On the existence of maximum likelihood estimates for presence-only data

Trevor J. Hefley[1,2]* and Mevin B. Hooten[3]

[1]*Department of Statistics and School of Natural Resources, University of Nebraska–Lincoln, 234 Hardin Hall, 3310 Holdrege Street, Lincoln, NE 68583, USA;* [2]*Headwaters Corporation, Platte River Recovery Implementation Program, 4111 4th Ave #6, Kearney, NE 68845, USA; and* [3]*U.S. Geological Survey, Colorado Cooperative Fish and Wildlife Research Unit, Department of Fish, Wildlife, and Conservation Biology, Colorado State University, Fort Collins, CO 80523, USA*

Summary

1. Presence-only data can be used to determine resource selection and estimate a species' distribution. Maximum likelihood is a common parameter estimation method used for species distribution models. Maximum likelihood estimates, however, do not always exist for a commonly used species distribution model – the Poisson point process.

2. We demonstrate the issue with conventional maximum likelihood mathematically, using a data example, and a simulation experiment and show alternative estimation methods.

3. We found that when habitat preferences are strong or the number of presence-only locations is small, by chance, maximum likelihood coefficient estimates for the Poisson point process model may not exist. We found that several alternative estimation methods can produce reliable estimates, but results will depend on the chosen method.

4. It is important to identify conditions for which maximum likelihood estimates are unlikely to be identifiable from presence-only data. In data sets where the maximum likelihood estimates do not exist, penalized likelihood and Bayesian methods will produce coefficient estimates, but these are sensitive to the choice of estimation procedure and prior or penalty term. When sample size is small or it is thought that habitat preferences are strong, we propose a suite of estimation procedures researchers can consider using.

Key-words: Bayesian inference, Firth correction, *Grus americana*, inhomogeneous Poisson point process, MAXENT, maximum likelihood, resource selection, species distribution model, use–availability, whooping crane

Introduction

Management of fish, wildlife, and plants requires knowledge of habitat characteristics that influence the distribution of abundance. Locations where a species occurred are a particularly common source of data. These types of data have been termed use–availability or presence-only data (hereafter referred to as presence-only data; McDonald 2013). Recently, multiple researchers have unified methods for analysing presence-only data by showing that many previously employed methods (e.g. logistic regression, MAXENT, resource selection functions) estimate parameters of an inhomogeneous Poisson point process (IPP) model (Warton & Shepherd 2010; Aarts, Fieberg & Matthiopoulos 2012; Dorazio 2012; Fithian & Hastie 2013; McDonald *et al.* 2013; Renner & Warton 2013; Warton & Aarts 2013). In addition, the IPP model has recently been extended and used to estimate resource selection from telemetry data (Johnson, Hooten & Kuhn 2013). Therefore, it is crucial to understand the properties of the IPP model.

Maximum likelihood is a commonly used parameter estimation method for the IPP model (Baddeley & Turner 2000; Aarts, Fieberg & Matthiopoulos 2012; Fithian & Hastie 2013; Johnson, Hooten & Kuhn 2013). Maximum likelihood has enjoyed great popularity because, given a statistical model, the estimation method is automatic to apply and produces nearly optimal inference (Efron 1986; Cole, Chu & Greenland 2014). However, for some models, maximum likelihood estimates (MLEs) do not always exist. For example, MLEs of coefficients in logistic and Poisson regression models do not exist for certain data configurations (Albert & Anderson 1984; Heinze & Schemper 2002; Owen 2007; Santos Silva & Tenreyro 2010). For logistic regression, Reineking & Schröder (2006) noted: '[t]his happens, for example, when all presence records occur at values larger than a certain threshold for one of the explanatory variables, and all absences occur below that threshold. Standard maximum likelihood estimation fails to converge in these situations, and at least one parameter estimate diverges to infinity'.

If a covariate predicts an outcome perfectly when using logistic regression, then the MLE of the estimated coefficient will be $-\infty$ or ∞ or depending on the data. Although having a

*Correspondence author. E-mail: thefley@huskers.unl.edu

covariate predict a Bernoulli event perfectly is a desirable property, infinite coefficient estimates are difficult to explain. For example, in the case of multiple covariates, comparison of the marginal influence is a challenge when one or more coefficient estimates are infinite. For logistic regression, the issue is known as 'complete separation' and is common in many applications. Because of the prevalence of complete separation, the literature developing and evaluating remedial methods is vast and introductions are given in many textbooks (e.g. Hastie, Tibshirani & Friedman 2009; Gelman *et al.* 2013).

A problem similar to complete separation occurs with the IPP model. This issue could be considered a special case of the results presented by Santos Silva & Tenreyro (2010) and may be a well-understood problem for some regression models (e.g. Poisson and logistic regression). To our knowledge, however, this has not been previously documented and explained for the IPP model and has gone unnoticed in the literature on species distribution modelling and resource selection. We demonstrate some conditions where the MLEs for the IPP model do not exist. We then explore possible solutions to the problem. Finally, we present a simulation experiment and illustrative analysis of location records from a critically endangered species – the whooping crane (*Grus americana*) – which motivated this study.

Materials and methods

THE INHOMOGENIOUS POISSION POINT PROCESS MODEL

The goal of the IPP model is to formulate a statistical link between a species distribution and some number (q) of covariates. Let **Y** be the matrix of n locations and $\mathbf{y}_i$ be a 1×2 vector of the location of each individual in the Euclidian space within region $\mathcal{A}$ where the species occurred. Further, let $\boldsymbol{x}(\mathbf{y}_i)' = \{1, x_1(\mathbf{y}_i),\ldots,x_q(\mathbf{y}_i)\}'$ be a vector of q covariates corresponding to ith location ($\mathbf{y}_i$). An IPP model links the presence-only locations to the covariates by modelling the limiting expected count ($\lambda(\mathbf{y})$) per unit area; $\lambda(\mathbf{y})$ is also referred to as the 'intensity' function). The intensity function is typically, but not always, modelled as a log-linear function of the q covariates:

$$\log(\lambda(\mathbf{y})) = \mathbf{x}(\mathbf{y})'\boldsymbol{\beta},$$

where $\boldsymbol{\beta}$ is a vector of the $q + 1$ model parameters. The log-likelihood of an IPP model is (Cressie 1993):

$$l(\boldsymbol{\beta}; \mathbf{Y}) = \sum_{i=1}^{n} \log \lambda(\mathbf{y}_i) - \int_{\mathbf{y} \in \mathcal{A}} \lambda(\mathbf{y})d\mathbf{y} - \log(n!). \qquad \text{eqn 1}$$

EXISTENCE OF MAXIMUM LIKELIHOOD ESTIMATES

The existence of the MLE for the elements of $\boldsymbol{\beta}$ is dependent on the data configuration (i.e. covariates at the presence-only locations and within the study region). Here, we choose a specific example to show that the MLE does not always exist. Let the intensity be a function of a single binary covariate (i.e. $\log(\lambda(\mathbf{y})) = \beta_0 + \beta_1 x_1(\mathbf{y})$ where $x_1(\mathbf{y}) = 0$ or 1). Assume that the binary covariate (x_1) within the region $\mathcal{A}$ follows a Bernoulli distribution. This assumption is equivalent to assuming that the study area contains two different habitat types. The frequency of the two habitat types within the study region depends on the parameter

p of the Bernoulli distribution. For this example, it is shown in Appendix S1 that

$$\hat{\beta}_0^{\text{MLE}} = \log\left(\frac{n - \sum_{i=1}^{n} x_1(\mathbf{y}_i)}{(1-p)|\mathcal{A}|}\right) \qquad \text{eqn 2}$$

and

$$\hat{\beta}_1^{\text{MLE}} = \log\left(\frac{\sum_{i=1}^{n} x_1(\mathbf{y}_i)}{p|\mathcal{A}|}\right) - \log\left(\frac{n - \sum_{i=1}^{n} x_1(\mathbf{y}_i)}{(1-p)|\mathcal{A}|}\right) \qquad \text{eqn 3}$$

where the superscript denotes maximum likelihood estimator. If by chance $x_1(\mathbf{y}_i) = 1$ for all $\mathbf{y} \in \mathbf{Y}$, then both $\hat{\beta}_0^{\text{MLE}}$ and $\hat{\beta}_1^{\text{MLE}}$ are undefined; as $\sum_{i=1}^{n} x_1(\mathbf{y}_i)$ approaches n, $\hat{\beta}_0^{\text{MLE}}$, diverges to $-\infty$ and $\hat{\beta}_1^{\text{MLE}}$ to ∞. In contrast, if by chance $x_1(\mathbf{y}_i) = 0$ for all $\mathbf{y} \in \mathbf{Y}$, then only $\hat{\beta}_1^{\text{MLE}}$ is undefined and as $\sum_{i=1}^{n} x_1(\mathbf{y}_i)$ approaches zero, $\hat{\beta}_1^{\text{MLE}}$ diverges to $-\infty$.

We have shown the MLE does not exist for a specific case where the covariate in the region $\mathcal{A}$ has a Bernoulli distribution. However, more general results can be borrowed from Owen (2007) relying on the connection of the IPP to infinitely imbalanced logistic regression (Fithian & Hastie 2013). Two important results are the following: (i) the MLE of a coefficient depends on the covariate through the sum of the covariate or equivalently the average value and (ii) to obtain a finite MLE, the average of the covariate must be surrounded by what is 'available' in the study region (Owen 2007). Using the example above to illustrate, it is clear from eqn 3 that the MLE of β_1 depends on $x_1(\mathbf{y}_i)$ at each presence location only through the sum $(\sum_{i=1}^{n} x_1(\mathbf{y}_i))$ or equivalently the average $(1/n \sum_{i=1}^{n} x_1(\mathbf{y}_i) = \overline{x_1(\mathbf{y})})$ of the covariates. Secondly, for the MLE to exist, the average must be bounded such that $0 < \overline{x_1(\mathbf{y})} < 1$. Note that if $\overline{x_1(\mathbf{y})}$ occurs at either end of the support (e.g. 0 or 1 in the example), the MLE of the coefficient will be infinite. This result is general and holds for both continuous and discrete covariates (Owen 2007). Intuitively, this result makes sense because if at all species locations the covariate recorded was the minimum or maximum of what was available, then it would be impossible to estimate how the species distribution changes due to the covariate.

MAXIMUM PENALIZED LIKELIHOOD ESTIMATION

Maximum penalized likelihood estimation (MPLE) has been used as a remedy to complete separation in logistic regression (Reineking & Schröder 2006; Hastie, Tibshirani & Friedman 2009). MPLE has also been widely used for estimation of the parameters in the IPP model (e.g. MAXENT; Phillips, Anderson & Schapire 2006; Renner & Warton 2013). MPLE penalizes the likelihood function for large values of the parameters. For example, the log-likelihood for the IPP model with a binary covariate (see Appendix S1) and a L_2 or ridge penalty is

$$l_p(\beta_0, \beta_1, \gamma_0, \gamma_1; \mathbf{Y}) = \sum_{i=1}^{n}(\beta_0 + \beta_1 x_1(\mathbf{y}_i)) - e^{\beta_0 + \beta_1}p|\mathcal{A}| \\ - e^{\beta_0}(1-p)|\mathcal{A}| - \log(n!) - \gamma_0\beta_0^2 - \gamma_1\beta_1^2. \qquad \text{eqn 4}$$

The penalty terms $\gamma_0\beta_0^2$ and $\gamma_1\beta_1^2$ have been added to the log-likelihood function. The magnitude of the penalty is determined by the tuning parameter γ_j ($0 \le \gamma_j < \infty$) and the square of the coefficient β_j. When using MPLE, it is common to set the tuning parameter for the intercept term to zero (i.e. $\gamma_0 = 0$) and standardize the covariates (Hastie, Tibshirani & Friedman 2009). Standardizing the covariates allows for a common tuning parameter to be applied to all coefficients (see BAYESIAN ESTIMATION for more information on covariate standardization).

Unlike with the standard MLE, we cannot derive closed form solutions (e.g. eqns 2 and 3) for the MPLEs of the IPP model.

In the absence of closed form solution, it is instructive to study the score function for β_1

$$s_p(\beta_1; \mathbf{Y}) = \frac{\partial}{\partial \beta_1} l_p(\beta_0, \beta_1, \gamma_0, \gamma_1; \mathbf{Y}) = \sum_{i=1}^{n} x_1(\mathbf{y}_i) - e^{\beta_0 + \beta_1} p|\mathcal{A}| - 2\gamma_1 \beta_1.$$

eqn 5

Setting $s_p(\hat{\beta}_1; \mathbf{Y}) = 0$ and solving for $\hat{\beta}_1$ yields the MPLE of β_1 (Cole, Chu & Greenland 2014). When $\gamma_1 > 0$, the MPLE is finite for all $\mathbf{y} \in \mathbf{Y}$. The value of $\hat{\beta}_1$ that solves $s_p(\hat{\beta}_1; \mathbf{Y}) = 0$, however, is dependent on γ_1; parameter estimation has been transferred from β_1 to γ_1.

Estimating γ_j can be viewed as a continuous form of model selection and requires out-of-sample data (Hooten & Hobbs 2015). In most cases, a true out-of-sample data set is unavailable and procedures such as cross-validation are used to estimate γ_j. Regardless of the procedure, the goal is to determine the value of γ_j that maximizes the predictive ability of the model. There are many ways to measure predictive ability, but to our knowledge, no measure will produce satisfactory coefficient estimates when the MLE is infinite. For example, the predictive log-likelihood is a commonly used measure of predictive accuracy. This involves calculating $l(\beta_0, \beta_1; \mathbf{Y})$ of eqn 1 using out-of-sample data for a path of values of γ_1. If the out-of-sample data suffers from the same problems as the in-sample data – which it will unless a true replicate out-of-sample data set is available – then the predictive accuracy of the model will monotonically increase as γ_1 goes to zero; when $\gamma_1 = 0$, MPLE are equivalent to the MLE (i.e. $\hat{\beta}_1 = \pm\infty$).

If a true out-of-sample data set is unavailable, it may be impossible to estimate γ_1 using data-driven techniques when the within-sample MLE for β_1 is infinite. Furthermore, it would be more useful to analyse the out-of-sample data assuming the MLE is finite. An alternative solution is to simply choose a value for γ_1. For example, the species distribution modelling software MAXENT (Phillips, Anderson & Schapire 2006) uses an IPP model with a LASSO penalty (e.g. eqn 4 with only a penalty of $\gamma_1|\beta_1|$), with the tuning parameter being 'chosen without any consideration for predictive performance of the model at hand, but rather based entirely on the number of presence cells' (Renner & Warton 2013). At present, there appears to be little guidance on how to choose the regularizing tuning parameter when it cannot be estimated from data.

BAYESIAN ESTIMATION

Even when the MLE does not exist, the data and the likelihood function are informative. Bayesian inference combines the information in the likelihood function and prior information to estimate IPP model parameters and can be used to estimate parameters when the MLE does not exist or is unstable (i.e. sensitive to small changes in the data). Bayesian inference has been applied to logistic regression to obtain coefficient estimates when complete or partial separation occurs (Gelman *et al.* 2008, 2013). In this application, the goal of Bayesian inference is to obtain estimates of coefficients that are finite and stable.

A close connection exists between Bayesian and penalized likelihood estimation methods (Hastie, Tibshirani & Friedman 2009; Cole, Chu & Greenland 2014; Hooten & Hobbs 2015). For example, the posterior mode assuming a normal prior with an expected value of zero ($\beta_j \sim N(0, \sigma_j^2)$) is equivalent to the estimates obtained assuming a ridge penalty when using MPLE (i.e. $\gamma_j = 1/2\sigma_j^2$ in eqn 4); a similar relationship exists for the LASSO penalty (Hastie, Tibshirani & Friedman 2009). As the connection between Bayesian and penalized likelihood estimation methods alludes to, the priors and hyperparameters (σ_j^2) heavily influence coefficient estimates for β_j when the MLEs are infinite.

Although coefficient estimates are heavily influenced by the priors when the MLEs are infinite, Bayesian inference differs from penalized likelihood methods in two important ways: (i) guidance on how to choose the hyperparameters for the prior, or equivalently the penalty γ_j term, is available, and (ii) inference is based on Bayes theorem which provides a formal probability-based tool for constraining models. When using MPLE, the tuning parameter γ_j lacks interpretation as prior information and if forced to choose a value of γ_j, it is not clear when or how this should be done. In contrast, the Bayesian perspective makes it clear that prior information is constraining the model parameters. Furthermore, to maintain acyclicity in the Bayesian 'graph', the only rule in choosing priors is not to use in-sample data (Hooten & Hobbs 2015); in the case that out-of-sample data are unavailable and cross-validation is going to fail, priors should be chosen before analysing the data.

Choosing priors before analysing the data may be difficult in practice. We suspect that most practitioners will estimate parameters using maximum likelihood and decide to use Bayesian estimation only if they encounter infinite estimates. An alternative to choosing a specific prior before analysing the data is to use a so-called weakly informative default prior (Gelman *et al.* 2008). In general, a weakly informative prior will produce parameter estimates that are similar to MLEs when the MLEs are finite and stable, but will result in sensible estimates when the MLEs are infinite or unstable. Many default priors have been proposed for logistic regression which could be applied to IPP model (e.g. Gelman *et al.* 2008).

An important consideration when applying a weakly informative default prior or penalized likelihood estimation is standardization of the covariates. We refer the reader to Gelman *et al.* (2008) for an excellent discussion on the topic for logistic regression which can be applied to the IPP model. We note, however, that standardization of covariates for the presence-only locations may present a problem when the MLE is infinite. When the MLE is infinite, the standard deviation of the covariates at the presence-only locations may be equal to zero. If standardization is required, we recommend that covariates be standardized to the scale of the study area or scale of the presence-only location and the study area (i.e. covariates obtained from the quadrature or Monte Carlo locations used to approximate the integral in eqn 1).

FIRTH BIAS CORRECTION

Firth (1993) proposed a method to remove the first-order term from the asymptotic bias in MLEs; infinite coefficient estimates are an extreme case of bias, and the Firth correction has been proposed as a solution for complete separation in logistic regression (Heinze & Schemper 2002). The Firth correction can be cast as a MPLE or from a Bayesian perspective. The Firth correction involves adding the penalty term $1/2\log(|I(\beta)|)$ to the log-likelihood (eqn 1) where $I(\beta)$ is the information matrix. Under a Bayesian paradigm, the penalty term is equivalent to assigning a Jeffreys invariant prior to all model parameters (Firth 1993; Gelman *et al.* 2013). For our example, we derived closed form solutions and show (Appendix S1) that the MPLE or mode of the posterior distribution is

$$\hat{\beta}_0^{\text{Firth}} = \log\left(\frac{n - \sum_{i=1}^{n} x_1(\mathbf{y}_i) + \frac{1}{2}}{(1-p)|\mathcal{A}|}\right)$$

eqn 6

and

$$\hat{\beta}_1^{\text{Firth}} = \log\left(\frac{\sum_{i=1}^{n} x_1(\mathbf{y}_i) + \frac{1}{2}}{p.|\mathcal{A}|}\right) - \log\left(\frac{n - \sum_{i=1}^{n} x_1(\mathbf{y}_i) + \frac{1}{2}}{(1-p)|\mathcal{A}|}\right)$$

eqn 7

where the superscript denotes that estimator is the Firth type. Clearly, $\hat{\beta}_0^{\text{Firth}}$ and $\hat{\beta}_1^{\text{Firth}}$ exist and are finite for all values of the covariate $x_1(\mathbf{y}_i)$

for all $\mathbf{y} \in \mathbf{Y}$. Unlike the penalty term for MPLE or equivalent priors for Bayesian inference, the Firth bias correction (or Jefferys invariant prior) has no tuning parameter (or hyperparameter) that must be chosen. In addition, the Firth bias correction for the logistic regression model is invariant to linear transformations in the covariates (Chen, Ibrahim & Kim 2009). As a result, linear transformations, such as standardization of the covariates, will not influence inference about the coefficients.

WHOOPING CRANE DATA

Whooping cranes are an endangered migratory avian species occurring in captivity and a single self-sustaining wild population that currently totals 250–300 individuals. This population overwinters in and around Aransas National Wildlife Refuge in southern Texas, USA, and nests during the summer in and around Wood Buffalo National Park of Canada. Each fall and spring, whooping cranes migrate approximately 4000 km as individuals or in small groups. These migrations include several stopovers that may last from a few hours for several weeks. Such stopovers during migration provide much needed energy recuperation and are critical to the survival of whooping cranes. Management of migratory habitat is a focus of a multistate–federal cooperative agreement focused on the Central Platte River Valley in Nebraska, USA (Freeman 2010). A basic question that must be answered to facilitate management is to determine whether whooping cranes avoid areas where human disturbances occur, and if so, at what scale. Also of interest is the influence of habitat covariates.

The data used in this analysis were described in more detail by Hefley *et al.* (2014) and are available on Dryad Digital Repository (Hefley *et al.* 2013). The data consist of 32 whooping crane group locations recorded in the state of Nebraska, USA, from 2000 to 2012. At each location, the proportion of aquatic habitat and development within a 100-m-radius buffer was calculated (hereafter referred to as the aquatic and development covariates). For the 32 crane group locations, the development covariate consisted of 31 zeroes and a single value of 0·086 which resulted in an average proportion of development of 0·0027 (SD = 0·0151). For comparison, the average proportion of development at 10 000 random locations within the state of Nebraska was 0·035 (SD = 0·112; range = [0,1]). The aquatic covariate averaged 0·396 (SD = 0·432; range = [0,1]). At 10 000 random locations within the state of Nebraska, the average proportion of aquatic habitat was 0·025 (SD = 0·117; range = [0,1]). The correlation between the two covariates at the 32 presence locations was −0·06, and the correlation at the 10 000 random locations was −0·03.

WHOOPING CRANE DATA ANALYSIS

We used the IPP model to analyze the 32 whooping crane presence-only locations with intensity function

$$\log(\lambda(\mathbf{y})) = \beta_0 + \beta_1 \texttt{aquatic} + \beta_2 \texttt{development} \qquad \text{eqn 8}$$

where $\texttt{aquatic}$ and $\texttt{development}$ are the proportion of aquatic habitat and development within a 100-m buffer, respectively. We estimated the β_j using maximum likelihood, maximum penalized likelihood, Bayesian, and Firth-type estimation. For Bayesian estimation, we used the weakly informative default prior proposed by Gelman *et al.* (2008), which was a Cauchy distribution with centre of 0 and scale of 10 and 2·5 for the intercept and covariates, respectively (with scaling as described in Gelman *et al.* (2008)).

Bayesian estimation with a Cauchy prior is equivalent to MPLE using a penalty proportional to the Cauchy likelihood and appropriate tuning parameters (i.e. $\gamma_j = 1/\sigma_j$, where σ_j is the scale parameter of the Cauchy distribution; Gelman *et al.* 2013); hereafter, we refer to the posterior mode of β_j assuming a Cauchy prior or the equivalent MPLE as $\hat{\beta}_j^{\text{Bayes/MPLE}}$.

Several authors have demonstrated that modified generalized linear models can be used to obtain IPP model coefficient estimates (Berman & Turner 1992; Baddeley & Turner 2000; Aarts, Fieberg & Matthiopoulos 2012; Fithian & Hastie 2013; McDonald 2013). The modifications involve the so-called infinite weights, or alternatively McDonald (2013) shows that logistic regression is just a trick to maximize the IPP likelihood under certain assumptions. We rely on the results of Fithian & Hastie (2013) and use infinitely weighted logistic regression to obtain coefficient estimates and 95% asymptotic confidence or credible intervals (CIs). Using infinitely weighted logistic regression, $\hat{\beta}_j^{\text{MLE}}$, $\hat{\beta}_j^{\text{Bayes/MPLE}}$, and $\hat{\beta}_j^{\text{Firth}}$ can be estimated using the $\texttt{glm}(\ldots)$, $\texttt{bayesglm}(\ldots)$, and $\texttt{logistf}(\ldots)$ functions, respectively, in R (R Core Team 2014). A tutorial with R code implementing the whooping crane data analysis and reproducing Table 1 and Figs 1 and S1 appears in Appendix S2.

As in Hefley *et al.* (2014), we approximated the integral in the likelihood (eqn 1) using the Monte Carlo method with values of the aquatic and development covariates from 10 000 random locations within the state of Nebraska. The number of random locations used to approximate the integral was determined to be adequate by observing the MLE for the aquatic ($\hat{\beta}_1^{\text{MLE}}$) and development ($\hat{\beta}_2^{\text{MLE}}$) coefficient using 100 random locations and increasing the number of random locations in step sizes of 10 (Northrup *et al.* 2013). Both $\hat{\beta}_1^{\text{MLE}}$ and $\hat{\beta}_2^{\text{MLE}}$ stabilized when at least 6000 random locations were used. We used 10 000 random locations in all analyses to be conservative (Fig. S1).

According to the results of Owen (2007) and theory presented in this study, we should expect that $\hat{\beta}_2^{\text{MLE}}$ would be infinite if the data analysed were the 31 whooping crane locations where the development covariate was zero. An infinite estimate of $\hat{\beta}_2^{\text{MLE}}$ should occur because the

Table 1. Coefficient estimates and 95% asymptotic confidence or credible interval (CI) for the aquatic (β_1) and development (β_2) covariate of an inhomogeneous Poisson point process model used to analyse presence-only whooping crane location records. Point estimates were obtained using maximum likelihood ($\hat{\beta}_j^{\text{MLE}}$), Bayesian/maximum penalized likelihood ($\hat{\beta}_j^{\text{Bayes/MPLE}}$), and Firth-type ($\hat{\beta}_j^{\text{Firth}}$) estimation. The coefficients were estimated using the full data set ($n = 32$) in which the development covariate was a collation of 31 zeroes and a single value of 0·086 and a reduced data set ($n = 31$) where the development covariate was a collation of 31 zeroes

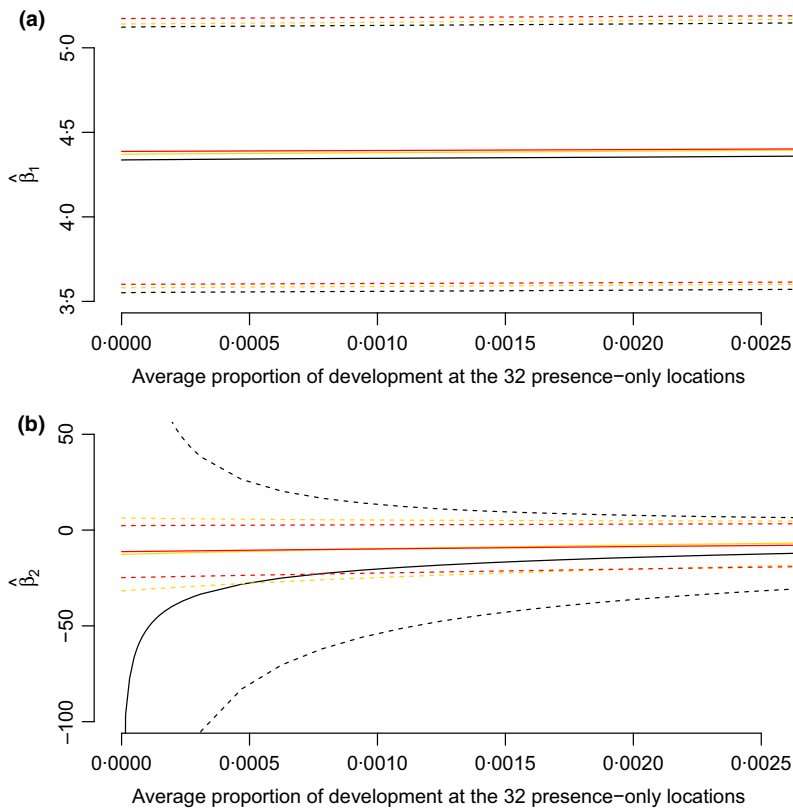| Parameter	n	Point estimate	Lower CI	Upper CI
$\hat{\beta}_1^{\text{MLE}}$	32	4·36	3·57	5·15
$\hat{\beta}_1^{\text{Bayes/MPLE}}$	32	4·36	3·57	5·15
$\hat{\beta}_1^{\text{Firth}}$	32	4·38	3·60	5·15
$\hat{\beta}_1^{\text{MLE}}$	31	4·32	3·52	5·12
$\hat{\beta}_1^{\text{Bayes/MPLE}}$	31	4·33	3·53	5·13
$\hat{\beta}_1^{\text{Firth}}$	31	4·33	3·55	5·12
$\hat{\beta}_2^{\text{MLE}}$	32	−11·98	−30·41	6·44
$\hat{\beta}_2^{\text{Bayes/MPLE}}$	32	−7·87	−19·03	3·30
$\hat{\beta}_2^{\text{Firth}}$	32	−6·73	−18·18	4·73
$\hat{\beta}_2^{\text{MLE}}$	31	−∞	−∞	∞
$\hat{\beta}_2^{\text{Bayes/MPLE}}$	31	−11·13	−24·72	2·46
$\hat{\beta}_2^{\text{Firth}}$	31	−12·45	−31·35	6·45

(a)

(b)

Fig. 1. Maximum likelihood estimates (black lines), Bayesian/maximum penalized likelihood estimates (red lines), and Firth-type estimates (gold lines) and 95% confidence intervals (dashed lines) for the aquatic ($\hat{\beta}_1$) and development ($\hat{\beta}_2$) coefficients obtained from the full ($n = 32$) presence-only location records of whooping cranes. The single non-zero value of the development covariate was decreased from the measured value of 0·086 to zero; therefore, the average of the development covariate decreased from 0·0027 to zero. The black dots show the maximum likelihood estimates obtained from the unaltered data set. Note, lines were shifted slightly to not overlap.

average proportion of development for these 31 locations is exactly zero and lies on the boundary of what is available (i.e. the development covariate must be ≥0). To test this theory for a continuous covariate and to explore how the other estimation methods preformed, we analyzed the 31 whooping crane locations where the development covariate was zero using the same methods as the full ($n = 32$) data set. To examine the sensitivity of coefficient estimates to small changes in the covariates, we estimated the aquatic and development coefficients and 95% CIs using all methods for the full ($n = 32$) data set; however, we incrementally varied the value of the single nonzero development covariate from its measured value of 0·086 to zero so that when plotted against the average value of the development covariate the relationship appeared continuous (Fig. 1).

SIMULATION STUDY

We conducted a simulation experiment to better understand the properties of the maximum likelihood, maximum penalized likelihood, Bayesian, and Firth-type estimators for the IPP model. We simulated presence-only locations (**Y**) from an IPP distribution. The intensity function ($\lambda(\mathbf{y})$) depended on a single discrete covariate $x_1(\mathbf{y})$. The environment in which the species could have occurred in (x_1) was generated from a Bernoulli distribution with $p = 0.5$. When $p = 0.5$, habitat type '1' is expected to be equally common within $\mathcal{A}$ as habitat type '0' (here $|\mathcal{A}| = 1$). We conducted the simulation study using the following three scenarios: 1). $\beta_0 = 4$ and $\beta_1 = -4$; 2). $\beta_0 = 8$ and $\beta_1 = -8$; and 3). $\beta_0 = 8$ and $\beta_1 = -4$. The first scenario was designed to test the properties of the estimator when sample size was small and habitat preferences were (relatively) weak resulting in MLEs that are infinite for most data sets. The second scenario was designed to test the properties of the estimators when sample size was large and habitat preferences were strong. Similar to the first scenario, this should result in MLEs that are infinite for most data sets. The third scenario was designed to test the

properties of the estimators when sample size was large and habitat preferences were relatively weak and should result in MLEs that are finite for nearly all data sets. We generated 1000 simulated data sets for each scenario.

We used the same estimation procedures to obtain point estimates and 95% CIs as was used for the whooping crane data analysis. We expect researchers to be concerned about the uncertainty of the coefficient estimates and would expect that the coverage probabilities of the CIs are close to the nominal level. To test this, we reported the coverage probability and averaged CI length of β_1 for each scenario. Annotated R code implementing the simulation study and capable of reproducing all simulation results and Fig. 2 is in Appendix S3.

Results

WHOOPING CRANE DATA ANALYSIS

The estimated coefficient for the aquatic covariates and associated 95% CIs were nearly equal for all estimation methods using the full ($n = 32$) and reduced ($n = 31$) data sets (Table 1). For the full data set, the estimated coefficient and 95% CIs for the development covariates were about 40% less in magnitude and shorter in length using Bayesian, maximum penalized likelihood, and Firth-type estimation when compared to the MLE and CIs. When the reduced data set was analysed, the MLE for the development coefficient was $-\infty$ and the Bayesian, MPLE, and Firth-type estimates were finite and close to the MLE obtained from the full data set (Table 1).

Varying the development covariate had minimal effect on $\hat{\beta}_1^{\text{MLE}}$, $\hat{\beta}_1^{\text{Bayes/MPLE}}$, $\hat{\beta}_1^{\text{Firth}}$, and associated 95% CIs (Fig. 1a). When the single nonzero value of the development covariate
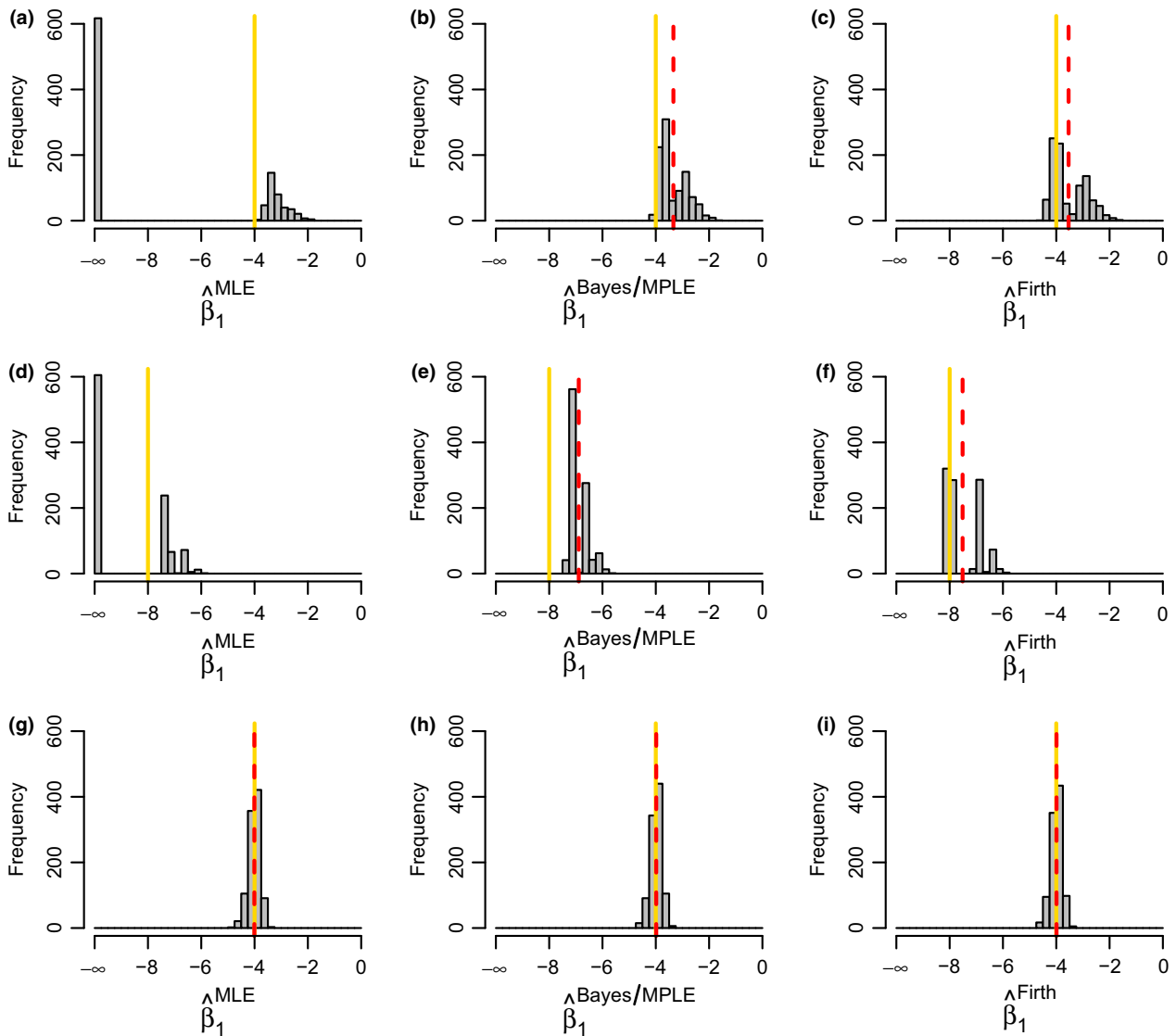
Fig. 2. Results from 1000 simulated presence-only data sets under three different scenarios. Panels show the maximum likelihood estimates ($\hat{\beta}_1^{MLE}$), Bayesian/maximum penalized likelihood estimates ($\hat{\beta}_1^{Bayes/MPLE}$), and Firth-type estimates ($\hat{\beta}_1^{Firth}$) for the regression coefficient of the inhomogeneous Poisson point process model. Scenario 1 (a,b,c) and scenario 2 (d,e,f) were designed to test the estimation procedures when the MLE was infinite for most simulated data sets. The third scenario (g,h,i) was designed to test the estimation procedures when the MLE was finite for all simulated data sets. The average number of presence-only locations in each data set was 28, 1491, and 1517 for scenarios 1, 2, and 3, respectively. The vertical gold line shows the true value of the coefficients, and the red line shows the average of the 1000 estimates shown in each histogram. Note, the lower mode in b, c, e, and f coincides with simulations for which the MLE was undefined.

was decreased from its measured value of 0·086 to zero, $\hat{\beta}_2^{MLE}$ and the lower 95% CI limit diverged nonlinearly to $-\infty$ and the upper CI limit diverged to ∞ (Fig. 1b). The $\hat{\beta}_2^{Bayes/MPLE}$ and $\hat{\beta}_2^{Firth}$ always existed, but decreased and 95% CIs widened as the development covariate decreased. The magnitude of the decrease in $\hat{\beta}_2^{Bayes/MPLE}$, $\hat{\beta}_2^{Firth}$, and widening of the CIs was small in comparison with that of $\hat{\beta}_2^{MLE}$ (Fig. 1b).

SIMULATION STUDY

The average number of presence-only locations in each data set was 28, 1491, and 1517 for scenarios 1, 2, and 3, respectively. The $\hat{\beta}_1^{MLE} = -\infty$ in >60% of the simulation data sets

under the first ($\beta_0 = 4$ and $\beta_1 = -4$) and second ($\beta_0 = 8$ and $\beta_1 = -8$) scenarios. In the first and second scenarios, when the $\hat{\beta}_1^{MLE}$ was finite, the estimates were always greater than the true value (Fig. 2a, d). The $\hat{\beta}_1^{Bayes/MPLE}$ and $\hat{\beta}_1^{Firth}$ were always finite and reasonably close to the true value (Fig. 2b,c,e,f). The $\hat{\beta}_1^{MLE}$, $\hat{\beta}_1^{Bayes/MPLE}$, and $\hat{\beta}_1^{Firth}$, in the third scenario ($\beta_0 = 8$ and $\beta_1 = -4$), were always finite and nearly identical (Fig. 2g–i). Coverage probabilities of the 95% CIs for all three estimators were similar and between 0·91 and 0·96 for all three scenarios, except for the second scenario where the CI for $\hat{\beta}_1^{Bayes/MPLE}$ had a coverage probability of 0·61 (see Appendix S3 for detailed results). The average length of the 95% CIs varied greatly for the first and second scenarios. The first scenario had

an average 95% CI length of ∞ for $\hat{\beta}_1^{\text{MLE}}$, 3·85 for $\hat{\beta}_1^{\text{Bayes/MPLE}}$, and 4·65 for $\hat{\beta}_1^{\text{Firth}}$. The second scenario had an average 95% CI length of ∞ for $\hat{\beta}_1^{\text{MLE}}$, 3·06 for $\hat{\beta}_1^{\text{Bayes/MPLE}}$, and 4·55 for $\hat{\beta}_1^{\text{Firth}}$. For the third scenario, all three estimation procedures had an average 95% CI length between 0·75 and 0·77.

Discussion

MAXIMUM LIKELIHOOD ESTIMATION

We demonstrated the nonexistence of the MLE for special cases of the IPP model mathematically and by analyzing real and simulated data. We found that MLEs were infinite when the average value of the covariates for all locations occurs at either end of the range of what is available – regardless of the number of presence-only locations. This result agrees with that of Owen (2007), who proved the result for infinitely imbalanced logistic regression, which can be used to obtain MLEs of coefficients for the IPP model (Fithian & Hastie 2013).

In practice, we expect a species will not absolutely avoid a particular value of a covariate within the study area. If this is true, the data configurations that result in infinite MLEs are most likely to occur when sample size is small; by chance, a data set with only a small number of locations has a higher probability of containing only locations that occur at the extreme value of the covariate. Another situation where this problem is likely to occur is when the effect of the covariate is large. For example, if the 'true' value of the coefficient is large, it is more likely that all of the locations will occur in areas that contain extreme values of the covariate.

We are aware of another setting where the MLE will be non-identifiable for the IPP model. This occurs when two or more covariates (or some linear combination) obtained at the presence-only locations are perfectly correlated (Santos Silva & Tenreyro 2010). The issue of collinearity is well known, and remedial measures are equally well described (Dormann *et al.* 2013). For presence-only data, it is important to examine correlation in the covariates obtained at the presence-only locations separate from the 'background' or 'available' locations (i.e. the quadrature or Monte Carlo locations) to determine whether collinearity will be an issue.

ALTERNATIVE ESTIMATION METHODS

We considered maximum penalized likelihood, Bayesian, and Firth-type estimation as alternative methods that can be used when the MLE is infinite (or nearly so). The weakly informative default prior proposed by Gelman *et al.* (2008) and the bias correction proposed by Firth (1993), provided nearly identical results when applied to our whooping crane example (Table 1) and scenarios 1 and 3 of the simulation (Fig. 2). It could be argued that the Firth-type estimator outperformed the maximum likelihood and Bayesian estimators in scenario 2 (Fig. 2d,e,f and in-text results). Although this is the case, we argue that it is difficult to make general recommendations as to when one estimation method is preferable; it will depend on the specific problem (e.g. are finite coefficients desired?); what type of prior information is available (e.g. is expert elicitation feasible?); how this information is transformed into a penalty or prior that can be used to constrain the coefficient estimates; and finally, what the 'true' values of the coefficients are and how well this is reflected by the choice of priors or penalties.

The variety of potential remedies for presence-only data and the IPP model echoes the lack of consensus for how to handle the problem in more general settings (e.g. Firth 1993; Heinze & Schemper 2002; Gelman *et al.* 2008; Hastie, Tibshirani & Friedman 2009). Methods, such as those presented by Firth (1993), certainly have more justification when little or no prior information is available. Regardless of the method employed, the coefficient estimate will be sensitive to the choice of prior or penalty when the MLE is infinite. Although this may be troublesome, it is encouraging to note that the issue, in general, has been successfully addressed in many applied studies that employed logistic or Poisson regression. It is straightforward to check for similar issues in ecological studies that rely on point process data. Armed with the set of tools we described herein, ecologists can then ameliorate the problem of infinite MLEs in a way that is best matched with the goals of their study.

Data accessibility

The source data for these analyses are archived in the Dryad Digital Repository (Hefley *et al.* 2013).

References

Aarts, G., Fieberg, J. & Matthiopoulos, J. (2012) Comparative interpretation of count, presence-absence and point methods for species distribution models. *Methods in Ecology and Evolution*, **3**, 177–187.

Albert, A. & Anderson, J.A. (1984) On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, **71**, 1–10.

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

Berman, M. & Turner, T. (1992). Approximating point process likelihoods with GLIM. *Journal of the Royal Statistical Society Series C Applied Statistics*, **41**, 31–38.

Chen, M., Ibrahim, J.G. & Kim, S. (2009) Properties and implementation of Jeffreys's prior in binomial regression models. *Journal of the American Statistical Association*, **103**, 1659–1664.

Cole, S.R., Chu, H. & Greenland, S. (2014) Maximum likelihood, profile likelihood, and penalized likelihood: a primer. *American Journal of Epidemiology*, **179**, 252–260.

Cressie, N. (1993) *Statistics for Spatial Data*. Wiley, New York, NY, USA.

Dorazio, R. (2012) Predicting the geographic distribution of a species from presence-only data subject to detection errors. *Biometrics*, **68**, 1–25.

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G. *et al.* (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, **36**, 27–46.

Efron, B. (1986) Why isn't everyone Bayesian. *American Statistician*, **40**, 1–11.

Firth, D. (1993) Bias reduction of maximum likelihood estimates. *Biometrika*, **80**, 27–38.

Fithian, W. & Hastie, T. (2013) Finite-sample equivalence in statistical models for presence only data. *Annals of Applied Statistics*, **7**, 1917–1939.

Freeman, D.M. (2010) *Implementing the Endangered Species Act on the Platte Basin Water Commons*. University Press of Colorado, Boulder, CO, USA.

Gelman, A., Jakulin, A., Pittau, M.G. & Su, Y.-S. (2008) A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, **2**, 1360–1383.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. & Rubin, D.B. (2013) *Bayesian Data Analysis*, 3rd edn. Chapman and Hall/CRC, Boca Raton, FL, USA.

Hastie, T., Tibshirani, R. & Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edn. Springer, New York, NY, USA.

Hefley, T.J., Baasch, D.M., Tyre, A.J. & Blankenship, E.E. (2013) Data from: Correction of location errors for presence-only species distribution models. *Dryad Digital Repository*, doi:10.5061/dryad.h81s5.

Hefley, T.J., Baasch, D.M., Tyre, A.J. & Blankenship, E.E. (2014) Correction of location errors for presence-only species distribution models. *Methods in Ecology and Evolution*, **5**, 207–214.

Heinze, G. & Schemper, M. (2002) A solution to the problem of separation in logistic regression. *Statistics in Medicine*, **21**, 2409–2419.

Hooten, M.B. & Hobbs, N.T. (2015). A guide to Bayesian model selection for ecologists. *Ecological Monographs*, doi: 10.1890/14-0661.1.

Johnson, D.S., Hooten, M.B. & Kuhn, C.E. (2013) Estimating animal resource selection from telemetry data using point process models. *Journal of Animal Ecology*, **82**, 1155–1164.

McDonald, T.L. (2013) The point process use-availability or presence-only likelihood and comments on analysis. *Journal of Animal Ecology*, **82**, 1174–1182.

McDonald, L.L., Manly, B.F., Huettmann, F. & Thogmartin, W.E. (2013) Location-only and use-availability data: analysis methods converge. *Journal of Animal Ecology*, **82**, 1120–1124.

Northrup, J., Hooten, M.B., Anderson, C. & Wittemyer, G. (2013) Practical guidance on characterizing availability in resource selection function under use-availability design. *Ecology*, **94**, 1456–1463.

Owen, A.B. (2007) Infinitely imbalanced logistic regression. *Journal of Machine Learning Research*, **8**, 761–773.

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

R Core Team. (2014). R: a language and environment for statistical computing. Retrieved from http://www.r-project.org/

Reineking, B. & Schröder, B. (2006) Constrain to perform: regularization of habitat models. *Ecological Modelling*, **193**, 675–690.

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.

Santos Silva, J.M.C. & Tenreyro, S. (2010) On the existence of the maximum likelihood estimates in Poisson regression. *Economics Letters*, **107**, 310–312.

Warton, D. & Aarts, G. (2013) Advancing our thinking in presence-only and used-available analysis. *Journal of Animal Ecology*, **82**, 1125–1134.

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.

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1. Analytical results.

Appendix S2. Tutorial with data and R code for the analysis of the locations records of whooping cranes and to reproduce Table 1, Figs 1, and S1.

Appendix S3. Tutorial with R code for the simulation experiment and to reproduce Fig. 2.