0.184)	3.992 (0.183)
Semilogit	1.909 (0.113)	1.910 (0.113)	1.909 (0.113)
Logistic-1	1.869 (0.124)	1.871 (0.124)	1.869 (0.124)
Logistic-2	2.767 (0.109)	2.802 (0.109)	2.766 (0.109)
Quadratic	3.813 (0.440)	3.819(0.440)	3.813 (0.440)
Gaussian	16.046 (0.825)	16.047 (0.824)	16.046 (0.825)

the log-linear functions. Obviously, the slope estimates of the LI and LK method are nearly the same.

When all methods were fitted with the log-linear functions in Table 3, not only are the slope estimates of the LI and LK methods nearly the same, but they are also the same for the CLK method. The resulting relative probabilities of presence from these three models are all proportional to the true probability of presence, by a ratio of $1/\log\hat{\beta}_0$, estimated from the CLK method. Meanwhile, in most of our simulated species, the estimates fitted by a log-linear function in general have better performance compared to the estimates fitted with either a logit or complementary log-log functions. This was only violated for species Logistic-2, where the true probability function is logit-linear but the data was fitted with the log-linear function.

The AUC values of the LI and LK methods, which are shown to have poor predictions, are surprisingly the same as the well-fitted CLK method (e.g., constant: 0.628, linear: 0.630, Logistic-1: 0.651, and Logistic-2: 0.907). This is due to the fact that the AUC is a rank-based measurement, while the LI/LK methods have preserved the ranks of the actual probabilities. The caution of using AUC as a measure of model accuracy in estimating the probability of presence will be addressed in the Discussion section.

Table 2 Mean of $\hat{\beta}_0$ and $\hat{\beta}_1$ for LI, LK, and CLK, fitted with logit-linear function from Eq. 1 (standard errors provided in parentheses)

	LK-$\hat{\beta}0$	LI-$\hat{\beta}0$	LK-$\hat{\beta}1$	LI-$\hat{\beta}1$	CLK-$\hat{\beta}0$	CLK-$\hat{\beta}1$
Constant	− 2.345 (1.242)	−2.442 (0.294)	0.262 (1.400)	− 0.001 (0.134)	− 0.858 (0.088)	0.0008 (0.177)
Linear	− 1.529 (0.271)	− 1.527 (0.268)	3.119 (0.767)	3.128 (0.770)	−2.628 (0.085)	1.634 (0.147)
Exponential	−5.716 (0.559)	− 5.724 (0.574)	4.356 (0.289)	4.355 (0.290)	− 4.550 (0.212)	5.664 (0.316)
Semilogit	− 2.958 (0.550)	− 2.963 (0.558)	2.422 (0.305)	2.421 (0.307)	− 2.048 (0.113)	3.452 (0.218)
Logistic-1	− 2.821 (0.539)	− 2.822 (0.541)	2.491 (0.467)	2.491 (0.467)	− 3.991 (0.084)	1.985 (0.132)
Logistic-2	− 4.055 (0.243)	− 4.056 (0.243)	8.073 (0.777)	8.074 (0.775)	− 4.050 (0.223)	8.105 (0.462)
Quadratic	− 1.700 (1.997)	− 1.488 (0.966)	10.202 (4.535)	10.096 (4.698)	−1.484 (0.157)	6.036 (0.751)
Gaussian	− 3.030 (0.677)	− 3.016 (0.614)	5.362 (7.341)	5.502 (7.315)	− 5.042 (0.302)	22.983 (1.467)

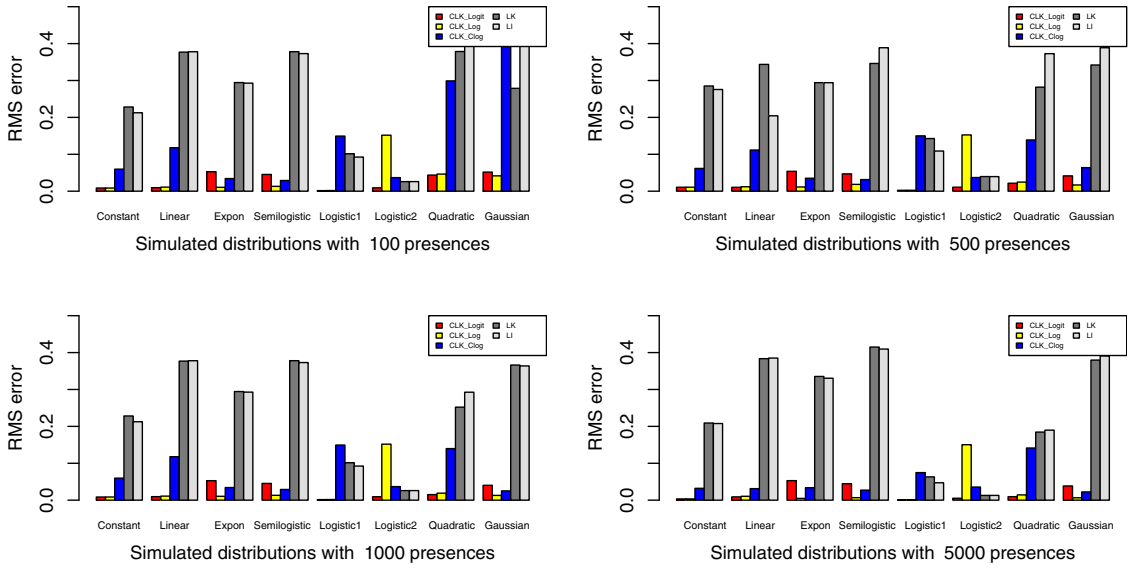

Fig. 5 Validation root mean square (RMS) errors of the LI, LK both fitted using the logit link functions (two gray columns) and the CLK method using the logit (CLK_Logit, red), log-linear (CLK_Log,yelllow), and the complementary log-log (CLK_Clog,blue) functions, for different numbers of species presences (100, 500, 1000, and 5000 presences)

Although it is hard to see what the LK or LI method have estimated in Table 2, this ambiguity, however, becomes clear when we plot the relative probability of presence, i.e., the ratios $\frac{p(y=1|x,\hat{\beta})}{\hat{\pi}}$ of the LK estimates fitted with both the logit and log-linear functions (Fig. 6). Comparing these ratios with the CLK estimates, we see that these ratios are all similar to each other, regardless of the functional form of the link function and which type of likelihood (full vs. the conditional) have been used to fit the PB data. It further confirms that the LK/LI method can provide a good estimate of the relative probability of presence, when no extra information is available on either the RSPF conditions (Lele and Keim 2006; Solymos and Lele 2016) or the species' prevalence. Also, there are some "erratic" curves observed for the LI and LK estimates in Figs. 3 and 6 for the species with a Gaussian distribution. These estimates were again simply caused by the non-identifiability problem in the LI and LK methods.

Discussion

In this paper, we have studied some commonly used methods for modeling species probability of presence with PB data. These methods include the LI (Lancaster and Imbens 1996), LK (Lele and Keim 2006; Royle et al. 2012), Lele (Lele 2009), EM (Ward et al. 2009), SB (Phillips and Elith 2011), SC (Steinberg and Cardell 1992), MAXENT (Phillips and Dudík 2008), and point process models (Warton and Shepherd 2010; Chakraborty et al. 2011). Firstly, we have shown that it is the conditional/partial likelihood, i.e., Eq. 3 that is actually employed for modeling PB data, where the species prevalence is in general unknown. The LI, LK, Lele, MAXENT, and the conditional PPM model, built upon the conditional/partial likelihood, alone (without extra information) can only be used to make inference about the relative probability of presence. Other methods, such as the Poisson-generalized linear regression, logistic regression, and the PPM, can only estimate the probability of reporting (as opposed to the probability of presence), due to the lack of appropriate information on the true population prevalence or number of true presences (Fithian and Hastie 2013; Dorazio 2014; Renner et al. 2015). In order to estimate the actual probabilities of species' presence, extra information is needed, such as the parametric structure of the probability function or the species' prevalence π. The methods of SC, EM, and SB require a pre-determined value of π, while the LI and Lele methods need the RSPF conditions to be satisfied (Lele and Keim 2006; Solymos and Lele 2016) (see end of "The relationship between LI and LK methods"). Otherwise, the LI, LK, and Lele methods intrinsically have identification problem in estimating all the parameters relevant to the true probabilities of presence. However, the parametric RSPF conditions are often hard to meet in practice, and this has led to much controversy as to whether the LK method is capable of estimating the actual probabilities of presence in modeling PB data (Ward et al. 2009; Phillips and Elith 2011; Hastie and Fithian 2013).

Under these circumstances, a revised version of the LK method is proposed in "A unified Constrained LK (CLK) method", where the PB data is augmented with an additional

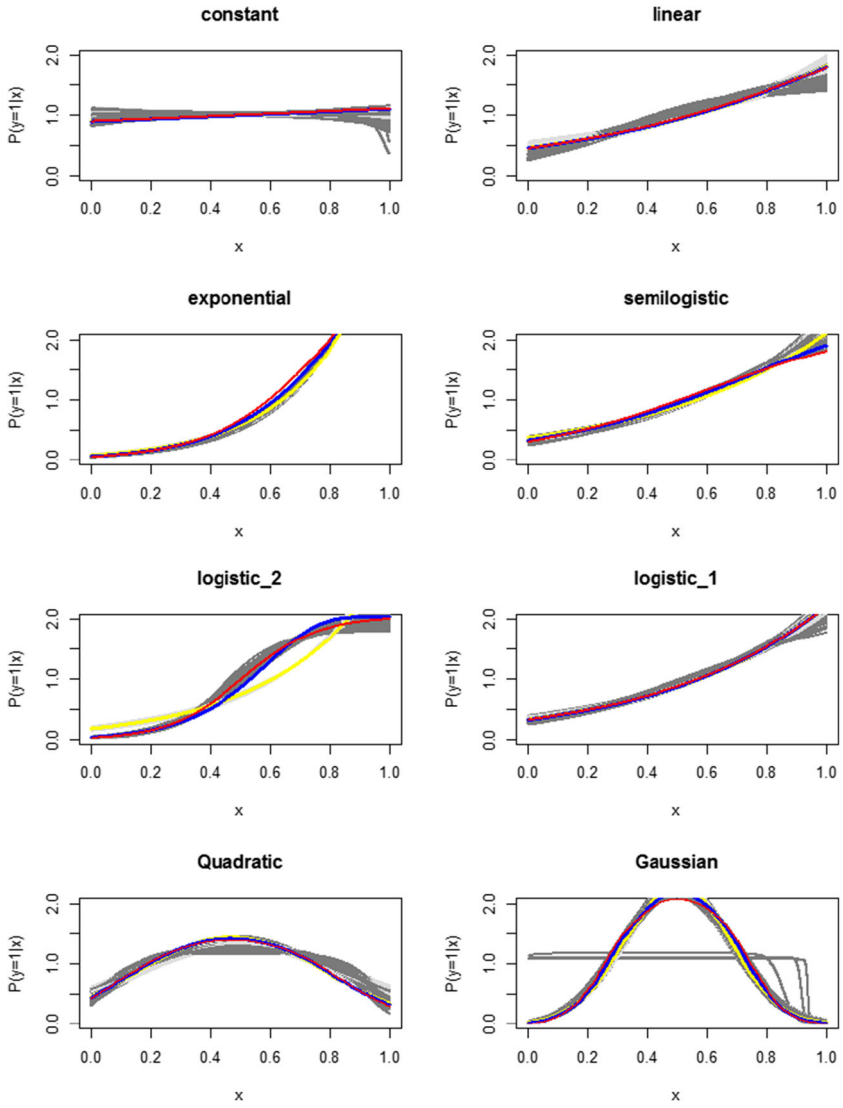

Fig. 6 The ratio between the estimated probability of presence $p(y = 1|x, \hat{\beta})$ and the estimated population prevalence $\hat{\pi}$, are fitted with the LK method using logit (gray 1 line) and log-linear function (gray 2 line) over 100 simulations. The ratio is also fitted with the CLK_Logit (red line), CLK_Log (yellow line), and CLK_Clog methods (blue line), using one randomly selected simulation (due to resemblance among replications)

datum on the species' prevalence π. The introduction of the constraint in the CLK method guarantees a unique estimate for the probabilities of presence, which well approximate the true probabilities of presence when the supplied prevalence is close to the true one. This unique estimate is just one of the multiple solutions obtained from the LI, LK, and the MAXENT methods. The CLK method makes the controversial LI/LK approaches, as well as the conditional likelihood of PPM (or MAXENT) method, comparable to other well-performing methods (SC, EM, and SB).

One may argue that the CLK method requires the population prevalence π, which is sometimes hard to obtain or estimate in practice. For our purposes, the CLK method proposed in this paper serves more as a technical generalisation tool to gain insight into modeling PB data, and to look at the connection of seemingly different methods. On the other hand, the information of population prevalence can be obtained independently from either pilot studies or other types of data, for example, the PA survey data or the complementary expert map. There have been a few recent studies on the combination of PB and PA data (Dorazio 2014; Fithian et al. 2015; Koshkina et al. 2017). These combined methods can estimate the absolute probability of presence successfully, by gaining the information of population prevalence from PA data. One might also expect that if an estimate of species prevalence is given, it should be a simple matter of normalizing all probabilities by a scale factor, say $\frac{p(x=1|y)}{f(x)}\hat{\pi}$, to obtain the true probabilities of presence at any site. Such a procedure though, is not very useful, as it does not leave us with a working model to understand how the covariates impact the probability of presence, i.e., the normalization will not give us the correct values of the coefficients β. Meanwhile, estimates from the LI/LK methods may have

larger standard error compared to the CLK methods (see Table 2). Hence, these methods may erroneously reject those influential covariates that have significant impact on the species' probability of presence in statistical inference.

In the simulation studies, we have used the predicted probabilities and the validation RMS error to assess and compare the performance of different models. There have been several recognized features of the AUC value (under the ROC curve) that prevents its use as a measure of model accuracy in spatial distribution modeling (Lobo et al. 2007). The latter paper has pointed out that "AUC scores ignore the actual probability values, being insensitive to transformations of the predicted probabilities that preserve their ranks." This is particularly evident in our study, where the proportional transformations of species occurrence probabilities, such as those by LI and LK methods, may dramatically change the prediction output but do not have any effect on the AUC scores. Therefore, there is a need for caution when using the AUC value to assess the goodness-of-fit of the distributions models, particularly when the probability values are of interest as in this paper.

In this paper, we have revisited some commonly used regression-based SDM methods for modeling species probability of presence with PB data, including the SB, SC, EM, LI, LK, Lele, as well as point process models and MAXENT method. In the past, there have been numerous serious attempts to find commonalities in these different methods, and these have been reported in the statistical and ecological literature (Lele and Keim 2006; Keating and Cherry 2004; Lele 2009; Ward et al. 2009; Warton and Shepherd 2010; Baddeley et al. 2010; Aarts et al. 2012; Fithian and Hastie 2013; Renner and Warton 2013; Phillips and Elith 2013; Hastie and Fithian 2013; Solymos and Lele 2016). From our study, we can conclude that all these different methods, regardless of using extra information or not and their popularity, are essentially the same for estimating the relative probabilities of presence for PB data. Furthermore, these methods also have similar performance in estimating the absolute probability of presence, when the same additional information is provided. In particular, this paper has proposed a constrained LK (CLK) method as a unification of these better known existing approaches, with less theory involved and greater ease of implementation. Compared to other well-performing methods, the CLK does not specify any particular link function; instead, it can use any of the commonly used link functions, such as the logit, log, or the complementary log-log functions. More importantly, it provides such a generalization that each of the SDM approaches discussed in this paper, e.g. SB, SC, EM, LI, LK, Lele, and the Poisson point process method, is either equivalent to, or a special cases of, the CLK method.

Funding information This study is supported by the Australian Research Council with the grant DP150102472.

Appendix A: Equivalence of EM, SB, and the LI methods

In the following, we demonstrate how the expectation-maximisation (EM) method (Ward et al. 2009) and the scaled binomial loss model (SB) (Phillips and Elith 2011) are essentially the same as the LI method through simple mathematical derivations.

Ward et al. (2009) let $z = 1$ and $z = 0$ denote the observed presences and background data, respectively. Note that when $z = 1$, we know $y = 1$. However, when $z = 0$, we do not know whether $y = 0$ or $y = 1$. The EM method proposed the following likelihood for the presence-background data (Ward et al. 2009):

$$\begin{aligned} L(\eta|z, X) &= \prod_i P(z_i|s_i = 1, x_i) \qquad (\boldsymbol{EM}) \\ &= \prod_i \left(\frac{\frac{n_1}{\pi n_0} e^{\eta(x_i)}}{1 + (1 + \frac{n_1}{\pi n_0}) e^{\eta(x_i)}} \right)^{z_i} \\ &\quad \times \left(\frac{1 + e^{\eta(x_i)}}{1 + (1 + \frac{n_1}{\pi n_0}) e^{\eta(x_i)}} \right)^{1 - z_i} . \end{aligned} \tag{A.1}$$

Here, n_0 is the number of observed presences denoted by $z = 1$, and n_1 the number of background points denoted by $z = 0$. The notation $s = 1$ is a construct of case-control modeling and indicates that this observation is in the presence-background data sample. The logit link function is used to model the true probability of presence, i.e., $\log \frac{p(y=1|x)}{1-p(y=1|x)} = \eta(x)$. As the information on the true presence y is missing, direct maximisation of this likelihood is difficult. The EM technique is implemented on the full likelihood of both the true and observed presences, with the missing y imputed with its expectation.

Now examine the SB method, in which the probability used in the likelihood function is defined as $P_{\mathrm{UA}}(s = 1|x) = \frac{1}{1+r+\exp(-\eta(x)+\ln r)}$ (Phillips and Elith 2011), where r equals $\frac{1-f_p}{f_p}\pi$ through the sampling probability of the presence points f_p. The sampling probability f_p in the SB

method can be rewritten as $f_p = \frac{n_0}{n_1+n_0}$, and it therefore gives $r = \frac{n_0}{n_1}\pi$. The probability $P_{\mathrm{UA}}(s = 1|x)$ can be rewritten as follows:

$$\begin{aligned} P_{\mathrm{UA}}(s=1|x)\ (\boldsymbol{SB}) &= \frac{1}{1+r+\exp(-\eta(x)+\ln r)} \\ &= \frac{1}{1+\frac{\pi n_0}{n_1}+\frac{\pi n_0}{n_1}e^{-\eta(x)}} \\ &= \frac{\frac{n_1}{\pi n_0}e^{\eta(x)}}{1+(1+\frac{n_1}{\pi n_0})e^{\eta(x)}}. \end{aligned} \tag{A.2}$$

It is obvious that $P_{\mathrm{UA}}(s = 1|x)$ used in the SB method is exactly the same as $P(z|s = 1, x)$ of the EM method. Instead of working indirectly on the likelihood of the observed data (as the EM method), the SB method directly maximizes the observed likelihood function from the outset, by using a modification of the standard binomial loss function.

In the LI method, each observed presence is drawn uniformly with the probability h, the same as f_p defined in the SB method. The likelihood function $L(\beta, \pi, h)$ is constructed on the probability R_{1n} through $L(\beta, \pi, h) = \prod_i R_{1n}(\beta, \pi, h)^{z_i}(1 - R_{1n}(\beta, \pi, h))^{1-z_i}$, where $R_{1n} = \frac{(h/\pi)P(y=1|x,\beta)}{(h/\pi)P(y=1|x,\beta)+1-h}$ (Lancaster and Imbens 1996). When $p(y = 1|x, \beta)$ takes the same logit-linear function as the EM and SB methods, i.e., $P(y = 1|x, \beta) = \frac{e^{\eta(x)}}{1+e^{\eta(x)}}$, one finds that

$$R_{1n}\ (\mathbf{LI}) = \frac{\frac{n_1/\pi}{n_1+n_0}\frac{e^{\eta(x)}}{1+e^{\eta(x)}}}{\frac{n_1/\pi}{n_0+n_1}\frac{e^{\eta(x)}}{1+e^{\eta(x)}}+\frac{n_0}{n_0+n_1}} = \frac{\frac{n_1}{\pi n_0}e^{\eta}(x)}{1+(1+\frac{n_1}{\pi n_0})e^{\eta}(x)}. \tag{A.3}$$

Obviously, $R_{1n}(\boldsymbol{LI}) = P(z|s = 1, x)(\boldsymbol{EM}) = P_{\mathrm{UA}}(s = 1|x)(\boldsymbol{SB})$, i.e., the probabilities on which the likelihood functions were formulated, are the same for these three seemingly different methods. The difference lies in the extra information required: the SB and EM methods need a pre-determined value of π, while the LI method treats π as one of the unknown parameters. In order for the LI method to identify π, the identifiability conditions listed in Lele and Keim (2006) and Solymos and Lele (2016) have to be satisfied. The numerical examples in Lancaster and Imbens (1996) paper work well, because they satisfy these parametric identifiability conditions.

Table 4 The table summarizes the key symbols used in Lele (2009) and our paper, making them comparable to each other

Definition	Lele (2009)	Our paper
Probability of presence	$\pi(X_i, \beta)$	$p(y_I = 1|x_i, \beta)$
Population prevalence	$P(\beta)$	π
Number of presences	N	n_1
Number of background	M	n_0
Sampling probability	$w = \frac{N}{N+M}$	$h = \frac{n_1}{n_1+n_0}$

Appendix B: Equivalence between the Lele (2009) and LI (1999) methods

Lele's partial likelihood is given in Lele (2009; Eq. 2) as follows:

$$\begin{aligned} \mathrm{PL}(\beta) = &\prod_{i=1}^{N}\frac{w\pi(X_i^U,\beta)}{w\pi(X_i^U,\beta)+(1-w)P(\beta)} \\ &\times\prod_{j=1}^{M}\frac{(1-w)P(\beta)}{w\pi(X_j^A,\beta)+(1-w)P(\beta)} \end{aligned} \tag{B.1}$$

Before we show the equivalence between this partial likelihood (PL) in Lele (2009) and that of Lancaster and Imbens (1996), we summarized the comparable notations used by these different approaches.

The likelihood function $L_1(\beta, \pi, h)$ in Lancaster and Imbens (1999) is constructed from the probability R_{1n} through $L_1(\beta, \pi, h) = \prod_i R_{1n}(\beta, \pi, h)^{z_i}(1 - R_{1n}(\beta, \pi, h))^{1-z_i}$, where $R_{1n} = \frac{(h/\pi)P(y=1|x,\beta)}{(h/\pi)P(y=1|x,\beta)+1-h}$. Using the comparable notations in Table 4, we can easily rewrite the partial likelihood function of Lele (2009) as follows:

$$\begin{aligned} \mathrm{PL}(\beta) &= \prod_{i=1}^{N}\frac{w\pi(X_i^U,\beta)}{w\pi(X_i^U,\beta)+(1-w)P(\beta)}\prod_{j=1}^{M}\frac{(1-w)P(\beta)}{w\pi(X_j^A,\beta)+(1-w)P(\beta)} \\ &= \prod_{i=1}^{n_1}\frac{\mathrm{hp}(y=1|x_i,\beta)}{\mathrm{hp}(y=1|x_i,\beta)+(1-h)\pi}\prod_{j=1}^{n_0}\frac{(1-h)\pi}{\mathrm{hp}(y=1|x_j,\beta)+(1-h)\pi} \\ &= \prod_{i=1}^{n_1}\frac{(h/\pi)p(y=1|x_i,\beta)}{(h/\pi)p(y=1|x_i,\beta)+(1-h)}\prod_{j=1}^{n_0}\frac{(1-h)}{(h/\pi)p(y=1|x_j,\beta)+(1-h)} \\ &= \prod_i R_{1n}(\beta,\pi,h)^{z_i}(1-R_{1n}(\beta,\pi,h))^{1-z_i} \\ &= L_1(\beta,\pi,h). \end{aligned} \tag{B.2}$$

Therefore, both the Lele (2009) and the LI methods are based on exactly the same likelihood functions.

Appendix C: Statistical mechanism underlying the constrained LK(CLK) method

If we look further using Lagrange multipliers for the proposed CLK method, the Lagrange function is as follows:

$$L(p_i, \lambda) = \sum_{i=1}^{n_1} \{\log p_i - \log \pi_0\} - \lambda \left(\frac{\sum_{i=1}^{n_0} p_i}{n_0} - \pi_0 \right), \tag{C.1}$$

with the constant Lagrange multiplier λ. Calculate the gradient of Eq. C.1 with respect to p_i and λ respectively,

$$\begin{aligned} \nabla_{p_i,\lambda} L(p_i, \lambda) &= \left(\frac{\partial L}{\partial p_i}, \frac{\partial L}{\partial \lambda} \right) \\ &= \left(\frac{n_1}{p_i} - \frac{\lambda}{n_0}, \frac{\sum_{i=1}^{n_0} p_i}{n_0} - \pi_0 \right). \end{aligned}$$

Solving $\nabla_{p_i,\lambda} L(p_i, \lambda) = 0$, shows that the estimate of the probability of presence is equal to the population prevalence, i.e., $\hat{p}_i = \pi_0$.

Likelihood estimates of the SB and the SC methods

The log-likelihood of the SB method (Phillips and Elith 2011) is

$$U(p_i) = \sum_{i=1}^{n_1} \log P_{\mathrm{UA}} + \sum_{i=1}^{n_0} \log(1 - P_{\mathrm{UA}}).$$

Here p_i is the logit function $p_i = \frac{1}{1+\exp(-\eta(x))}$, and $P_{\mathrm{UA}} = \frac{1}{1+r+\exp(-\eta(x)+\ln r)} = \frac{1}{1+r/p_i}$. Taking the derivative of $U(p_i)$ with respect to p_i, setting it to zero, and the score function for p_i becomes $n_0 P_{\mathrm{UA}} = n_1(1 - P_{\mathrm{UA}})$, i.e., $\frac{n_0}{1+r/p_i} = \frac{n_1 r/p_i}{1+r/p_i}$. In Phillips and Elith (2011), r is defined as $r = \frac{1-f_p}{f_p}\pi_0$, which is equivalent to $\frac{n_0}{n_1}\pi_0$, given the sampling proportion of the presence only points $f_p = \frac{n_1}{n_1+n_0}$. It therefore yields $\hat{p}_i = \frac{n_0}{n_1} r = \pi_0$.

As for the SC method (Steinberg and Cardell 1992), the log-likelihood function is as follows:

$$L(p_i) = \frac{1}{n_0} \sum_{i=1}^{n_0} \log(1 - p_i) + \frac{\pi_0}{n_1} \sum_{i=1}^{n_1} \log \frac{p_i}{1 - p_i}.$$

Similarly, solving the score function for p_i leads to the estimate of $\hat{p}_i = \pi_0$.

Therefore, the proposed CLK method obtains the same estimate as the SC (Steinberg and Cardell 1992) and SB (Phillips and Elith 2011) methods for p_i, which are all estimated to be equal to the pre-determined probability of prevalence π_0.

References

Aarts G, Fieberg J, Matthiopoulos J (2012) Comparative interpretation of count, presence-absence and point methods for species distribution models. Methods Ecol Evol 3(1):177–187. https://doi.org/10.1111/j.2041-210X.2011.00141.x

Baddeley A, Berman M, Fisher N, Hardegen A, Milne R, Schuhmacher D, Shah R, Turner R (2010) Spatial logistic regression and change-of-support in Poisson point processes. Electron J Stat 4:1151–1201. https://doi.org/10.1214/10-EJS581

Baddeley A, Rubak E, Turner R (2015) Spatial point patterns: methodology and applications with R. Chapman and Hall/CRC Press

Chakraborty A, Gelfand AE, Wilson AM, Latimer AM, Silander JA (2011) Point pattern modelling for degraded presence-only data over large regions. J R Stat Soc: Ser C: Appl Stat 60(5):757–776. https://doi.org/10.1111/j.1467-9876.2011.00769.x

Cressie N, Wikle CK (2011) Statistics for spatio-temporal data. Wiley, Hoboken

Dorazio RM (2012) Predicting the geographic distribution of a species from presence-only data subject to detection errors. Biometrics 68(4):1303–1312. https://doi.org/10.1111/j.1541-0420.2012.01779.x

Dorazio RM (2014) Accounting for imperfect detection and survey bias in statistical analysis of presence-only data: imperfect detection and survey bias in presence-only data. Glob Ecol Biogeogr 23(12):1472–1484. https://doi.org/10.1111/geb.12216

Drake J, Richards R (2017) Estimating environmental suitability. bioRxiv. https://doi.org/10.1101/109041

Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. J Anim Ecol 77(4):802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Fithian W, Elith J, Hastie T, Keith DA (2015) Bias correction in species distribution models: pooling survey and collection data for multiple species. Methods Ecol Evol 6(4):424–438. https://doi.org/10.1111/2041-210X.12242

Fithian W, Hastie T (2013) Finite-sample equivalence in statistical models for presence-only data. Ann Appl Stat 7(4):1917–1939. https://doi.org/10.1214/13-AOAS667

Guillera-Arroita G, Lahoz-Monfort JJ, Elith J, Gordon A, Kujala H, Lentini PE, McCarthy MA, Tingley R, Wintle BA (2015) Is my species distribution model fit for purpose? Matching data and models to applications. Glob Ecol Biogeogr 24(3):276–292. https://doi.org/10.1111/geb.12268

Guisan A, Edwards TC, Hastie T (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecol Model 157(2):89–100. https://doi.org/10.1016/S0304-3800(02)00204-1

Hastie T, Fithian W (2013) Inference from presence-only data; the ongoing controversy. Ecography 36(8):864–867. https://doi.org/10.1111/j.1600-0587.2013.00321.x

Keating KA, Cherry S (2004) Use and interpretation of logistic regression in habitat-selection studies. J Wildl Manag 68(4):774–789. https://doi.org/10.2193/0022-541X(2004)068[0774:UAIOLR]2.0.CO;2

Koshkina V, Wang Y, Gordon A, Dorazio R, White M, Stone L (2017) Integrated species distribution models: combining presence-background data and site-occupancy data with imperfect detection.

Methods in Ecology and Evolution, pp 420–430. https://doi.org/10.1111/2041-210X.12738
Lancaster T, Imbens GW (1996) Case-control studies with contaminated controls. J Econ 70(1):145–160
Lele SR (2009) A new method for estimation of resource selection probability function. J Wildl Manag 73(1):122–127. https://doi.org/10.2193/2007-535
Lele SR, Keim JT (2006) Weighted distributions and estimation of resource selection probability functions. Ecology 87(12):3021–3028
Lobo JM, Jiménez-Valverde A, Real R (2007) Auc: a misleading measure of the performance of predictive distribution models. Glob Ecol Biogeogr 17(2):145–151. https://doi.org/10.1111/j.1466-8238.2007.00358.x
Møller J, Waagepetersen RP (2003) Statistical inference and simulation for spatial point processes. Chapman and Hall/CRC, Boca Raton
Ovaskainen O, Roy DB, Fox R, Anderson BJ (2016) Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models 7(4), 428–436. https://doi.org/10.1111/2041-210X.12502
Pearce JL, Boyce MS (2006) Modelling distribution and abundance with presence-only data. J Appl Ecol 43(3):405–412. https://doi.org/10.1111/j.1365-2664.2005.01112.x
Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. Ecol Model 190(3-4):231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026
Phillips SJ, Dudík M (2008) Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31(2):161–175
Phillips SJ, Elith J (2011) Logistic methods for resources selection functions and presence-only species distribution models. In: Proceedings of the 25th AAAI conference on artificial intelligence. San Francisco, California, USA, pp 1384–1389
Phillips SJ, Elith J (2013) On estimating probability of presence from use-availability or presence-background data. Ecology 94(6):1409–1419
Renner IW, Elith J, Baddeley A, Fithian W, Hastie T, Phillips SJ, Popovic G, Warton DI (2015) Point process models for presence-only analysis. Methods Ecol Evol 6(4):366–379. https://doi.org/10.1111/2041-210X.12352
Renner IW, Warton DI (2013) Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. Biometrics 69(1):274–281. https://doi.org/10.1111/j.1541-0420.2012.01824.x
Royle JA, Chandler RB, Yackulic C, Nichols JD (2012) Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. Methods Ecol Evol 3(3):545–554. https://doi.org/10.1111/j.2041-210X.2011.00182.x
Solymos P, Lele SR (2016) Revisiting resource selection probability functions and single-visit methods: Clarification and extensions. Methods Ecol Evol 7(2):196–205. https://doi.org/10.1111/2041-210X.12432
Steinberg D, Cardell N (1992) Estimating logistic regression models when the dependent variable has no variance. Commun Stat Theory Methods 21(2):423–450. https://doi.org/10.1080/03610929208830787
Sugiyama M, Suzuki T, Kanamori T (2012) Density ratio estimation in machine learning. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781139035613
Ward G, Hastie T, Barry S, Elith J, Leathwick JR (2009) Presence-only data and the EM algorithm. Biometrics 65(2):554–563. https://doi.org/10.1111/j.1541-0420.2008.01116.x
Warton DI, Blanchet FG, O'Hara RB, Ovaskainen O, Taskinen S, Walker SC, Hui FKC (2015) So many variables: joint modeling in community ecology 30(12), 766–779. https://doi.org/10.1016/j.tree.2015.09.007
Warton DI, Shepherd LC (2010) Poisson point process models solve the pseudo-absence problem for presence-only data in ecology. Ann Appl Stat 4(3):1383–1402. https://doi.org/10.1214/10-AOAS331
Ypma J (2014) R interface to NLopt. The comprehensive R archive network. https://cran.r-project.org/web/packages/nloptr/. Accessed 13 Jul 2017