2013

Correction of location errors for presence-only species distribution models

Trevor Hefley
University of Nebraska at Lincoln, trevorhefley@msn.com

David M. Baasch
University of Nebraska - Lincoln, baaschd@headwaterscorp.com

Andrew J. Tyre
University of Nebraska at Lincoln, atyre2@unl.edu

Erin E. Blankenship
University of Nebraska-Lincoln, erin.blankenship@unl.edu

Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2013 doi: 10.1111/2041-210X.12144

Correction of location errors for presence-only species distribution models

Trevor J. Hefley[1]*, David M. Baasch[2], Andrew J. Tyre[3] and Erin E. Blankenship[4]

[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, 4111 4th Avenue, Suite 6, Kearney, NE 68845, USA; [3]School of Natural Resources, University of Nebraska–Lincoln, 416 Hardin Hall, 3310 Holdrege Street, Lincoln, NE 68583, USA; and [4]Department of Statistics, University of Nebraska–Lincoln, 343B Hardin Hall North, 3310 Holdrege Street, Lincoln, NE 68583, USA

Summary

1. Species distribution models (SDMs) for presence-only data depend on accurate and precise measurements of geographical and environmental covariates that influence presence and abundance of the species. Some data sets, however, may contain both systematic and random errors in the recorded location of the species. Environmental covariates at the recorded location may differ from those at the true location and result in biased parameter estimates and predictions from SDMs.

2. Regression calibration is a well-developed statistical method that can be used to correct the bias in estimated coefficients and predictions from SDMs when the recorded geographical location differs from the true location for some, but not all locations. We expand the application of regression calibration methods to SDMs and provide illustrative examples using simulated data and opportunistic records of whooping cranes (*Grus americana*).

3. We found we were able to successfully correct the bias in our SDM parameters estimated from simulated data and opportunistic records of whooping cranes using regression calibration.

4. When modelling species distributions with data that have geographical location errors, we recommend researchers consider the effect of location errors. Correcting for location errors requires that at least a portion of the data have locations recorded without error. Bias correction can result in an increase in variance; this increase in variance should be considered when evaluating the utility of bias correction.

Key-words: *Grus americana*, inhomogeneous Poisson point process, location errors, measurement error, Nebraska, opportunistic sightings, Public Land Survey System, regression calibration, whooping crane

Introduction

A prerequisite to successful management of fish and wildlife populations is determining environmental features that influence presence and population abundance. To answer this question, ecologists, statisticians and computer scientists have developed an impressive array of sampling methods and statistical tools (Manly *et al.* 2002; Tyre *et al.* 2003; Pearce & Boyce 2006; Phillips, Anderson & Schapire 2006; Elith & Leathwick 2009); however, rare or locally extinct species present a challenge because feasible sampling protocols would produce few, if any, records of presence. An alternative approach involves the analysis of presence-only records that are collected opportunistically. Opportunistic presence-only records are accounts of where a species occurred that, in general, are collected haphazardly (e.g. museum records) or lack information on sampling effort (Elith & Leathwick 2007). For example, the United States Fish and Wildlife Service (USFWS) has constructed and maintained a data base containing locations of all confirmed sightings of whooping cranes (*Grus americana*), a critically endangered species in North America (Austin & Richert 2001). Whooping cranes are one of the rarest avian species, and a large proportion of sightings are not obtained from research efforts, but rather are reported by members of the public.

Recently, multiple authors have unified methods for analysing presence-only data by showing that many previously developed methods (e.g. MAXENT, logistic regression) are approximating an inhomogeneous Poisson point process model (IPP; Warton & Shepherd 2010; Dorazio 2012; Fithian & Hastie 2013; Renner & Warton 2013). This unification, and future extensions using the IPP, will reduce confusion within and between statisticians and ecologists. Limitations to the analysis of presence-only data, such as sampling bias and errors in location records, however, still exist. Sampling bias has received much attention (Araújo & Guisan 2006; Phillips *et al.* 2009; Dorazio 2012; Hefley *et al.* 2013; Kramer-Schadt

*Correspondence author. E-mail: thefley@huskers.unl.edu *et al.* 2013; Monk 2013); however, little has been done to account for and correct the bias introduced by errors in location records (Graham *et al.* 2007).

Error in location occurs when the recorded geographical location is different from the true location. For studies using radio or global position system (GPS) telemetry, the effects of errors in location have been acknowledged, but are typically ignored because the tracking technology used to collect the data provides precision much greater than the environmental and geographical scales of interest (Montgomery *et al.* 2010; Montgomery, Roloff & Hoef 2011). Although there is no single natural scale at which species' distribution patterns should be studied, ideally the appropriate scale would be dictated by the goals of the study and knowledge of the species and not by the quality of the data (Bradter *et al.* 2013). For opportunistically collected presence-only data, however, the imprecision of location records may be of concern because the errors in location can be large compared with the scales of interest (Barry & Elith 2006). Most often, presence-only records are used with a geographical information system to derive environmental covariates that are assumed to influence species' presence and abundance. Imprecise location records, however, can result in covariates at the recorded location that are different from those at the true location. In general, errors in location can result in biased predictions and estimates of SDM coefficients when the location error is large compared with the scale of environmental and geographical covariates. We explore the effects of location errors on regression coefficient estimates obtained from SDMs using simulated and real data and offer a remedial method for analysing records such as opportunistic sightings of the whooping crane.

Materials and methods

WHOOPING CRANE DATA

Whooping cranes are an endangered migratory avian species that occur in a single self-sustaining wild population that currently totals 200–300 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 to several weeks. Such stopovers during migration provide much needed rest and food and are critical to the survival of whooping cranes. Restoration and preservation of migratory habitat has been a focus of a multistate, federal cooperative agreement focused on the central Platte River Valley in Nebraska, USA (Freeman 2010). A prerequisite for successful habitat restoration and preservation along the central Platte River Valley is determining environmental conditions that influence the distribution of whooping cranes during migration.

Opportunistic sightings have been recorded by the USFWS since 1943 for the state of Nebraska, USA (Austin & Richert 2001). The accuracy of the recorded locations of the opportunistic sightings, however, is highly variable. Some of the locations have near perfect geographical location obtained with a GPS. Other locations were identified according to the Public Land Survey System (PLSS) at the section level, which identifies the location of the crane group as the centre of a 2·59 km^2 area (Fig. 1).

We performed two analyses. For the first analysis, we used all crane groups reported opportunistically from 2000 to 2012 when the birds were not flying and had a recorded location that was obtained with a GPS. This resulted in a total of 32 crane group locations. For this sample, we assumed the locations were measured perfectly or that the error in locations was minimal and ignorable. We derived environmental covariates from the 2006 National Landover Cover Dataset (NLCD; Fry *et al.* 2011). We constructed 100-, 250- and 500-m-radius buffers around each crane group and calculated the proportion of aquatic habitat (amalgamation of land class 90 and 95) and development (amalgamation of land class 21, 22, 23 and 24) within each buffer. We chose three buffer sizes to allow for a range of measurement error, because we expected the magnitude of the bias in coefficient estimates to be positively related to the amount of measurement error and, hence, inversely related to the size of the buffer. We chose two environmental covariates based on *a priori* knowledge that a majority of whooping crane observations occurred in or near aquatic habitats and whooping cranes may be sensitive to developed area. For the second analysis, we modelled all observations from 2000 to 2012 that were obtained when the birds were not flying and had location recorded with a GPS or location accuracies listed as a PLSS section. This resulted in a total of 68 crane group location records.

We do not contend that any part of the analysis presented here is a complete or comprehensive representation of factors that influence the distribution of whooping crane groups. In particular, the data used in our analysis are appropriate to model apparent species' distribution, not the true species' distribution as we did not attempt to correct for

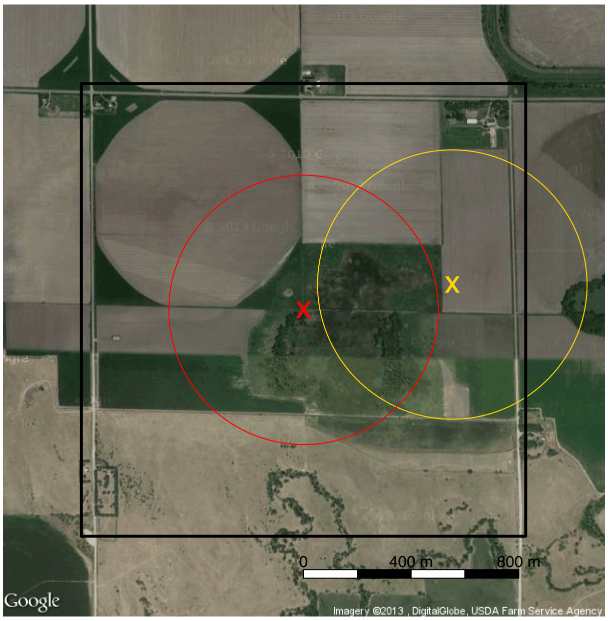

Fig. 1. Satellite photo illustrating the recorded accuracy of an opportunistic whooping crane group reported in Nebraska, USA. The black box approximately delineates a section of land (2·59 km^2) as classified by the Public Land Survey System (PLSS). The gold 'x' is the location of a whooping crane group recorded with a global position system (GPS) with a 500-m-radius buffer (gold circle). The red 'x' represents the centre of the PLSS section with a 500-m-radius buffer (red circle). Of 68 whooping crane group records from 2000–2012, 32 had locations recorded with a GPS and 36 locations were recorded at the centre of the PLSS section.

sampling bias (Kéry 2011; Fithian & Hastie 2013; Hefley *et al.* 2013). Sampling bias occurs when the probability that a whooping crane group is reported depends on environmental covariates (Dorazio 2012; Hefley *et al.* 2013). Sampling bias is not unique to our whooping crane data set, but likely exists in many presence-only data sets. Instead, our goal was to determine the effects of location errors on SDM results and explore remedial methods; considering a simplified analysis allowed us to accomplish this goal. Ignoring sampling bias does not limit the usefulness of our study, because the effects of location error would be present if sampling bias were corrected for and the remedial methods we develop could be used with or without a correction for sampling bias. Furthermore, we supported our empirical results with a simulation study where the true relationships between the environmental covariates and the presence-only locations were known.

THE EFFECTS OF ERRORS IN COVARIATES AND REGRESSION CALIBRATION

The effects of errors in covariates can be difficult to determine except when simple linear regression models with a single covariate are used. With multiple covariates and nonlinear effects, the effects of errors in covariates are complex and difficult to describe (Carroll, Ruppert & Stefanski 1995). We proceed by describing the effects of errors in covariates for simple linear regression; however, we present this only as a heuristic, and it should be emphasized that our results do not necessarily apply to SDMs.

In simple linear regression, when estimating the effect covariate x has on the response variable y, the covariate is assumed to be measured perfectly. Introducing random error into the covariate results in coefficient attenuation (i.e. coefficient estimates are closer to zero). The effects of systematic errors on regression coefficients can be more serious. Consider the example where the response y depends on the covariate x. Instead of measuring x, $w = bx + c$ is measured, where b is the systematic bias in the variability of the covariate and c is the systematic bias in the numerical value of the covariate. In the case $b = 1$ and $c \neq 0$, the regression coefficient estimates would be unbiased; however, the estimated intercept would be biased. In the case $b \neq 1$, estimates of the regression coefficient will be biased and the magnitude and direction of the bias will depend on the numerical value of b.

Combining both random and systematic error, the observed covariate w can be written as $w = bx + c + e$, where e is the random measurement error. From this example, it is clear that linear regression can be used to model the expected value of the true covariate x, given the measured covariate w ($\mathrm{E}[x|w]$). The model predicting $\mathrm{E}[x|w]$ is known as a calibration model. For presence-only observations without exact locations, w_{error} is the observed covariate (i.e. the value of the covariate at the recorded locations). The calibration model is used to predict or estimate the expected value of the covariate given the measured covariate ($E[x_{predict}\widehat{|W}_{error}]$). The prediction or estimate of $\mathrm{E}[x_{predict}\,|w_{error}]$ is then used as the covariate in the SDM and will result in corrected (with respect to location error) SDM coefficient estimates. This method, known as regression calibration, has a long history of use in measurement error models and is potentially applicable to any regression model (Carroll, Ruppert & Stefanski 1995). To implement regression calibration, a prediction of $\mathrm{E}[x|w]$ is needed, but the relationship between x and w does not need to be linear or univariate and a wide array of modelling techniques could be used (Carroll, Ruppert & Stefanski 1995).

Regression calibration, however, requires a sample of covariates from exact locations (x_{exact}) measured without error and accuracy-degraded locations (w_{exact}). For many presence-only data sets with errors in locations, a sample of exact and degraded locations could be easily obtained. For example, if some location estimates in an opportunistic sightings data base were obtained using a GPS, degrading those locations based on a known mechanism such as the centre of a PLSS section may be a feasible means of obtaining data to build a calibration model.

We must emphasize, however, that systematic error in geographical space may not necessarily result in systematic error in an environmental space; similarly, the reverse holds true. For example, the geographical error introduced by recording the location as the centre of a PLSS section may produce random errors in the geographical covariates (i.e. the latitude and longitude of the location). Within the study area, development (e.g. houses and roads) is most often on the edges of the PLSS section because roads typically surround each section (Fig. 1). The centre of the PLSS section is generally as far as possible from development; therefore, we would expect the development covariate to contain systematic error.

SPECIES DISTRIBUTION MODEL

We analysed data comprised of opportunistic whooping crane group locations reported in Nebraska using an IPP model. Our IPP model is similar to a generalized linear model with a Poisson response distribution in that the environmental covariates affect the relative intensity of crane group abundance through the log link function. We can write the linear predictor in our IPP as:

$$\log(\lambda) = \beta_0 + \beta_1 \times \text{aquatic} + \beta_2 \times \text{development}, \qquad \text{eqn 1}$$

where λ is the intensity, β_0 is the intercept and the remaining β_is are regression coefficients for each environmental covariate at a fixed scale (i.e. 100-, 250- or 500-m-radius buffer in our analysis). In general, β_0 is not identifiable from presence-only data and is not necessarily needed to direct habitat management decisions (i.e. to estimate coefficients; Fithian & Hastie 2013). Instead of the true intensity, λ would represent the relative intensity and would describe how relative intensity of crane group abundance changes in response to the covariates. The IPP likelihood function contains an integral that can be difficult or impossible to solve. Solving this integral is similar to determining the number and location of pseudoabsences when using logistic regression or maximum entropy methods. The IPP differs from these methods; however, in that, the integral is defined over the entire region from which the presence-only data could have been reported; in our example, this area is the state of Nebraska (Warton & Shepherd 2010). We approximated the integral and estimated regression coefficients using maximum likelihood by infinitely weighted logistic regression with 10 000 Monte Carlo integration points and weights of 10 000 (see Appendix S1 for annotated R code; Fithian & Hastie 2013). We varied the number of Monte Carlo integration points and found that coefficient estimates stabilized at or before 10 000 points. We therefore chose to use 10 000 Monte Carol integration points. The location of the Monte Carlo integration points was the same for all of our analyses. We used program R (version 2.15.2) for all statistical computations (R Development Core Team 2013).

EFFECTS OF LOCATION ERRORS

To test the effect location errors had on the covariates in our IPP-SDM, we used the 32 crane group records that had locations estimated with a GPS (henceforth, exact locations). We degraded the exact locations by using the centre of the PLSS section as the location instead of the exact location (henceforth, degraded locations; Fig. 1; Nebraska

Department of Natural Resources 1995). We used the degraded locations to simulate the geographical location error present in the full data set. The average distance between exact locations and degraded locations was 557 m (SD = 454 m). As a metric of comparison, we also degraded the exact locations by adding independent bivariate normal random error (henceforth, randomly degraded locations). We considered two levels of random error: small (σ = 100 m) and large (σ = 1000 m). We chose values of σ for the small and large levels of random location accuracy degradation so that the distance between the exact and section level degraded locations was approximately in between the expected distances of the small and large randomly degraded locations, which were 125 and 1254 m, respectively. For this analysis, the two environmental covariates were not highly correlated ($R^2 < 0{\cdot}10$) for the exact locations and all levels of accuracy degradation.

REGRESSION CALIBRATION

For the 32 exact locations, we used linear regression to model the true environmental covariates (x_{exact}) obtained from the exact locations using covariates obtained from the accuracy-degraded locations (w_{exact}). Regression calibration required a prediction or estimate of the expected value of the true covariates conditional on the observed covariate. For our example, $E[\widehat{x_{predict}|W_{error}}]$ was the predicted value of the covariates given the observed covariate w_{error} based on the estimated linear regression equation obtained from the exact locations. We then used $E[\widehat{x_{predict}|W_{error}}]$ as the environmental covariates in the IPP model. This procedure results in corrected coefficient estimates for the IPP model assuming that the calibration model predicts $E[\widehat{x_{predict}|W_{error}}]$ well. We note that any measurable covariates could be used to predict the true covariate and that several methods exist for complex, multidimensional and nonlinear relationships (Carroll, Ruppert & Stefanski 1995).

Although regression calibration resulted in corrected regression coefficient estimates for the IPP model, obtaining corrected measures of coefficient uncertainty, such as standard errors (SEs) and confidence intervals (CIs), required additional effort. We used a two-phase, nonparametric bootstrap algorithm to correct measures of coefficient uncertainty (Efron & Tibshirani 1994; Haukka 1995). The two-phase nonparametric bootstrap algorithm integrated over the uncertainty in the covariate measurement error model and provided SEs and CIs that were corrected for small sample size (Haukka 1995). Such small sample size corrections would be required when the presence-only sample results in non-asymptotic sampling distributions of the IPP model parameters. Although bootstrapping required extra effort, researchers should test the asymptotic assumptions associated with conventional asymptotic SEs and CIs estimates especially when the sample size is small (Efron & Tibshirani 1994). Below, we present the two-phase nonparametric bootstrap algorithm for the IPP model (or any SDM) corrected for covariate measurement error.

1 Calculate environmental covariates (x_{exact}) for the sample of exact locations.

2 Degrade location accuracy of exact locations simulating the accuracy degradation in the presence-only data with location error and calculate environmental covariates (w_{exact}).

3 Draw a single bootstrap sample from x_{exact} and w_{exact}.

4 Model the bootstrap sample of x_{exact} using w_{exact} as the covariate.

5 Predict the true environmental covariate ($x_{predict}$) from the model in step four using the observed covariate (w_{error}) from the location records with errors.

6 Combine x_{exact} and $x_{predict}$ and draw a single bootstrap sample from the combination.

7 Fit the IPP model with the bootstrap sample from step six and save the coefficient estimates.

Repeat steps three through seven to obtain b bootstrap estimates of IPP model parameters or predictions. For all of our analyses, we used b = 1000 and obtained 95% CIs from the equal-tailed percentiles of the bootstrap samples. In our algorithm, bootstrap sample refers to a sample of the original data that has the same number of data entries as the original data, but is sampled with replacement (Efron & Tibshirani 1994). It should be noted that for the IPP model, the bootstrap resampling is applied only to the presence-only data and not the integration points. An annotated example with R code implementing the two-phase, nonparametric bootstrapping algorithm for the IPP is provided in Appendix S1 & S2.

COMPARSION

We compared coefficient estimates and 95% CIs from the analysis of the exact locations (n = 32) under various levels of location accuracy degradation (section, small and large) and our full data set (n = 68) with and without correction for 100-, 250- and 500-m-radius buffers. Correcting for location errors can result in estimates of regression coefficients with larger variances and wider CIs. Attempts to correct for bias should always be accompanied by an examination of the resultant increase in variance, and choosing the level of bias correction should be viewed as a bias–variance trade-off (Carroll, Ruppert & Stefanski 1995). Comparing the coefficient estimates and associated CIs allowed us to accomplish this goal in an interpretable manner, although the comparison would also be valid, albeit less interpretable using our example, for predictions (e.g. heat map of λ).

SIMULATION STUDY

To better understand the effects of location error on the relationship between the distribution of species abundance and habitat covariates derived from locations with error, we conducted a simulation study. We simulated presence-only records using an inhomogeneous Poisson point process distribution over the spatial domain of the state of Nebraska. The inhomogeneous Poisson point process distribution corresponded to the IPP model likelihood of our SDM. Similar to the IPP-SDM used in our analysis of the whooping crane data (eqn 1), the natural log of the intensity (log(λ)) of the inhomogeneous Poisson point process distribution can be written as a linear function of the environmental covariates:

$$\log(\lambda) = 3.875 + 5 \times \text{aquatic} + 0 \times \text{development}. \qquad (2)$$

For our simulation, we calculated the environmental covariates as the proportion of each land class within a 500-m buffer. We chose the numerical values of the coefficients to be similar to the results of the analysis of the whooping crane data. We set the coefficient for the development covariate equal to zero because we wanted to explore the effects of location error when no true effect existed. We chose a 500-m-radius buffer because we felt the analysis of the whooping crane data was most interesting statistically and ecologically at this scale (see Results and Discussion). The size of the calibration sample (i.e. x_{exact}) was 32, the same as the full analysis, and we used 100 simulated data sets. The IPP-SDM and methods used to estimate the coefficients from the simulated data were exactly the same as were used on the whooping crane data.

Results

EXACT LOCATIONS

When the exact location was known, coefficient estimates for the aquatic covariate (β_1 in eqn 1) were 4·36, 5·44 and 6·66 for the 100-, 250- and 500-m-radius buffer, respectively (Fig. 2). Coefficient estimates for the development covariate (β_2 in eqn 1) were −11·98, −6·88 and 0·82 for the 100-, 250- and 500-m-radius buffer, respectively (Fig. 2). Coefficient estimates for the aquatic covariate from data with location errors were similar to that obtained from the exact locations, except the coefficient estimate for locations with larger errors (σ = 1000 m) was attenuated. In general, coefficient estimates for the development covariate were attenuated when errors in location were present and ignored (Fig. 2). Note, however, this was not the case for the development coefficient for the 500-m-radius buffer size, which was 0·82 when the location was known exactly, but −3·19 when the errors in location were at the PLLS section level. The smallest attenuation of estimated regression coefficients occurred when the accuracy degradation was small (σ = 100 m; Fig. 2). When the location accuracy was degraded to the PLSS section level, the regression coefficients were similar or, in some cases, larger in magnitude when compared to the coefficients when accuracy deterioration was large (σ = 1000 m; Fig. 2). The bias caused by errors in locations generally decreased as the size of the buffer increased. When regression calibration was used to correct for location errors, all coefficient estimates were similar, if not identical to the second decimal place, to the coefficient estimates obtained when the location was known exactly (Fig. 2).

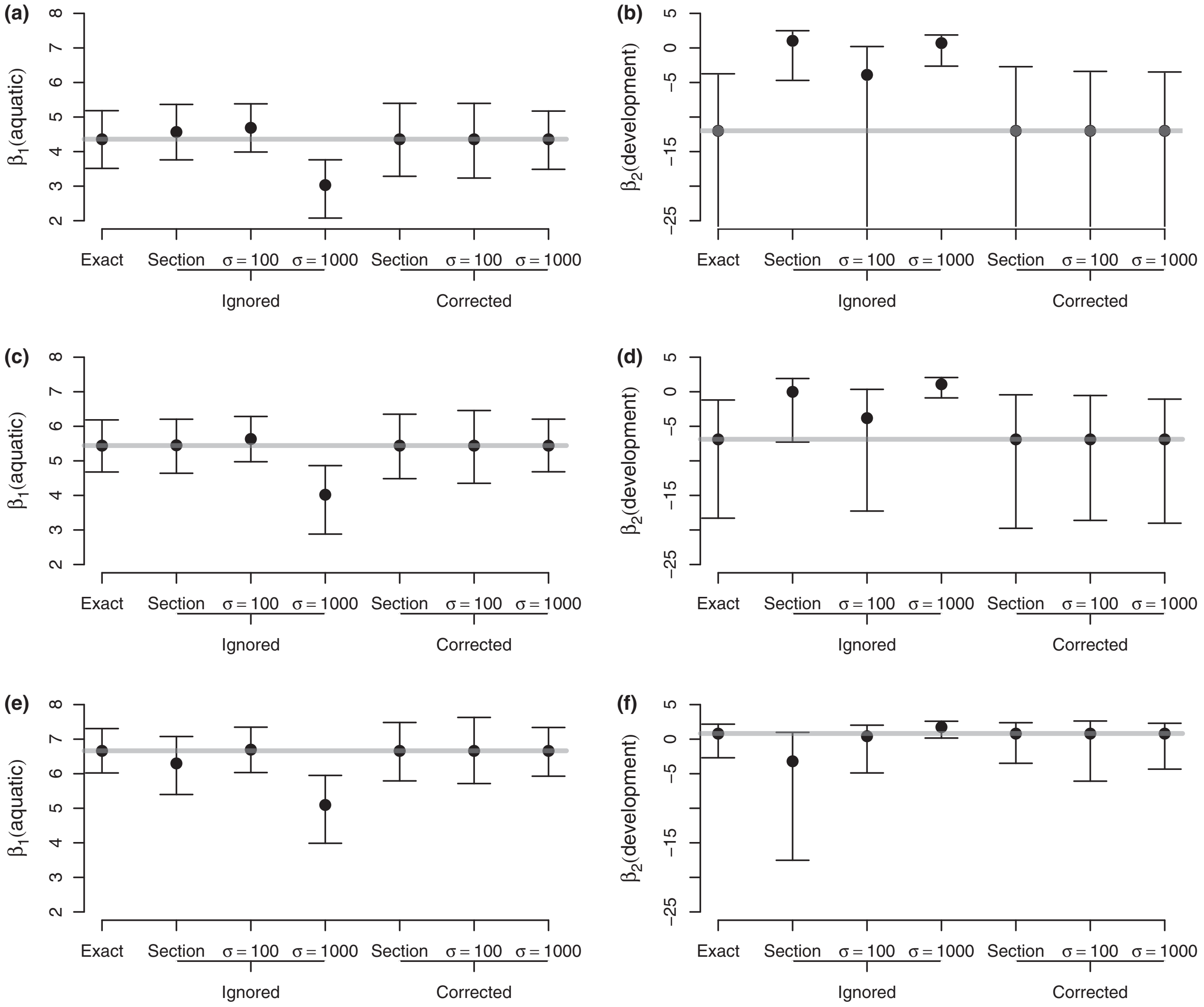

Fig. 2. Estimated inhomogeneous Poisson point process regression coefficients for aquatic habitat (β_1) and development (β_2) with 95% confidence intervals (CIs) estimated from whooping crane locations recorded with a global position system (Exact, n = 32) and three varying levels of simulated accuracy. Environmental covariates were calculated as the proportion of habitat type within a 100-m- (a and b), 250-m- (c and d), and 500-m- (e and f) radius buffer. Section locations were degraded in accuracy by recording the location as the centre of the Public Land Survey System section. The σ = 100 and σ = 1000 were degraded in accuracy by adding independent bivariate normal location errors to the exact locations with standard errors of 100 and 1000 m, respectively. The grey line represents coefficient estimates from an analysis of the 32 exact locations. Note: lower limit of 95% confidence intervals (CIs) for the development covariate at the 100-m-radius buffer extend beyond the range shown in the figure.

The CIs for all aquatic habitat coefficient estimates were similar in width, although slightly wider when the location error was corrected for. In contrast, the CIs for the development coefficients for the 100-m-radius buffer were wide except when location error was large or at the section level (Fig. 2). The CIs for the development coefficients at the 100-m-radius buffer size were wide because the empirical distribution was skewed with heavy tails. In general, the width of the CIs decreased as buffer size increased, and the CIs were wider when location error was corrected for; however, the increase in CI width, when compared to the exact locations, was not large.

FULL DATA SET

When location error was ignored, coefficient estimates for the aquatic habitat covariate obtained in the analysis of the full data set (n = 68) were slightly attenuated when compared to estimates obtained when location error was corrected for (Fig. 3). Both corrected and uncorrected coefficient estimates for aquatic habitat were smaller than the coefficient estimates obtained when only exact locations (n=32) were analysed (Fig. 3). The differences between the estimated coefficients for the aquatic covariate, however, were generally small (Fig. 3). Coefficient estimates for the development covariate when location errors were ignored were strikingly different from the corrected estimates and estimates obtained from the exact locations (Fig. 3). The difference between the development coefficient estimates when location error was ignored and corrected for was of the same sign and generally of the same magnitude when compared to the analysis of the exact locations with simulated location error at the section level (c.f. Exact and Section vs. Ignored and Corrected; Figs 2 and 3).

SIMULATION RESULTS

Our simulations resulted in an average of 67·7 (SD = 9·0) presence-only locations. When the exact location of the simulated presence-only data was used to derive the aquatic and development covariates, the distributions of the coefficient estimates from the IPP-SDM ($\bar{\beta}_1$ = 4.90 and $\bar{\beta}_2$ = −1.06) were centred, relative to the variability in the estimates, near the true values of 5·0 and 0·0, respectively (Fig. 4). When the location of the presence-only location was recorded as the centre of the PLSS section, the aquatic coefficient estimates were attenuated ($\bar{\beta}_1$ = 4.13); the development coefficient, however, was very biased with an average value of $\bar{\beta}_2$ = −11.40. Calibrated regression was successfully able to correct for the bias with only a small increase in the variability of the coefficients (Fig. 4).

Discussion

We found that random errors in location can result in biased regression coefficient estimates for the IPP model. This might be expected as a general result for the IPP-SDM, because as the random error in the covariates tends to infinity the IPP is reduced to a homogeneous poison process (i.e. coefficients are reduced to zero; Dobrushin 1963; Cressie & Wikle 2011). In general, results from our analyses that incorporated small (σ = 100 m) and large (σ = 1000 m) levels of random accuracy deterioration tended to support this conclusion (Fig. 2).

For the whooping crane data analysis, we might have expected the effect of development would depend on the scale examined. For example, in our study area, most PLSS sections (a 2·59 km^2 geometrically square area) were surrounded by roads and rural development usually occurs next to roads

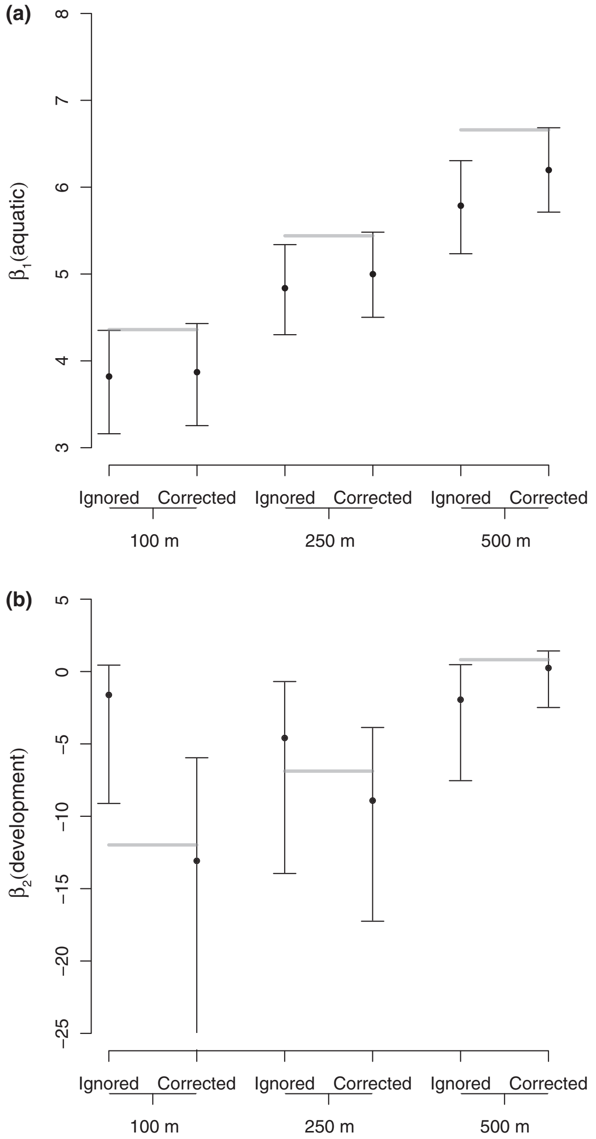

Fig. 3. Inhomogeneous Poisson point process (IPP) regression coefficients for aquatic habitat (a) (β_1) and development (b) (β_2) with 95% confidence intervals (CIs) estimated from opportunistic whooping crane locations (n = 68). Environmental covariates were calculated as the proportion of habitat type within a 100 m, 250 m and 500-m-radius buffer. Axis ticks labelled 'Ignored' indicate the IPP regression coefficients were estimated with no correction for location errors, whereas plots labelled 'Corrected' indicate coefficients were corrected using regression calibration. The grey lines represent IPP regression coefficient estimates obtained from 32 whooping crane group locations that were recorded with a global positioning system (i.e. 'Exact' point estimates and grey lines from Fig. 2). Note: lower limit of the 95% CI for the corrected development covariate at the 100-m-radius buffer extends beyond the range shown in the figure.

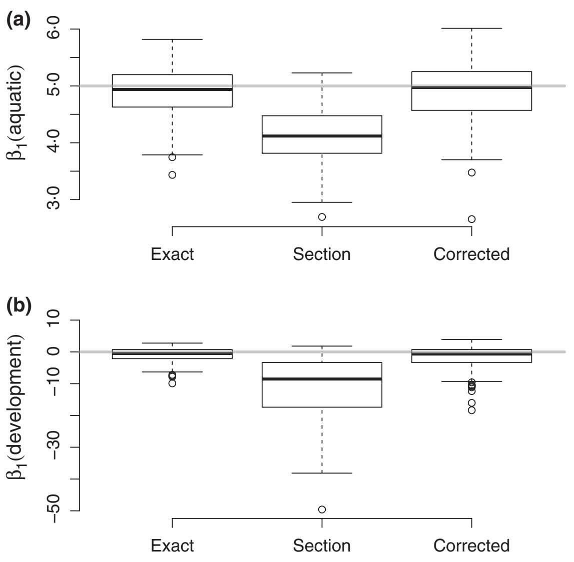

Fig. 4. Simulation results from presence-only data ($\bar{n}$ = 67·7) when the location is recorded exactly (Exact) and at the centre of the Public Land Survey System section (Section) in which the point occurred (see Fig. 1). Coefficients for aquatic habitat (a) and development (b) were estimated using an inhomogeneous Poisson point process species distribution model. Each box and whisker plot corresponds to the maximum likelihood estimate from 100 simulated data replicates. Grey lines show the true coefficient value. Environmental covariates were calculated as the proportion of habitat type in a 500-m-radius buffer.

(Fig. 1). It would have been relatively easy for whooping cranes to avoid areas of development within a 100 and 250 m radius, but more difficult to avoid development within the 500 m radius. Unless the exact location of the whooping crane group was near the centre of the PLSS section, it would be difficult to avoid a small amount of development; by recording the location as the centre of the PLSS section, the 500 m radius buffer will, in most situations, contain little or no developed areas (Fig. 1).

The coefficient estimates and CIs for the exact locations (Fig. 2) and corrected estimates from the full data set (Fig. 3) support the conclusion that whooping cranes avoid development within a 100 and 250 m radius, but are indifferent to development at 500 m. When the location of the crane group is recorded as the centre of the PLSS section, we observed coefficient attenuation at the 100- and 250-m-radius buffer sizes, likely due to random error, and a negative bias for the development coefficient at the 500-m-radius buffer size due to systematic error. This result is strongly supported by comparisons of the analysis of exact locations and the full data set (Figs 2 and 3) and further supported by the simulation study that shows coefficient estimates for development at the 500-m-radius buffer are negative when the location is recorded in the middle of the PLSS section even when the true value was known to be zero. From the simulation, the average value of the coefficients for development at the 500-m-radius buffer was −1·11 when location was known exactly compared with −11·40 when the location was recorded as the centre of the PLSS section (Fig. 4). Given the true effect was zero and that the coefficient represents a change in relative log intensities, −11·40 is a large number representing a change in intensity of 29 437 times greater between an area that is 100% development when compared to an area that is 0% development (i.e. $\frac{e^{-1.11}}{e^{-11.40}}$).

We were encouraged to find that regression calibration successfully reduced the bias in coefficient estimates caused by the errors in locations for all levels of accuracy degradation. We did not expect regression calibration to perform well at the 100-m-radius buffer due to the relatively large size of the location errors in comparison to the scale examined. The reduction in bias, however, was not free as the regression calibration correction resulted in an increase in variance of parameter estimates and thus wider CIs (Fig. 2). For some covariates, such as the aquatic covariate in our analysis, the bias caused by errors in location may be minimal and correction may not be warranted. We suggest researchers and managers consider the study goals in the light of the bias–variance trade-off when using regression calibration. For example, if the goal is to make predictions using the IPP regression coefficient estimates, calibrated regression could reduce or eliminate the bias in estimates of relative intensity. In the case of prediction, bias correction may be worthwhile for small buffer sizes; however, it would be important to communicate the increased uncertainty associated with the predictions due to the bias reduction.

The coefficient estimates from the exact locations and the full data set may have been influenced by sampling bias (Hefley *et al.* 2013). For example, the result that both corrected and uncorrected coefficient estimates for aquatic habitat were smaller than the coefficient estimate obtained when only exact locations were analysed (Fig. 3) may be a result of differing sampling bias between the two data sets. However, verifying this conclusion would be difficult, if not impossible because it would require an estimate of sampling bias (Hefley *et al.* 2013).

Lastly, our methods implicitly assume that the covariates can be measured without error. For example, if the exact location of a whooping crane group is known, we can measure the two habitat covariates exactly or, at least, with minimal error. This may not be true for analyses deriving covariates from sparse or interpolated environmental data (i.e. where the covariate at the true location is a prediction, not a measurement). Our methods do not address this additional error and is an area of needed research (Foster, Shimadzu & Darnell 2012).

Conclusion

When possible, we recommend field biologists to expend additional effort to obtain accurate location estimates. For our example, it seems reasonable that the accuracy of the location records for whooping cranes could be increased with minimal effort. When analysing presence-only records, corrective methods such as regression calibration may be the only option to explore the effects of and possibly correct for errors in the location data. Alternatively, we could have only used the 32 exact locations in our analysis or ignored the location error. Using the 32 exact locations would have resulted in a loss of 53·7% of the data. Our practical experience with wildlife biologists and managers suggests analysis of the full data set would be more desirable for informing conservation decisions. Ignoring location error and analysing the full data set would have resulted in a different conclusion. For example, based on a 95% CI covering zero, by ignoring location errors, we would have concluded that whooping crane group abundance is not related to the proportion of development within a 100-m-radius buffer, when in fact the effect is negative (c.f. Exact and Section; Fig. 2).

Whether one chooses to correct for location errors or not depends on the specifics of the data collection process, the available data and the geographical and environmental space the species occupies. The effects of location errors on coefficient estimates can be difficult or impossible to anticipate without additional contextual information (e.g. Fig. 1). Even if there is additional information available, the effect location errors have on coefficient estimates will become very complex as more covariates are added to the SDM and as more complex relationships between the covariates and intensity of abundance are explored. Regardless, calibrated regression can reduce the inherent biases in the data, but the method requires some exact location records and knowledge of the mechanism of accuracy degradation.

Acknowledgements

We thank all the citizens who reported sightings of whooping cranes and the United States Fish and Wildlife Service for maintaining the data base. We thank R. Grosse for developing NLCD data layers used in this analysis. We thank five anonymous reviewers, the Associate Editor and K. Lawry for comments that improved the quality of this manuscript. This research was supported by funding from the Platte River Recovery Implementation Program and the National Science Foundation Integrative Graduate Education and Research Traineeship (NSF-DGE-0903469).

Data accessibility

The source data for these analyses are archived in the Dryad Digital Repository doi: 10.5061/dryad.h81s5.

References

Araújo, M.B. & Guisan, A. (2006) Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, **33**, 1677–1688.

Austin, J. & Richert, A. (2001). A comprehensive review of observational and site evaluation data of migrant whooping cranes in the United States, 1943–99. Retrieved July 17, 2013, from http://www.npwrc.usgs.gov/resource/birds/wcdata/pdf/wcdata.pdf

Barry, S. & Elith, J. (2006) Error and uncertainty in habitat models. *Journal of Applied Ecology*, **43**, 413–423.

Bradter, U., Kunin, W.E., Altringham, J.D., Thom, T.J. & Benton, T.G. (2013) Identifying appropriate spatial scales of predictors in species distribution models with the random forest algorithm. *Methods in Ecology and Evolution*, **4**, 167–174.

Carroll, R.J., Ruppert, D. & Stefanski, L.A. (1995). *Measurement Error in Nonlinear Models*. Chapman and Hall, London, UK.

Cressie, N. & Wikle, C.K. (2011). *Statistics for Spatio-Temporal Data*. John Wiley & Sons, Hoboken, New Jersey, USA.

Dobrushin, R.L. (1963) On Poisson laws for distributions of particles in space. *Ukrains'kyi Matematychnyl Zhurnal*, **8**, 127–134.

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

Efron, B. & Tibshirani, R.J. (1994). *An Introduction to the Bootstrap*. Chapman and Hall/CRC, Boca Raton, Florida, USA.

Elith, J. & Leathwick, J. (2007) Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. *Diversity and Distributions*, **13**, 265–275.

Elith, J. & Leathwick, J.R. (2009) Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 677–697.

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

Foster, S.D., Shimadzu, H. & Darnell, R. (2012). Uncertainty in spatially predicted covariates: is it ignorable? *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **61**, 637–652.

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

Fry, J., Xian, G., Jin, S., Dewitz, J., Homer, C., Yang, L., Barnes, C., Herold, N. & James, W. (2011) Completion of the 2006 national land cover database for the conterminous United States. *Photogrammetric Engineering & Remote Sensing*, **77**, 858–864.

Graham, C.H., Elith, J., Hijmans, R.J., Guisan, A., Townsend Peterson, A. & Loiselle, B.A. (2007). The influence of spatial errors in species occurrence data used in distribution models. *Journal of Applied Ecology*, **45**, 239–247.

Haukka, J.K. (1995) Correction for covariate measurement error in generalized linear models–a bootstrap approach. *Biometrics*, **51**, 1127–1132.

Hefley, T., Tyre, A., Baasch, D. & Blankenship, E. (2013). Non-detection sampling bias in marked presence-only data. *Ecology and Evolution*. doi: 10.1002/ece3.887. in press.

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.

Kéry, M. (2011) Towards the modelling of true species distributions. *Journal of Biogeography*, **38**, 617–618.

Kramer-Schadt, S., Niedballa, J., Pilgrim, J.D., Schröder, B., Lindenborn, J., Reinfelder, V. *et al.* (2013) The importance of correcting for sampling bias in MaxEnt species distribution models (M. Robertson, Ed.). *Diversity and Distributions*, **19**, 1366–1379.

Manly, B.F., McDonald, L., Thomas, D.L., McDonald, T.L. & Erickson, W.P. (2002). *Resource Selection by Animals: Statistical Design and Analysis for Field Studies*. Kluwer Academic Publishers, Dordrecht, the Netherlands..

Monk, J. (2013). How long should we ignore imperfect detection of species in the marine environment when modelling their distribution? *Fish and Fisheries*. doi: 10.1111/faf.12039.

Montgomery, R., Roloff, G.J. & Hoef, J.M.V. (2011) Implications of ignoring telemetry error on inference in wildlife resource use models. *Journal of Wildlife Management*, **75**, 702–708.

Montgomery, R., Roloff, G.J., Ver Hoef, J.M. & Millspaugh, J.J. (2010) Can we accurately characterize wildlife resource use when telemetry data are imprecise? *Journal of Wildlife Management*, **74**, 1917–1925.

Nebraska Department of Natural Resources. (1995). Nebraska sections boundary database. Retrieved January 16, 2013, from http://www.dnr.ne.gov/databank/PLSS

Pearce, J.L. & Boyce, M.S. (2006) Modelling distribution and abundance with presence-only data. *Journal of Applied Ecology*, **43**, 405–412.

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

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

R Development Core Team. (2013). R: a language and environment for statistical computing. Retrieved July 15, 2013, from http://www.r-project.org/

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.

Tyre, A.J., Tenhumberg, B., Field, S.A., Niejalke, D., Parris, K. & Possingham, H.P. (2003). Improving precision and reducing bias in biological surveys: estimating false-negative error rates. *Ecological Applications*, **13**, 1790–1801.

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.

Received 29 September 2013; accepted 12 November 2013
Handling Editor: David Warton

Supporting Information

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

Appendix S1. Annotated R code used to simulate and implement the inhomogeneous Poisson point process species distribution model, calibrated regression and two-phase bootstrap algorithm.

Appendix S2. Description of simulated data in Appendix S1.