Methods in Ecology and Evolution

BRITISH ECOLOGICAL SOCIETY

Methods in Ecology and Evolution 2017, **8,** 1566–1573 doi: 10.1111/2041-210X.12793

Bias correction of bounded location errors in presence-only data

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

[1]*Department of Statistics, Kansas State University, Manhattan, KS, USA;* [2]*Marine Mammal Laboratory, Alaska Fisheries Science Center, National Oceanic and Atmospheric Administration, Seattle, WA, USA; and* [3]*U.S. Geological Survey, Colorado Cooperative Fish and Wildlife Research Unit, Department of Fish, Wildlife, and Conservation Biology, Department of Statistics, Colorado State University, Fort Collins, CO, USA*

Summary

1. Location error occurs when the true location is different than the reported location. Because habitat characteristics at the true location may be different than those at the reported location, ignoring location error may lead to unreliable inference concerning species–habitat relationships.

2. We explain how a transformation known in the spatial statistics literature as a change of support (COS) can be used to correct for location errors when the true locations are points with unknown coordinates contained within arbitrary shaped polygons.

3. We illustrate the flexibility of the COS by modelling the resource selection of Whooping Cranes (*Grus americana*) using citizen contributed records with locations that were reported with error. We also illustrate the COS with a simulation experiment.

4. In our analysis of Whooping Crane resource selection, we found that location error can result in up to a five-fold change in coefficient estimates. Our simulation study shows that location error can result in coefficient estimates that have the wrong sign, but a COS can efficiently correct for the bias.

Key-words: citizen science, ecological fallacy, errors-in-variables, hierarchical model, MAXENT, Poisson point process, resource selection, species distribution model, Whooping Crane

Introduction

Determining the habitat preferences of a species is an important component to conserving the world's biodiversity. For example, understanding how variables such as temperature and vegetation influence the distribution of a species is necessary to predict the impact of threats such as climate change and habitat destruction (Elith, Kearney & Phillips 2010; Guisan *et al.* 2013). Accordingly, biogeographers, ecologists, wildlife biologist, statisticians and the machine learning community have developed models that facilitate inference on the habitat characteristics that individuals of a species prefer (Manly *et al.* 2004; Elith & Leathwick 2009; Hefley & Hooten 2016).

Observed locations of species occurrence are a common source of data used for modelling distributions and resource selection. These type of data are termed use-availability or presence-only data (hereafter referred to as presence-only data; McDonald *et al.* 2013) and can arise from a variety of sampling mechanisms such as citizen science and animal tracking data obtained from telemetry devices (e.g. Hefley & Hooten 2016; Hooten *et al.* 2017). An important property of presence-only data, whether citizen-reported sightings of a rare species or locations obtained from satellite-based telemetry, is that errors can occur when the true locations are different than the reported locations (Graham *et al.* 2008; Montgomery *et al.* 2010; Montgomery, Roloff & Ver Hoef 2011; Mitchell, Monk & Laurenson 2017). Location errors are problematic because habitat characteristics at the reported locations might be different than the habitat at the true locations. For example, Brost *et al.* (2015) used a resource selection function to determine the influence of bathymetry on harbour seal (*Phoca vitulina*) habitat use. However, over 66% of the recorded presence-only locations from the Argos telemetry data used by Brost *et al.* (2015) occurred within inaccessible regions on land, where bathymetry could not be calculated unless location error was explicitly modelled (see Fig. 2 in Brost *et al.* 2015). Location error is also prevalent in citizen science data, the locations of which are often reported as the nearest road, town or other landscape feature. For example, citizen-reported sightings of the critically endangered Whooping Crane (*Grus americana*) have been used to guide site selection for wind energy projects (Belaire *et al.* 2014), determine the influence of human disturbances (Hefley *et al.* 2014), and evaluate a protected area that includes a critical habitat designation under the U.S. Endangered Species Act (Hefley *et al.* 2015b). The studies of Belaire *et al.* (2014) and Hefley *et al.* (2014, 2015b) relied on opportunistic records of Whooping Cranes that were reported as the centroid of administrative units. As shown in Hefley *et al.* (2014),

*Correspondence author. E-mail: thefley@ksu.edu

location errors compromise statistical inference concerning species–habitat relationships.

Regression calibration can be used to account for location error, but the method has four main limitations: (i) a subset of the presence-only data must have exact locations (Hefley *et al.* 2014), (ii) the bias correction is not guaranteed to be optimal (Carroll *et al.* 2006), (iii) accounting for heterogenous location error is challenging, and (iv) specialized computational algorithms are required to properly account for uncertainty (e.g. a double bootstrap). Although the third and fourth challenges are surmountable, the second challenge may affect inference and the first challenge may cause an impasse; if at least a subset of exact locations are not available, then bias caused by location error cannot be corrected using regression calibration. Furthermore, it may be impossible to verify records that are reported to have exact locations for studies that rely on citizen science data (e.g. Hefley *et al.* 2014).

We present an approach that corrects for location error in presence-only data and resolves the four limitations associated with regression calibration. We accomplish this by presenting the Whooping Cranes records that motivated our study and reviewing a unified approach for the analysis of presence-only data. Next, we introduce a change of support (COS) that can be used to account for location error. Then we demonstrate the COS using the Whooping Crane data and conduct a simulation study to evaluate the properties of the method. Lastly, we provide future direction by connecting location error in presence-only data to the common practice of aggregating exact locations of individuals into survey units when using count or presence–absence data to determine species–habitat relationships. In all cases, the bias due to aggregation or location error can be interpreted as an ecological fallacy when the inference about resource selection at the location of individuals differs from the conclusions at the survey units (Robinson 1950; Bradley, Wikle & Holan 2017).

Materials and methods

WHOOPING CRANE DATA

Whooping Cranes are an endangered migratory bird with a single wild population of 329 individuals [293–371, 95% confidence interval (CIs)] as of winter 2015–2016 (U.S. Fish and Wildlife Service 2016; see table 1 of Butler, Harris & Strobel 2013 for historical estimates). 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 in Alberta and the Northwest Territories, Canada. Approximately 4000 km migrations are typically undertaken in small groups and include multiple stopovers that last from several hours to several weeks (Pearse *et al.* 2015, 2017). Stopovers provide much needed rest and food during the migration and are critical to the survival of individual Whooping Cranes and the species.

In 1978, a portion of the Central Platte River Valley in Nebraska (USA) was designated as critical habitat for the Whooping Crane under the U.S. Endangered Species Act (Fig. 1a; U.S. Fish and Wildlife Service 1978). One of the management objectives of the Platte River Recovery Implementation Program, a multistate-federal cooperative agreement, is to acquire and manage stopover habitat for Whooping Cranes within the critical habitat area and surrounding region (Fig. 1a; Freeman 2010). Within the region, habitat acquisition and management decisions are based on a set of 'minimum habitat criteria' such as the distance to nearest disturbance feature (defined as distance from a point in any direction to the nearest disturbance feature; Platte River Recovery Implementation Program 2012).

The presence-only Whooping Crane locations used in our study were described by Hefley *et al.* (2015b) and the data are available from the Dryad Digital Repository (Hefley *et al.* 2015a). For the purposes of this study, we limit our analysis to the locations of 120 groups of Whooping Cranes reported within a 6280 km^2 region that contains the critical habitat (Fig. 1a). A challenge associated with these data are that the location of most groups ($n = 103$) were reported as the centres of quarter-, half- or full-sections of land, which are administrative units (i.e. polygons) designated by the Public Land Survey System (Fig. 2a; Nebraska Department of Natural Resources 2016). Consequently, the true location of the Whooping Crane group is only known to be contained within a polygon having an area of roughly 0·65 km^2 (quarter-section), 1·3 km^2 (half-section), and 2·6 km^2 (full-section; Figs 1b and 2a). Relatively few groups ($n = 17$) were reported with locations obtained using a global positioning system.

We derived two covariates, distance to the nearest developed area and distance to water, from the 30 × 30 m^2 resolution 2011 National Landcover Cover Database (NLCD; Figs 1c,d and 2b,c; Homer *et al.* 2015). For the purposes of this study, we defined development as the amalgamation of land classes 21–24 (open space, low-, medium- and high-intensity development) and water as the amalgamation of land classes 11 (open water), 90 (woody wetlands), and 95 (emergent herbaceous wetlands). Location error of Whooping Crane groups is problematic because of the spatial structure of the two covariates. For example, roads, houses and other types of development typically occur on the boundary of a section of land (Fig. 2a). Locations of Whooping Crane groups are often reported as the centre of a section of land, which typically results in the maximum possible distance from development within the study area (Fig. 2b).

POINT PROCESS MODEL

Multiple researchers have unified methods for analyzing presence-only data by showing that many commonly employed modelling approaches (e.g. logistic regression, MAXENT, resource selection functions) estimate, with varying degrees of accuracy, parameters of an inhomogeneous Poisson point process (IPP) distribution (Warton & Shepherd 2010; Aarts, Fieberg & Matthiopoulos 2012; Renner & Warton 2013; Renner *et al.* 2015; Hefley & Hooten 2016). The IPP distribution describes the location of a random number of individuals (or groups) within a geographic area. The IPP is constructed by assuming the geographic distribution of individuals can be explained by an intensity function, $\lambda(\mathbf{s})$, defined in continuous space (where $\mathbf{s}$ is a single location contained within the study area). For example, the intensity function is commonly specified using a linear combination of location-specific covariates with

$$\log(\lambda(\mathbf{s})) = \beta_0 + \mathbf{x}(\mathbf{s})'\boldsymbol{\beta}, \qquad \text{eqn 1}$$

where β_0 is the intercept, $\mathbf{x}(\mathbf{s})$ is a $p \times 1$ vector that contains covariates at location $\mathbf{s}$, and $\boldsymbol{\beta} \equiv (\beta_1, \ldots, \beta_p)'$ is a vector of regression coefficients.

The support of a distribution is the set of points (locations) where the distribution is defined. For example, there are an infinite number of points within the study area where Whooping Crane groups could occur (Fig. 1a), and the continuous spatial support of the IPP distribution appropriately matches that of the true locations. Conversely, the

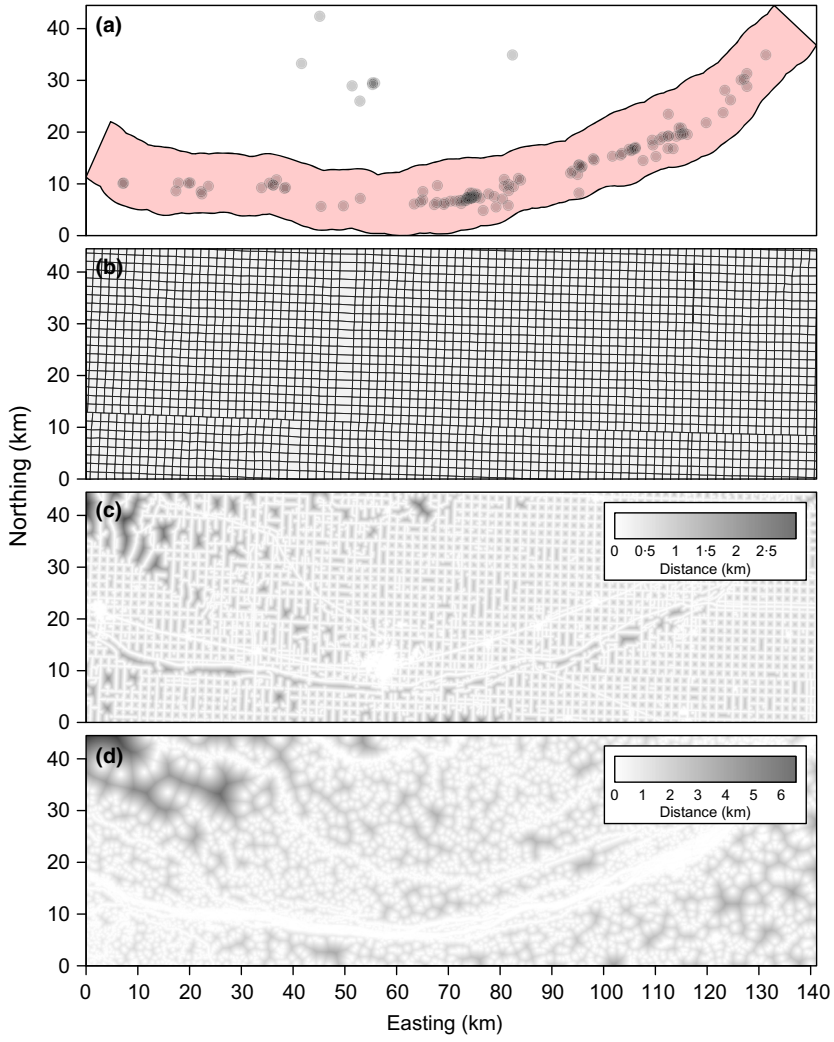

Fig. 1. Locations of opportunistic records ($n = 120$) of Whooping Crane groups in Nebraska (USA) from 1988 to 2012 (panel a). The 120 records were the presence-only data used in our analysis. The area shaded red is the Platte River Recovery Implementation Program associated habitat area which contains the critical habitat designated for the Whooping Crane. The study area is divided into irregular polygons designated as sections under the Public Land Survey System (panel b). The two covariates used in our analysis include distance to nearest development (panel c) distance to nearest water (panel d).

support of the reported locations is discrete (or areal in a spatial context) because there are a finite number of polygons (sections of land) that Whooping Crane groups could occur within (Gotway & Young 2002).

An important property of the IPP distribution is the support can be changed from continuous geographic space to discrete space, simply by integrating the intensity function over the area containing the true locations. More precisely, after a COS, the number of individuals (n_j) within non-overlapping polygons of arbitrary shape ($\mathcal{A}_j$) follows a Poisson distribution, with a rate parameter $\bar{\lambda}_j$, where

$$\bar{\lambda}_j = \int_{\mathcal{A}_j} \lambda(\mathbf{s}) d\mathbf{s}. \qquad \text{eqn 2}$$

As a result, when the true locations of individuals have unknown coordinates contained within known polygons, standard Poisson regression can be used to model the number of individuals within each polygon provided the rate parameter is calculated as shown in eqn 2. In most cases, the integral in eqn 2 must be approximated. For example, using a numerical quadrature approximation (Givens & Hoeting 2012, ch. 5)

$$\bar{\lambda}_j \approx \frac{1}{|\mathcal{A}_j|}\left(\frac{1}{Q_j}\sum_{q=1}^{Q_j}\lambda(\mathbf{s}_q)\right), \qquad \text{eqn 3}$$

where $|\mathcal{A}_j|$ is the area of $\mathcal{A}_j$ and Q_j is the number of (equally spaced) points on a grid that partitions the polygon $\mathcal{A}_j$. Conceptually, eqn 3 is the average value of the intensity function $\lambda(\mathbf{s})$ within $\mathcal{A}_j$ divided by the area of $\mathcal{A}_j$.

For the Whooping Crane example, the 6280 km^2 study area ($\mathcal{S}$) contains 2679 sections of land (i.e. $\mathcal{S} = \bigcup_{j=1}^{2679} \mathcal{A}_j$; Fig. 1b). Within each section of land ($\mathcal{A}_j$), the number of Whooping Crane groups ($n_1, n_2, \ldots, n_{2679}$), is the observed response. Of the 2679 sections of land in our study area (Fig. 1b), 2610 of the sections had zero Whooping Crane groups reported, 41 sections had 1 group, 19 sections had 2 groups and 9 sections had >2 groups.

To assist readers implementing the COS, a tutorial with R code illustrating the Whooping Crane data analysis and reproducing all results and figures is given in Appendix S1 of the Supporting Information. An in-depth presentation of the COS, including a detailed explanation of the IPP and why the COS works, is provided in Appendix S2.

SIMULATION EXPERIMENT

We conducted a simulation study to evaluate bias and efficiency in parameter estimation. We simulated presence-only data from the IPP distribution with a single covariate ($x(\mathbf{s})$) using the values of $\beta_1 = 1$ (i.e. $\log(\lambda(\mathbf{s})) = \beta_0 + \beta_1 x(\mathbf{s})$) on the unit square. We simulated two different sample sizes using $\beta_0 = 4{\cdot}25$ (small) and $\beta_0 = 6$ (large). For each sample size, we simulated three different types of covariates ($x(\mathbf{s})$) that

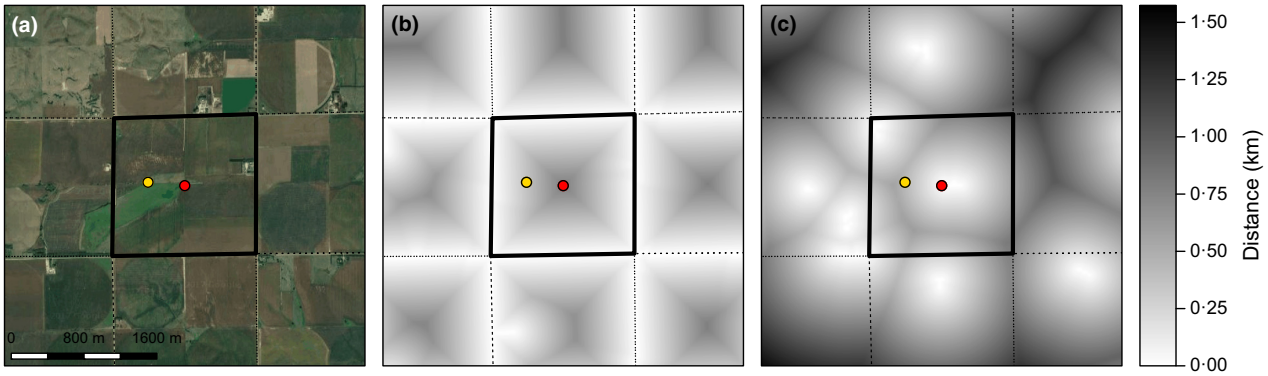

Fig. 2. Satellite photo illustrating the reported accuracy of a Whooping Crane group (panel a). The black boxes delineate sections of land. The yellow point (●) is the exact location of a Whooping Crane group reported with locations obtained with a global position system (GPS). The red point (●) represents the centre of the section. Of the 120 reported groups of Whooping Cranes used in our analysis, the location accuracy was: GPS (n = 17), quarter-section (n = 73), half-section (n = 11) and section (n = 19). The two covariates used in our analysis include distance to nearest disturbance (panel b) and water (panel c). The distance to nearest disturbance in panel b results in a regular pattern because most disturbance features (e.g. roads, houses, etc) occur on the boundary of a section of land.

resulted in 'habitat' with: (i) small-scale spatial autocorrelation (Fig. 3a); (ii) large-scale spatial autocorrelation (Fig. 3b); and (iii) large-scale spatial autocorrelation with near minimum values of $x(\mathbf{s})$ occurring within the centre of each polygon (Fig. 3c). Given that spatial autocorrelation is ubiquitous in spatial covariates, scenario 1 and 2 demonstrates the effect of different scales of autocorrelation on location error (Naimi *et al.* 2011). The third scenario is similar to the Whooping Crane example (i.e. distance to development is greatest at section centres) and shows that increasing levels of spatial autocorrelation in the covariate do not necessarily reduce the impact of location error (e.g. Naimi *et al.* 2011), when the location error itself has spatial structure relative to the covariate. For each scenario, we simulated location error by placing a grid with m = 100 (fine grain) and m = 25 (course grain) polygons to mimic sections of land and assumed that the 'reported' locations were the centre of each grid cell. For each simulated dataset, we fit the IPP distribution using maximum likelihood estimation to the presence-only data with the exact locations and data with location errors (i.e. cell centres; Fig. 3). We also fit the IPP distribution with the COS correction using data with location errors. In total, we conducted three different covariate scenarios (1, 2 and 3), with two different sample sizes (small and large), and two levels of aggregation (fine and course grain) for a total of 12 different settings. For each setting, we simulated 1000 datasets and assessed bias in the maximum likelihood estimates (MLEs) of β_1. We assessed efficiency by calculating the mean standard error of $\hat{\beta}_1$ from the COS correction using data with location errors divided by the mean standard error for $\hat{\beta}_1$ obtained from fitting the IPP distribution using data with exact locations (hereafter 'ratio of standard errors'). We also report the estimated coverage probability for the 95% CIs for $\hat{\beta}_1$ obtained from all model fits (i.e. exact, ignored and COS). A tutorial with R code implementing the simulation study and reproducing Figs 3 and S2–S4 is available in Appendix S3.

Results

WHOOPING CRANE DATA

When we fit the IPP model and ignored location error, we obtained estimates for the distance to nearest development of $\hat{\beta}_1 = 2{\cdot}2$ (1·7, 2·8; 95% CI) and distance to nearest water of $\hat{\beta}_2 = -5{\cdot}7$ (−6·8, −4·5). Using the COS to correct for location error, we obtained $\hat{\beta}_1 = 1{\cdot}4$ (0·6, 2·2) and $\hat{\beta}_2 = -29{\cdot}2$ (−46·4, −12·1). These results represent roughly a two-thirds-fold and five-fold estimated change in effect of distance to nearest development and water, respectively, when location error is corrected for (i.e. $\frac{1{\cdot}4}{2{\cdot}2} = 0{\cdot}64$ and $\frac{-29{\cdot}2}{-5{\cdot}7} = 5{\cdot}1$).

SIMULATION EXPERIMENT

When location error was ignored, the MLE of the regression coefficient ($\hat{\beta}_1$) was biased (Figs 3 and S2–S4) and coverage probabilities of the 95% CIs were ≤0·21 for all settings (Table 1). The bias was particularly pronounced in scenario 3 (i.e. large-scale spatial autocorrelation with near minimum values of $x(\mathbf{s})$ occurring within the centre of each polygon), where $\hat{\beta}_1$ was negative for every dataset from all settings when location error was ignored (Fig. 3c; note the true value was $\beta_1 = 1$). The COS correction resulted in unbiased $\hat{\beta}_1$ for all settings and coverage probabilities between 0·93–0·96 (Fig. 3 and S2–S4; Table 1). Overall, the COS correction was efficient; the ratio of standard errors was ≤2·5 for all simulation settings (Table 1). Given the collective results of all simulations, we present the 'worst case' (small sample size and a course grid resulting in large location errors) for the three covariate scenarios (Fig. 3). Detailed results for all simulations are presented in Table 1 and Figs S2–S4.

Discussion

Our analysis of the Whooping Crane presence-only data and simulation study show that location error can result in biased parameter estimates. In fact, biased estimates will always occur unless the location errors result in reported locations that have the same covariate values as the true locations. The magnitude and direction of the bias caused by location errors depends on the spatial structure of the covariates (e.g. Fig. 3; Hefley *et al.* 2014). Based on our simulation study, the COS is an efficient

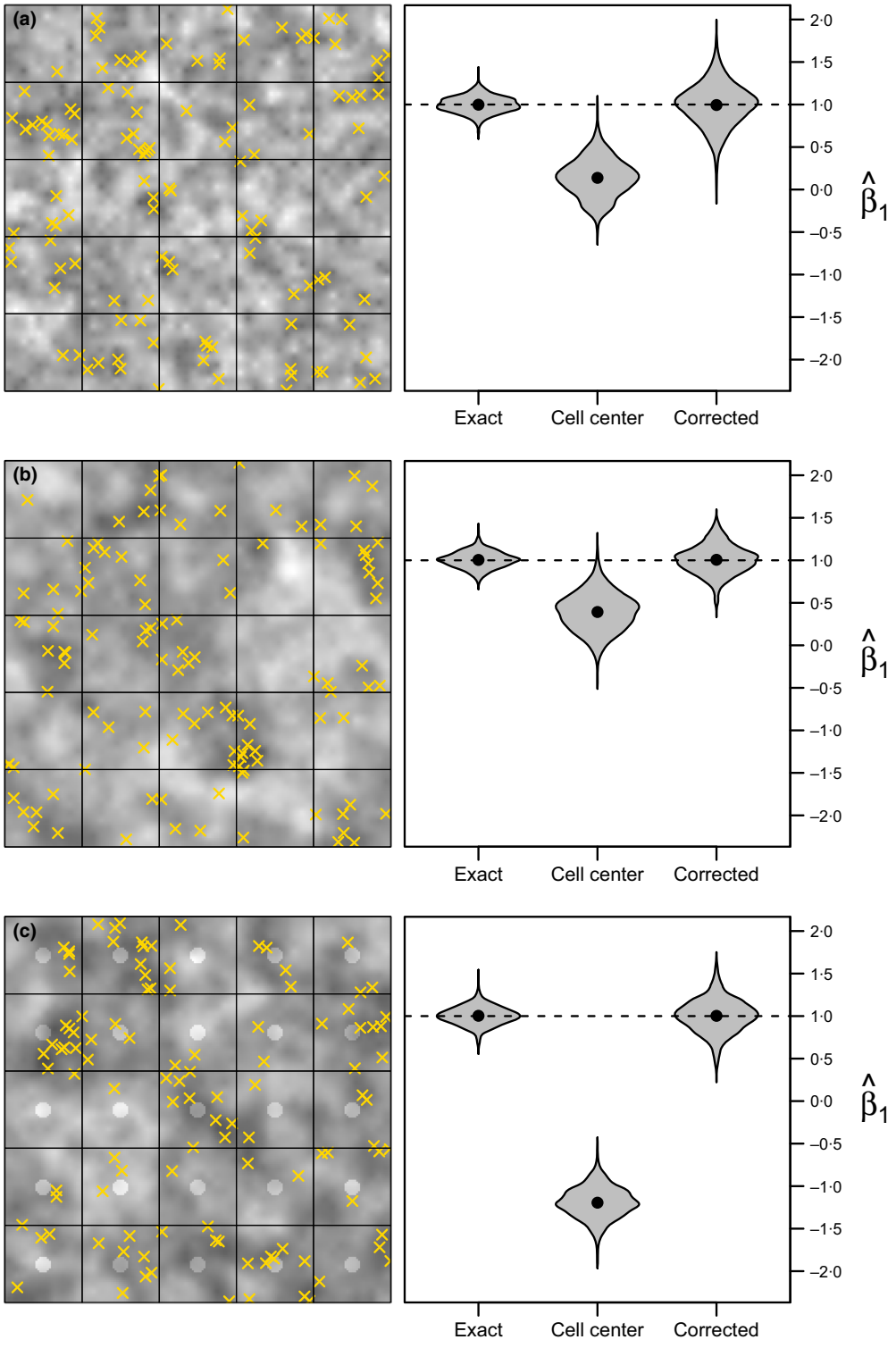

Fig. 3. Results of a simulation study that used three types of covariates for $x(\mathbf{s})$ which resulted in 'habitat' with: (i) small-scale spatial autocorrelation (panel a); (ii) large-scale spatial autocorrelation (panel b); and (iii) large-scale spatial autocorrelation with near minimum values of $x(\mathbf{s})$ occurring within the centre of each polygon (panel c; note that black shading is the maximum values of $x(\mathbf{s})$, white is the minimum of $x(\mathbf{s})$). The covariate $x(\mathbf{s})$ was used to simulate presence-only locations (gold ×). Maximum likelihood estimates for the regression coefficients ($\hat{\beta}_1$; violin plots) from 1000 data-sets for the IPP distribution using the exact locations (exact) and the locations with error caused by aggregating to the centre (cell centre). We also applied the COS correction (corrected), which used locations reported at the centre of each polygon, to obtain $\hat{\beta}_1$ for each dataset. For each type of autocorrelated covariate, we used multiple settings which included two different sample sizes (small and large) and two levels of location error (fine grain) and (course grain). Shown in this figure is the 'worst case' settings (small sample size and a course grid resulting in large location errors; see Figs S2–S4 for additional results).

Table 1. Results of a simulation study using three different types of covariates for $x(\mathbf{s})$ which resulted in 'habitat' with: (i) small-scale spatial autocorrelation (scenario 1; Fig. 3a); (ii) large-scale spatial autocorrelation (scenario 2; Fig. 3b); and (iii) large-scale spatial autocorrelation with near maximum values of $x(\mathbf{s})$ occurring within the centre of each polygon (scenario 3; Fig. 3c). For each covariate scenario we used two different sample sizes (small and large) and two levels of location error caused by aggregation (fine and course) for a total of 12 different settings. We report the average number of presence-only locations ($\bar{N}$) from 1000 datasets simulated for each setting and the estimated coverage probability (CP) for the 95% CIs obtained by fitting the IPP distribution using maximum likelihood estimation to the exact locations (exact), the locations reported as the centre (cell centre) and the COS correction (corrected). Also reported is efficiency, which was calculated as the mean standard error of $\hat{\beta}_1$ from the COS correction using data with location errors divided by the mean standard error of $\hat{\beta}_1$ obtained from fitting the IPP distribution using data with exact locations.

Scenario	Location error	Sample size	$\bar{N}$	CP (exact)	CP (cell centre)	CP (corrected)	Efficiency
1	Fine	Large	667·6	0·96	0·00	0·95	1·6
2	Fine	Large	666·8	0·96	0·01	0·96	1·3
3	Fine	Large	631·0	0·95	0·00	0·96	1·4
1	Fine	Small	115·6	0·95	0·01	0·95	1·6
2	Fine	Small	115·5	0·94	0·21	0·94	1·3
3	Fine	Small	109·4	0·95	0·00	0·96	1·4
1	Coarse	Large	665·3	0·96	0·00	0·95	2·5
2	Coarse	Large	663·4	0·95	0·00	0·95	1·7
3	Coarse	Large	651·3	0·95	0·00	0·95	1·7
1	Coarse	Small	115·6	0·94	0·00	0·94	2·5
2	Coarse	Small	114·8	0·95	0·03	0·94	1·7
3	Coarse	Small	112·6	0·94	0·00	0·93	1·7

technique that corrects location error, even when the location errors are relatively large (e.g. Fig. 3, Table 1). Location error for presence-only data can easily be accounted for using the COS when the true locations are unknown, but occur within known polygons. In situations where it is not known if the true locations are contained within polygons, the regression calibration approach of Hefley *et al.* (2014), movement constraint approach of Brost *et al.* (2015) or the errors-in-variables approach of Velásquez-Tibatá, Graham & Munch (2016) are viable alternatives.

The distribution of a species may have temporal dynamics. For example, Hefley *et al.* (2015b) used the IPP distribution to incorporate seasonal dynamics related to the migration of Whooping Cranes. The IPP distribution can be used in spatio-temporal species distribution models (Hefley & Hooten 2016). For spatio-temporal species distribution models, covariates at any given point in space may change over time. For example, the distance to nearest water may be dynamic seasonally or annually. We used the 2011 NLCD to derive the distance to nearest water even though NLCDs are available for the years 1992, 2001, 2006, 2011 and 2015. The distance to nearest water for the Whooping Crane locations collected in, for example, 2008 might be derived from the 'closest' available NLCD (i.e. 2006). Obtaining covariates associated with presence-only locations from covariates that are closest in time is tempting (i.e. using the 2006 NLCD to calculate distance to nearest water for the Whooping Crane locations collected in 2008), but temporal mismatch in covariates and observations may cause bias similar to location error.

While the COS can correct for location error, the covariates may also contain errors. For example, researchers may include spatial climate covariates obtained from climate models. A problem known as 'errors-in-variables' occurs when the predicted climate at a location is different than the true climate (Foster, Shimadzu & Darnell 2012). In addition to location errors, errors-in-variables can be accommodated using a hierarchical species distribution modelling framework (e.g. Stoklosa *et al.* 2015; Hefley & Hooten 2016; Velásquez-Tibatá, Graham & Munch 2016).

Researchers that use spatial covariates that are only available at a coarse resolution will face challenges even when the exact location of the species is known. For example, WorldClim provides a set of global climate variables that are predictions available on a 1×1 km^2 grid and are commonly used as covariates in species distribution models (Hijmans *et al.* 2005). The value of climate variables for an exact location record of a species will have the same value regardless of where the location falls within the 1×1 km^2 grid cell that contains the exact location. As a result, the true value of the climate variable where the species was located may be different than the value reported in the 1×1 km^2 grid cell. Climate could vary at a finer resolution and it is well-known that the resolution of covariates can influence results (e.g. Luoto, Virkkala & Heikkinen 2007; Scales *et al.* 2017). Although this problem is analogous to location error, the COS will not correct for the bias because the internal structure of the grid cells is unknown and results in the intensity function of the IPP being constant within each grid cell.

Although location error was the focus of our study, analyses that rely on citizen science data suffer from other biases related to the data collection process. For example, sampling bias occurs when the probability a location is reported is less than one and has spatial (or temporal) structure (e.g. Dorazio 2012; Hefley *et al.* 2013; Warton, Renner & Ramp 2013; Fithian *et al.* 2015). Simultaneously correcting for multiple biases, such as location error and sampling bias, is needed to obtain reliable statistical inference from presence-only data like the

Whooping Crane records used in our analysis (Hefley *et al.* 2015b; Hefley & Hooten 2016).

FUTURE DIRECTION

Count and presence–absence data arise from aggregating the exact locations of individuals into polygons (hereafter survey units; for example, Fig. 1·1 on p. 5 of Kéry and Royle 2016; Aarts, Fieberg & Matthiopoulos 2012; Hefley and Hooten 2016). For example, the European Breeding Bird Atlas makes presence–absence data available on a grid with 50×50 km^2 resolution (European Bird Census Council 2016). Popular ecological models for count and presence–absence data include the N-mixture approach of Royle (2004) and the occupancy model of MacKenzie *et al.* (2002) and Tyre *et al.* (2003). Models for count and presence–absence data typically use covariates that are the average values within the survey unit or values of the covariates at the centroid of the survey unit associated with each count or presence–absence observation. Consequently, inference is relative to how abundance or presence within the survey unit is influenced by covariates at the centroid or the average. As with location error in presence-only data, inference at the survey unit level may differ from the inference obtained using the covariates at the exact location of each individual.

Within a broader context, bias caused by mismatch in spatial scale is known as the ecological fallacy when conclusions at the exact locations of individuals differs from the inference at the aggregate level (Robinson 1950; Cressie & Wikle 2011, p. 197; Bradley, Wikle & Holan 2017). For example, suppose that we modelled presence–absence data within 1×1 km^2 quadrats using the elevation at the centre of the quadrats. If the true elevation was variable within the 1×1 km^2 quadrates, then the elevations where each individual was located may be different than the elevation at the centre of the quadrats. We expect that the effect of aggregating the exact locations for surveys of abundance or presence–absence will result in similar effects as location error in presence-only data.

Within the fields of spatial and spatio-temporal statistics, there are many well-established and recently developed methods for obtaining inference at the desired scale given data at another spatial scale (e.g. Gotway & Young 2002; Wikle & Berliner 2005; Bradley, Wikle & Holan 2016, 2017). Many of the statistical methods developed to address the modifiable areal unit problem, downscaling and change of support can be applied to ecological models when there is a mismatch in the scale at which the data were collected and inference is desired (e.g. Bradley, Wikle & Holan 2017). For example, when the locations of individuals are aggregated, as is common practice for surveys of abundance or presence–absence, and the goal is to obtain inference about the response of the species to covariates at the true locations of the individuals (i.e. resource selection), the well-established COS should be used. The COS involves linking the integrated intensity function to the expected count or probability of occurrence within the quadrats (Hefley & Hooten 2016).

Authors' contributions

T.J.H., B.M.B. and M.B.H. conceived the study. T.J.H. applied the statistical methods, conducted the simulation experiment and wrote the manuscript. B.M.B. developed the figures. All authors contributed substantially to revisions.

Acknowledgements

We thank the citizens who reported sightings of Whooping Cranes and the U.S. Fish and Wildlife Service for maintaining the database of records. We thank David Baasch, Joseph Northrup, Nelson Walker, the Associate Editor and two anonymous reviewers for comments that improved the quality of this manuscript. Any use of trade, firm or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Data accessibility

The Whooping Crane records are archived in the Dryad Digital Repository http://datadryad.org/resource/doi:10.5061/dryad.t6859 (Hefley *et al.* 2015a). Additional source data required to reproduce the results of this study (e.g. covariate layers) are archived in the Dryad Digital Repository https://doi.org/10.5061/dryad.82qd0 (Hefley, Brost & Hooten 2017).

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**(1), 177–187.

Belaire, J., Kreakie, B.J., Keitt, T. & Minor, E. (2014) Predicting and mapping potential Whooping Crane stopover habitat to guide site selection for wind energy projects. *Conservation Biology*, **28**(2), 541–550.

Bradley, J.R., Wikle, C.K. & Holan, S.H. (2016) Bayesian spatial change of support for count-valued survey data with application to the american community survey. *Journal of the American Statistical Association*, **11**, 472–487.

Bradley, J.R., Wikle, C.K. & Holan, S.H. (2017) Regionalization of multiscale spatial processes by using a criterion for spatial aggregation error. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, **79**(3), 815–832.

Brost, B.M., Hooten, M.B., Hanks, E.M. & Small, R.J. (2015) Animal movement constraints improve resource selection inference in the presence of telemetry error. *Ecology*, **96**, 2590–2597.

Butler, M.J., Harris, G. & Strobel, B.N. (2013) Influence of Whooping Crane population dynamics on its recovery and management. *Biological Conservation*, **162**, 89–99.

Carroll, R.J., Ruppert, D., Stefanski, L.A. & Crainiceanu, C.M. (2006) *Measurement Error in Nonlinear Models: A Modern Perspective*. CRC Press, Boca Raton, FL, USA.

Cressie, N. & Wikle, C. (2011) *Statistics for Spatio-Temporal Data*. Wiley, Hoboken, NJ, USA.

Dorazio, R.M. (2012) Predicting the geographic distribution of a species from presence-only data subject to detection errors. *Biometrics*, **68**(4), 1303–1312.

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**(1), 677–697.

Elith, J., Kearney, M. & Phillips, S. (2010) The art of modelling range-shifting species. *Methods in Ecology and Evolution*, **1**(4), 330–342.

European Bird Census Council (2016) European Breeding Bird Atlas. Available at: http://www.ebcc.info/new-atlas.html (accessed 11 December 2016).

Fithian, W., Elith, J., Hastie, T. & Keith, D.A. (2015) Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, **6**(4), 424–438.

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**(4), 637–652.

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

Givens, G.H. & Hoeting, J.A. (2012) *Computational Statistics*. John Wiley & Sons, Hoboken, NJ, USA.

Gotway, C.A. & Young, L.J. (2002) Combining incompatible spatial data. *Journal of the American Statistical Association*, **97**(458), 632–648.

Graham, C.H., Elith, J., Hijmans, R.J., Guisan, A., Townsend Peterson, A. & Loiselle, B.A. (2008) The influence of spatial errors in species occurrence data used in distribution models. *Journal of Applied Ecology*, **45**(1), 239–247.

Guisan, A., Tingley, R., Baumgartner, J.B.*et al.* (2013) Predicting species distributions for conservation decisions. *Ecology Letters*, **16**(12), 1424–1435.
Hefley, T.J. & Hooten, M.B. (2016) Hierarchical species distribution models. *Current Landscape Ecology Reports*, **1**(2), 87–97.
Hefley, T.J., Tyre, A.J., Baasch, D.M. & Blankenship, E.E. (2013) Nondetection sampling bias in marked presence-only data. *Ecology and Evolution*, **3**(16), 5225–5236.
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**(3), 207–214.
Hefley, T.J., Baasch, D.M., Tyre, A.J. & Blankenship, E.E. (2015a) Data from: Use of opportunistic sightings and expert knowledge to predict and compare Whooping Crane stopover habitat. *Dryad Digital Repository*, http://datadryad.org/resource/doi:10.5061/dryad.t6859
Hefley, T.J., Baasch, D.M., Tyre, A.J. & Blankenship, E.E. (2015b) Use of opportunistic sightings and expert knowledge to predict and compare Whooping Crane stopover habitat. *Conservation Biology*, **29**, 1337–1346.
Hefley, T.J., Brost, B.M. & Hooten, M.B. (2017) Data from: Bias correction of bounded location errors in presence-only data. *Dryad Digital Repository*, https://doi.org/10.5061/dryad.82qd0
Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. & Jarvis, A. (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**(15), 1965–1978.
Homer, C.G., Dewitz, J.A., Yang, L.*et al.* (2015) Completion of the 2011 National Land Cover Database for the conterminous United States-Representing a decade of land cover change information. *Photogrammetric Engineering and Remote Sensing*, **81**(5), 345–354.
Hooten, M.B., Johnson, D.S., McClintock, B.T. & Morales, J.M. (2017) *Animal Movement: Statistical Models for Telemetry Data*. CRC Press, Boca Raton, FL, USA.
Kéry, M. & Royle, J.A. (2016) *Applied Hierarchical Modeling in Ecology: Analysis of Distribution, Abundance and Species Richness in R and BUGS: Volume 1: Prelude and Static Models*. Academic Press, Cambridge, MA, USA.
Luoto, M., Virkkala, R., & Heikkinen, R.K. (2007) The role of land cover in bioclimatic models depends on spatial resolution. *Global Ecology and Biogeography*, **16**(1), 34–42.
MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Royle, A.J. & Langtimm, C.A. (2002) Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, **83**(8), 2248–2255.
Manly, B., McDonald, L., Thomas, D., McDonald, T.L. & Erickson, W.P. (2004). *Resource Selection By Animals: Statistical Design and Analysis for Field Studies*. Springer Science & Business Media, Berlin, Germany.
McDonald, L., Manly, B., Huettmann, F. & Thogmartin, W. (2013) Location-only and use-availability data: analysis methods converge. *Journal of Animal Ecology*, **82**(6), 1120–1124.
Mitchell, P.J., Monk, J. & Laurenson, L. (2017) Sensitivity of fine-scale species distribution models to locational uncertainty in occurrence data across multiple sample sizes. *Methods in Ecology and Evolution*, **8**(1), 12–21.
Montgomery, R.A., Roloff, G.J., Ver Hoef, J.M. & Millspaugh, J.J. (2010) Can we accurately characterize wildlife resource use when telemetry data are imprecise? *The Journal of Wildlife Management*, **74**(8), 1917–1925.
Montgomery, R.A., Roloff, G.J. & Ver Hoef, J.M. (2011) Implications of ignoring telemetry error on inference in wildlife resource use models. *The Journal of Wildlife Management*, **75**(3), 702–708.
Naimi, B., Skidmore, A.K., Groen, T.A. & Hamm, N.A. (2011) Spatial autocorrelation in predictors reduces the impact of positional uncertainty in occurrence data on species distribution modelling. *Journal of Biogeography*, **38**(8), 1497–1509.
Nebraska Department of Natural Resources (2016) Nebraska sections boundary database. Available at: http://www.dnr.ne.gov/boundaries-plss (accessed 11 December 2016).
Pearse, A.T., Brandt, D.A., Harrell, W.C., Metzger, K.L., Baasch, D.M. & Hefley, T.J. (2015) Whooping crane stopover site use intensity within the Great Plains. Technical report, US Geological Survey.
Pearse, A.T., Harner, M.J., Baasch, D.M., Wright, G.D., Caven, A.J. & Metzger, K.L. (2017) Evaluation of nocturnal roost and diurnal sites used by whooping cranes in the Great Plains, United States. Technical report, US Geological Survey.
Platte River Recovery Implementation Program (2012) Whooping Crane minimum habitat criteria descriptions. Available at: https://www.platteriverprogram.org/PubsAndData/ProgramLibrary/PRRIP%202012_WC%20Min%20Habitat%20Criteria_DRAFT.pdf (accessed 04 July 2016).
Renner, I.W. & Warton, D.I. (2013) Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. *Biometrics*, **69**(1), 274–281.
Renner, I.W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S.J., Popovic, G. & Warton, D.I. (2015) Point process models for presence-only analysis. *Methods in Ecology and Evolution*, **6**(4), 366–379.
Robinson, W.S. (1950) Ecological correlations and the behavior of individuals. *American Sociological Review*, **15**(3), 351–357.
Royle, J.A. (2004) N-mixture models for estimating population size from spatially replicated counts. *Biometrics*, **60**(1), 108–115.
Scales, K.L., Hazen, E.L., Jacox, M.G., Edwards, C.A., Boustany, A.M., Oliver, M.J., & Bograd, S.J. (2017) Scale of inference: on the sensitivity of habitat models for wide-ranging marine predators to the resolution of environmental data. *Ecography*, **40**(1), 210–220.
Stoklosa, J., Daly, C., Foster, S.D., Ashcroft, M.B. & Warton, D.I. (2015) A climate of uncertainty: accounting for error in climate variables for species distribution models. *Methods in Ecology and Evolution*, **6**(4), 412–423.
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**(6), 1790–1801.
U.S. Fish and Wildlife Service (1978) Determiniation of critical habitat for the Whooping Crane. *Federal Register*, **43**, 20938–20942.
U.S. Fish and Wildlife Service (2016) Whooping Crane survey results: winter 2015–2016. Available at: www.fws.gov/uploadedFiles/Region_2/NWRS/Zone_1/Aransas-Matagorda_Island_Complex/Aransas/Sections/What_We_Do/Science/Whooping_Crane_Updates_2013/WHCR%20Update%20Winter%202015-2016.pdf (accessed 04 July 2016).
Velásquez-Tibatá, J., Graham, C.H. & Munch, S.B. (2016) Using measurement error models to account for georeferencing error in species distribution models. *Ecography*, **39**(3), 305–316.
Warton, D.I. & Shepherd, L.C. (2010) Poisson point process models solve the pseudo-absence problem for presence-only data in ecology. *The Annals of Applied Statistics*, **4**(3), 1383–1402.
Warton, D.I., Renner, I.W., & Ramp, D. (2013) Model-based control of observer bias for the analysis of presence-only data in ecology. *PLoS One*, e79168.
Wikle, C.K. & Berliner, L.M. (2005) Combining information across spatial scales. *Technometrics*, **47**(1), 80–91.

Received 20 March 2017; accepted 5 April 2017
Handling Editor: Robert B. O'Hara

Supporting Information

Details of electronic Supporting Information are provided below.

Appendix S1. Tutorial and R code to reproduce the Whooping Crane data example and Figures 1 and 2.

Appendix S2. In-depth explanation of the change of support and Figure S1.

Appendix S3. Reproducible simulation study and results.