Statistical Report

Ecology, 0(0), 2020, e03204

Integrating distance sampling and presence-only data to estimate species abundance

Matthew T. Farr [1,2,4] David S. Green [1,2,3] Kay E. Holekamp,[1,2] and Elise F. Zipkin [1,2]

[1]*Department of Integrative Biology, Michigan State University, East Lansing, Michigan 48824 USA*
[2]*Ecology, Evolution, and Behavior Program, Michigan State University, East Lansing, Michigan 48824 USA*
[3]*Institute for Natural Resources, Oregon State University, Corvallis, Oregon 97331 USA*

Citation: Farr, M. T., D. S. Green, K. E. Holekamp, and E. F. Zipkin. 2020. Integrating distance sampling and presence-only data to estimate species abundance. Ecology 00(00):e03204. 10.1002/ecy.3204

Abstract. Integrated models combine multiple data types within a unified analysis to estimate species abundance and covariate effects. By sharing biological parameters, integrated models improve the accuracy and precision of estimates compared to separate analyses of individual data sets. We developed an integrated point process model to combine presence-only and distance sampling data for estimation of spatially explicit abundance patterns. Simulations across a range of parameter values demonstrate that our model can recover estimates of biological covariates, but parameter accuracy and precision varied with the quantity of each data type. We applied our model to a case study of black-backed jackals in the Masai Mara National Reserve, Kenya, to examine effects of spatially varying covariates on jackal abundance patterns. The model revealed that jackals were positively affected by anthropogenic disturbance on the landscape, with highest abundance estimated along the Reserve border near human activity. We found minimal effects of landscape cover, lion density, and distance to water source, suggesting that human use of the Reserve may be the biggest driver of jackal abundance patterns. Our integrated model expands the scope of ecological inference by taking advantage of widely available presence-only data, while simultaneously leveraging richer, but typically limited, distance sampling data.

Key words: *black-backed jackal; data integration; distance sampling; integrated modeling; integrated distribution model; integrated species distribution model; presence-only model.*

Introduction

Using multiple data sources can improve biological inferences and predictions on the abundance and dynamics of wildlife populations. Yet, inconsistencies in study designs, spatiotemporal extents, and/or observation processes of independent data sources can lead to challenges in analyses that make use of multiple data types. Integrated modeling (i.e., data integration) is a powerful framework that uses a wide variety of methods and data types to estimate species parameters describing demography and/or abundance within a unified analysis (Maunder and Punt 2013, Zipkin and Saunders 2018, Miller et al. 2019, Isaac et al. 2020). The general principle behind integrated modeling involves construction of a joint likelihood linking individual data sets through a common biological process such that one or more parameters are informed by multiple data sources (Fletcher et al. 2016). Recent developments in data integration take advantage of point process models to combine disparate data to estimate population density across space (Dorazio 2014, Fithian et al. 2015, Fletcher et al. 2016, Koshkina et al. 2017).

Spatial point process models analyze the density (i.e., intensity) of data points (Warton and Shepard 2010). These models have a long history in ecology, primarily being used to estimate population abundance relative to the effects of environmental covariates (Cressie 1993, Warton and Shepherd 2010). Many spatial point process models utilize unstructured presence-only data, a commonly collected data type in population studies (Phillips et al. 2009). Presence-only data are characterized by opportunistic or incidental sighting of individuals from a target population, typically recorded without a formal study design. While such information can be valuable, particularly for rare or elusive species, presence-only

Manuscript received 27 December 2019; revised 20 July 2020; accepted 7 August 2020. Corresponding Editor: Viviana Ruiz-Gutierrez.
[4] E-mail: farrmat1@msu.edu

data are generally of lower quality than data collected from structured sampling efforts because they contain sampling biases and lack direct information on absences (Phillips et al. 2009, Fithian et al. 2015). As such, applications of point process models using presence-only data can, at best, only generate estimates of relative (rather than absolute) abundance as they do not account for observation errors (i.e., nondetections are not recorded; Fithian et al. 2015). Newer modeling techniques using thinned point processes have focused on accounting for systematic sampling biases in presence-only data and on combining presence-only data with structured sampling data (e.g., detection–nondetection data, count data) thereby improving inferences that could be obtained from either data type alone (Dorazio 2014, Fithian et al. 2015, Fletcher et al. 2016, Koshkina et al. 2017).

We developed an integrated framework for combining presence-only and distance sampling data to estimate spatially explicit abundance patterns. We included effects of environmental covariates on the biological process as well as covariates that may differentially affect the observational processes generating the two data types. Unlike previous implementations of integrated point process models, we developed our model by discretizing space, which allowed us to readily implement our approach within a Bayesian framework. We validated our model across a wide range of realistic parameter values and quantities of data with a simulation study. We then demonstrated the utility of our approach with a case study of black-backed jackals (*Canis mesomelas*) in the Masai Mara National Reserve, Kenya. Certain regions of the Reserve experience minimal management enforcement and are thus strongly affected by anthropogenic disturbances, which may be benefitting jackals (Farr et al. 2019). By combining available distance sampling data with opportunistic presence-only data, we estimated abundance of black-backed jackals in the Reserve to evaluate spatially varying effects of disturbance using relevant covariates.

Model Description

Biological process

We modeled abundance of a target species using a spatial point process (Dorazio 2014, Renner et al. 2015) by assuming that abundance in region A, denoted N_A, is the realization of a Poisson process: $N_A \sim \text{Poisson}(\mu_A)$. The parameter μ_A is the mean expected abundance across region A, defined as: $\mu_A = \int_A \lambda(s)ds$ in which $\lambda(s)$ is the expected count at a specific location, s, within A. We discretized space within region A using pixels $1,2,\ldots,G$ (Baddeley et al. 2010). Discretizing space can be computationally costly but may be more tractable for modeling abundance because (1) observation processes are realized in discrete space and time and (2) environmental covariates are typically measured in discrete space and time.

Total abundance in each pixel, N_g (in which $g \in A$), is also a Poisson random variable, which can be denoted $N_g \sim \text{Poisson}(\lambda_g)$ where λ_g is the expected abundance of individuals in pixel g. Thus, expected population abundance in region A, μ_A, is approximated as $\sum_{g=1}^{G} \lambda_g$. Spatial variation in expected abundance is driven by covariates that change across pixels, allowing us to model the effects of a heterogeneous landscape or habitat. Pixel resolution should be defined to appropriately capture spatial variation in abundance and covariates specific to the system of interest (Baddeley et al. 2010). We modeled covariate effects on λ_g using a log-link function with a linear predictor describing the effects of each variable, $\log(\lambda_g) = \log(\lambda_0) + \boldsymbol{\beta} \cdot \boldsymbol{w}_g$. The parameter λ_0 is the intercept, which can be interpreted as the baseline intensity of region A, and $\boldsymbol{\beta}$ is the vector of effect parameters for each corresponding covariate $\mathbf{w}_g$ at pixel g. Variation in environmental covariates, $\mathbf{w}_g$, results in an inhomogeneous Poisson point process as expected abundance of the target species, λ_g, varies across the region (Koshkina et al. 2017).

Observation processes

Opportunistic sampling: presence-only data.—By definition, opportunistic sampling lacks a structured design and such data are generated when information is collected haphazardly. In the case of population studies, opportunistic sampling is generally in the form of "presence-only" data on either occurrence (i.e., the species is present) or abundance (i.e., the number of individuals of the species that are observed) of the target population in specific locations. With presence-only data, no information on non-occurrences (i.e., zero counts) or locations sampled are recorded. Presence-only data are common in museum and herbarium collections, public science programs, and auxiliary data collection (Phillips et al. 2009, Fithian et al. 2015). Additionally, biases in opportunistic sampling data can result from uneven sampling intensity across a study area. For example, many public science programs typically collect data near roads or urban areas and information on sampling intensity is not always recorded.

To link presence-only count data (i.e., instances when ≥ 1 individuals observed within a pixel are recorded) to true abundance, we corrected for observation errors (e.g., sampling bias, imperfect detection) using a spatially explicit Poisson thinning process (Dorazio 2014). We assumed that the presence-only count data cover a subset of area, B, of the region of interest, A ($B \subseteq A$). Counts recorded at a pixel, y_g, are the realization of a binomial process where $y_g \sim \text{binomial}(N_g, p_g)$, in which N_g is the true latent abundance in pixel g, with expected abundance, λ_g (as defined in the biological process model), and p_g is a value between 0 and 1 that describes the observation error. Observation error is the

combination of imperfect detection and variable sampling intensity. Thus, we assumed the number of observed individuals at a pixel is less than or equal to the number of individuals occurring within a pixel. This binomial-Poisson mixture model reduces to a thinned point process model: $y_g \sim \text{Poisson}\left(\lambda_g \cdot p_g\right)$ (Appendix S1). Spatial variation in observation error can be modeled using a logit-link function with a linear predictor, $\text{logit}(p_g) = \text{logit}(p_0) + \boldsymbol{\alpha} \cdot \mathbf{z}_g$, where p_0 is the intercept parameter, and $\boldsymbol{\alpha}$ is the vector of effect parameters of each of the corresponding covariates, $\boldsymbol{z}_g$, that vary by pixel g to represent both imperfect detection and variable sampling intensity.

Distance sampling.—Distance sampling is a technique used to estimate species abundance and/or density by recording counts of individuals along transects and using their distance from the transect line to estimate a detection function (Buckland et al. 1993). In distance sampling, detection is assumed to be perfect on the transect line and then decay as a function of increasing distance from the line. Calculating the detection probability allows for estimation of the number of individuals that were not observed, and hence true latent abundance of a species. Covariates can be included to account for variations in abundance and/or the detection process across sampled locations.

For distance sampling data, we again corrected for imperfect detection by using a spatially explicit Poisson thinning process (Appendix S1). The number of individuals observed, x_g, in each pixel g is the realization of a binomial process, $x_g \sim \text{binomial}\left(N_g, \pi_g\right)$ where π_g is the average detection probability in g (ranging between 0 and 1). This binomial-Poisson mixture model also reduces to a thinned point process model where $x_g \sim \text{Poisson}\left(\lambda_g \cdot \pi_g\right)$. We calculated the average detection probability of each pixel, π_g, using the distance from the pixel midpoint to the transect line and the half-normal distribution $\pi_g = \exp\left(-d_g^2/2\sigma_g^2\right)$, where d_g is the distance between the midpoint of pixel g and the transect line and σ_g is the scale parameter of the half-normal distribution for the pixel. Covariates that influence detection across pixels can be used to model σ_g with a log-link function. We assumed that distance sampling spans an area, C, within our region of interest that is less than or equal to the size of region A ($C \leq A$).

Integrated model

To form the integrated model, we assumed independence between presence-only and distance sampling data. Although the two data types describe the same population, we found that an assumption of independence among data sets minimized computational challenges and did not affect model inferences (Appendix S1). We linked the two observation processes by assuming a joint, underlying biological process for the latent abundance model describing the expected number of individuals at each pixel g, $\log(\lambda_g) = \log(\lambda_0) + \boldsymbol{\beta} \cdot \mathbf{w}_g$. We obtained the joint likelihood of the integrated model for the two observation processes, $L_{IM}(p_0, \boldsymbol{\alpha}, \sigma_g, \lambda_0, \boldsymbol{\beta} | \boldsymbol{z}_g, y_g, d_g, x_g, \boldsymbol{w}_g)$, *as the product of the separate likelihoods for presence-only data,* $L_{PO}(p_0, \boldsymbol{\alpha}, \lambda_0, \boldsymbol{\beta} | \boldsymbol{z}_g, y_g, \boldsymbol{w}_g)$, *and distance sampling data,* $L_{DS}(\sigma_g, \lambda_0, \boldsymbol{\beta} | d_g, x_g, \boldsymbol{w}_g)$. The parameters λ_0 and $\boldsymbol{\beta}$ describe the shared biological process, characterizing true latent abundance across the study area, which occurs independently of how data are collected.

Data from opportunistic and distance sampling collection processes may or may not come from the same area ($B = C$ or $B \neq C$) within region A. Ideally, at least a subset of region A will be sampled by both approaches ($B \cap C$) although this is theoretically not required. Additionally, the different data types may be collected at different scales leading to a spatiotemporal mismatch between data sets. To integrate data collected at different scales, a single spatiotemporal resolution must be achieved by rescaling the data through change-of-support procedures (Pacifici et al. 2019). Ideally, the biological process should be described at the smaller spatiotemporal resolution of the two data sets to allow for better numerical approximation of the point process (Baddeley et al. 2010). The data set with the larger resolution is then matched to the scale of the biological process through summation, $v_b \sim \text{Poisson}\left(\sum_{g=1}^{|b|} \lambda_g\right)$, where v_b is the observation at pixel b of the larger resolution data and $|b|$ is the number of smaller resolution pixels g within b ($g \in |b|$).

Model parameters can be estimated in either a frequentist or Bayesian framework using the joint likelihood. The discretization of space in our model facilitates the use of standard Bayesian software for parameter estimation (e.g., JAGS; Plummer 2003). Previous thinned spatial point process models using continuous space have relied on frequentist frameworks for analysis (Dorazio 2014, Fithian et al. 2015).

Simulation Study

We conducted a simulation study to evaluate the accuracy and precision of our integrated model (Data S1). We verified our model's ability to estimate abundance by examining its capacity to recover an intercept parameter, λ_0, and a single effect parameter, β_1 of a covariate w_g. Model performance was evaluated across a range of environmental conditions (i.e., λ_0, β_1) and observation error (i.e., p_0, α_1, z_g, σ) by drawing each parameter and covariate from a distribution of realistic values (Appendix S2). We varied the amount of distance sampling data by changing the subset of area C (i.e., area covered by distance sampling transects) as percentage of region A (i.e., 0%, 5%, 10%, 15%, and 20%) while assuming either high or low quantities of presence-only data, which we simulated by altering the intensity of opportunistic sampling (p_0, w_g; Appendix S2).

Our simulated data sets followed the assumptions laid out in the descriptions of the biological and observation processes (Fig. 1A, B). For a single simulation, we first drew the environmental parameter values to create true latent abundance values in region A. We then analyzed data sets generated using all combinations of data quantities (e.g., high and low presence-only with 0%, 5%, 10%, 15%, and 20% distance sampling). We ran 1,000 simulations for each of the 10 scenarios (10,000 total simulations) and estimated model parameters using a Bayesian framework within program JAGS (version 4.2.0; Plummer 2003) with the jagsUI wrapper (version 1.4.2; Kellner 2016) and program R (version 3.4.1; R Core Team 2017). We used Gelman-Rubin diagnostics to check for convergence. For each parameter, we saved true values along with the mean value from the estimated posterior distribution of each fitted model. We examined the bias in parameter estimates by calculating the estimated value minus the true value, such that positive values indicate overestimation and negative values indicate underestimation.

Our simulation results demonstrate that the quantity of each data type determines the accuracy and precision of λ_0 and β_1, and hence estimates of abundance (Fig. 1C–F). Absolute abundance cannot be estimated with presence-only data alone (i.e., 0% distance sampling coverage), as the intercept parameter (λ_0) is unidentifiable (Fig. 1C, D; Dorazio 2014, Koshkina et al. 2017). However, both the accuracy and precision of the estimated intercept improve significantly as the amount of integrated distance sampling data is increased. Although presence-only data can accurately estimate the effect of a covariate (β_1) on latent abundance (Fig. 1E, F), precision improves with increasing amounts of distance sampling data, especially when there is only a limited amount of presence-only data (Fig. 1F).

Although estimates of detection probability within the distance sampling component of the model can be accurately recovered, interpretation of the observation error for presence-only data (p_g and parameters therein) is not recommended because imperfect detection and sampling bias cannot be parsed apart (Appendix S1). As such, the estimates of p_0 and α_1 are not meaningful (Appendix S2) and become less reliable within an integrated framework as λ_g becomes identifiable because $y_g \sim \text{Poisson}\left(\lambda_g \cdot p_g\right)$ (Dorazio 2014). Thus, we caution readers against interpretation of specific parameters within the linear predictor of p_g.

Application

The objective of our case study is to identify and evaluate environmental factors influencing black-backed jackal abundance across disturbed and undisturbed regions of the Masai Mara National Reserve using the integrated presence-only and distancing sampling model. The Reserve is a protected area on the southern border of Kenya and is home to a diverse mammalian community. Despite its protected status, some species in the Reserve are declining (Green et al. 2018, Green et al. 2019). Inconsistency in enforcement of management regulations (i.e., active vs. passive enforcement) across the Reserve results in spatial variation in human disturbance. Consequently, some wildlife species may be declining, while others, such as the black-backed jackal, seem to be thriving despite, or perhaps because of, changes to the landscape (Green et al. 2018, Farr et al. 2019).

To assess wildlife populations in the Reserve, distance sampling was conducted at monthly intervals between July 2012 and March 2014, during which 145 jackal sightings were recorded. Distance sampling transects were designed to accommodate driving restrictions and did not adhere to the assumption of nonrandom sampling, which can affect abundance estimates (Buckland et al. 1993); however, previous work using these data did not find biases caused by the design of the transects (see Farr et al. 2019 and Appendix S3 for data collection methods). A previous community distance sampling model revealed higher jackal density in the disturbed eastern region than in the relatively undisturbed western region, but limited data prevented precise estimation of other covariates (Farr et al. 2019). During the same time period, opportunistic sampling of jackals occurred daily by researchers who were primarily observing spotted hyena (*Crocuta crocuta*) behaviors. There were 2,669 jackal sightings recorded during opportunistic sampling, which is over 18 times the number of observations with distance sampling.

We adapted the basic structure of our integrated model to address a change-of-support problem between the temporal and spatial scales of our two data types (Fig. 2A, Appendix S3). To match the temporal scales between the observation processes described by the two data sources (monthly distance sampling vs. daily presence-only), we summarized presence-only data at monthly intervals. To avoid double counting individuals across and within pixels during a month, we used a spatial resolution of 1 km^2 for the presence-only data and the maximum group size (i.e., total number of individual jackals seen together) across observations within each pixel, which resulted in a total of 1655 recorded sightings. Distance sampling protocols prevented double counting of individuals across and within pixels; thus, we used a spatial resolution of 2,500 m^2 (50 × 50 m), which allowed us to capture observational and biological variation within transects (Fig. 2A). Given that jackals live in pairs (Moehlman 1979), we assumed abundance could be described using a Poisson distribution. Species with larger social groupings may require a different distribution or a group size submodel. We modeled the biological process as $\log(\lambda_g) = \log(\lambda_0) + \boldsymbol{\beta} \cdot \mathbf{w}_g$ at the smaller spatial resolution (2,500 m^2). When changing spatial support to 1 km^2 to match the scale of the presence-only data, we summed the Poisson random variables to scale up the intensity function by a factor of 400 (2,500 m^2 •

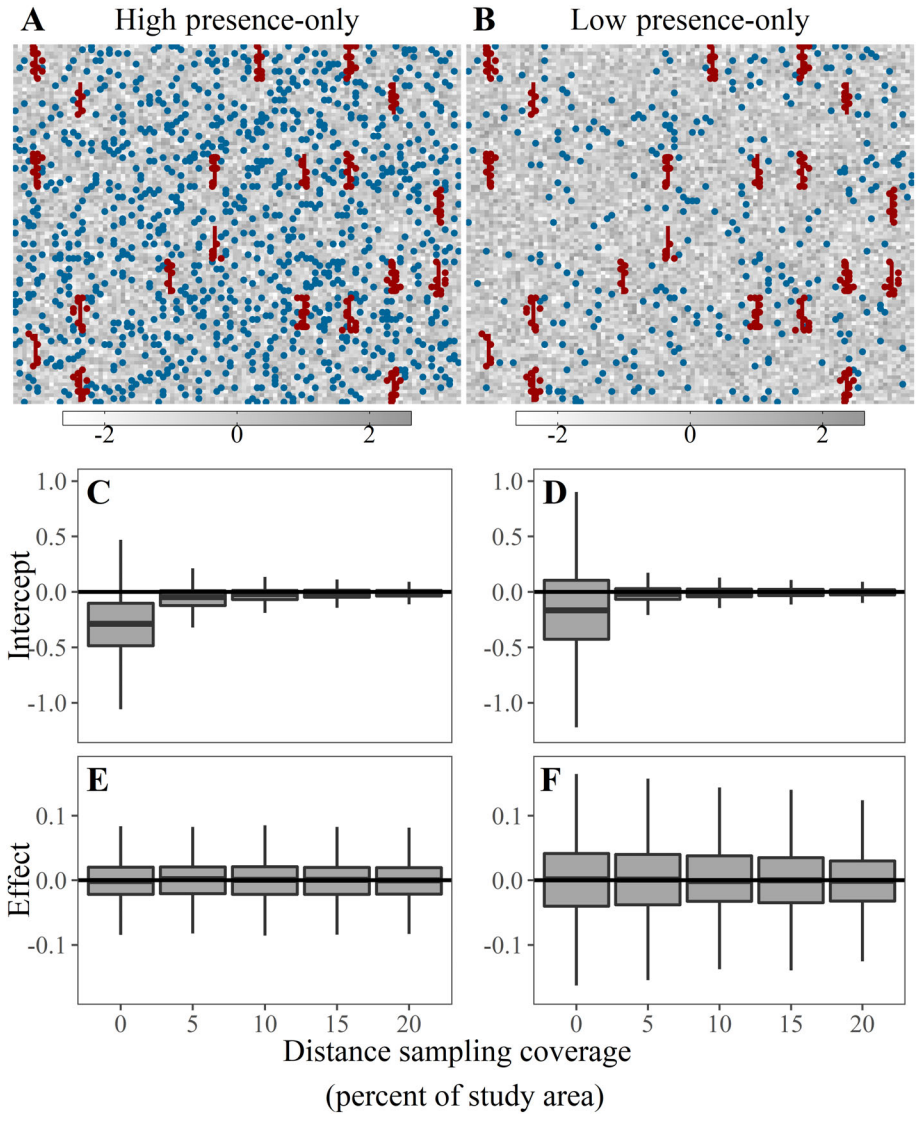

FIG. 1. (A, B) Visualization of the region of interest for a single simulation of high (left column; A) and low (right column; B) presence-only data shown with 20% distance sampling coverage. The gray background shows the spatial variation of a standardized environmental covariate that drives abundance. Blue and red dots represent presence-only and distance sampling data, respectively, and red lines show distance sampling transects. (C–F) Simulation results showing bias (estimated minus true value) for the intercept (C, D) and covariate effect (E, F) parameters for high (C, E) and low (D, F) presence-only data across a range (0–20%) of distance sampling coverage (0% coverage is presence-only data alone). The solid black lines at zero indicate no bias, where negative and positive values indicate under- and overestimations, respectively. Center lines of boxes show the median value of 1,000 simulations. Boxes contain the interquartile range. Vertical lines indicate values within ±1.5 times the interquartile range.

$400 = 1\ \text{km}^2$), $y_b \sim Poisson\left(\left(\sum_{g=1}^{400}\lambda_g\right)\bullet p_b\right)$, where b is a 1-km^2 pixel that contains 400 g pixels at 2,500 m^2. We also assumed that temporary emigration out of the study area was random and would thus have negligible effects on parameter estimation (Chandler et al. 2011).

We defined the study area as the combined extent of our two data sources (Appendix S3) and included in our model several environmental variables that might influence jackal abundance: disturbance regime coded as a binary indicator of whether a pixel was in the western (i.e., undisturbed) or the eastern (i.e., disturbed) region of the Reserve, distance to border as a proxy for the intensity of human disturbance, distance to a permanent water source, a vegetation index (i.e., normalized difference vegetation index [NDVI]), and an index of African lion (*Panthera leo*) density (the apex predator responsible for intraguild predation and competition, using kernel density estimate of lion sightings within each month), which were generally

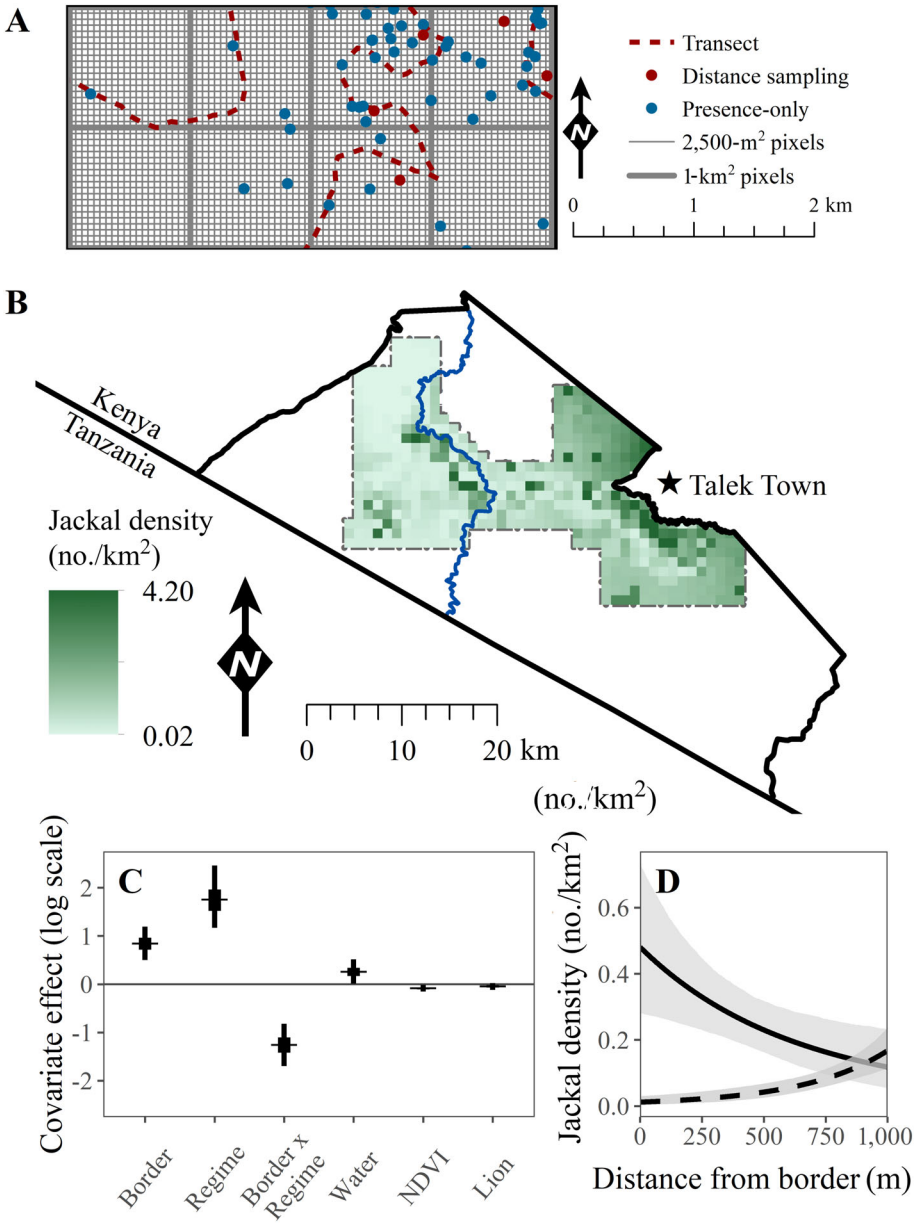

FIG. 2. (A) Visualization of the spatial change-of-support problem. Distance sampling occurs at a 2,500-m^2 resolution, while presence-only data are aggregated at a 1-km^2 resolution. (B) Spatially explicit abundance estimates (jackal density, no./km^2) of black-backed jackals in the Masai Mara National Reserve. Study area outlined with a thin dashed line. The undisturbed and disturbed regions are west and east of the Mara River (blue line), respectively. (C) Estimated covariate effects (log-scale) on black-backed jackal abundance. Small horizontal bars indicate posterior means; 50% and 95% credible intervals are displayed with thick and thin vertical bars, respectively. NDVI is the normalized difference vegetation index. (D) The relationship between jackal density and distance to border by management regime: eastern disturbed (solid line) vs. western undisturbed (dashed line) region. The shaded regions show 95% credible intervals.

lower in the disturbed region (Green et al. 2018, Farr et al. 2019). We also included an interaction effect between disturbance regime and distance to border, as we hypothesized that jackals within the disturbed region may be benefiting from being close to human activity. We included a random effect of pixel to account for variation among monthly replicates. To model the spatial heterogeneity in bias for opportunistic sampling, we added a spatially explicit and time-varying index of sampling intensity, as calculated by the sampling locations of hyena observations (n = 16,267) during the study period. We allowed the scale parameter for distance sampling to vary between disturbance regimes, as differences in grass height

(from differences in grazing intensity) may influence detection probability.

Using our integrated model (Data S2), we discovered that black-backed jackals have highest densities in the disturbed region closest to the border where human activity is greatest (Fig. 2B). The mean effect of management regime on jackal distribution was significantly positive (1.75; 95% CI = 1.17, 2.46; covariates are reported on the log scale; Fig. 2C), consistent with results using only distance sampling data demonstrating that jackal abundance is higher in the disturbed than undisturbed region (Farr et al. 2019). The mean effect of distance to border was positive (0.84; 95% CI = 0.50, 1.19; Fig. 2 C), but the interaction between distance to border and management regime was negative (−1.29; 95% CI = −1.71, −0.83; Fig. 2C). Thus, abundance of jackals within the disturbed region was highest along the border of the Reserve near the urban center of Talek Town, but within the undisturbed region, jackal abundance was lower near the border than in interior areas (Fig. 2D). Though the effects of NDVI (−0.09; 95% CI = −0.15, −0.02), distance to water source (0.26; 95% CI = 0.01, 0.51), and lion density (−0.04; 95% CI = −0.12, 0.02) on black-backed jackal abundance were precisely estimated, the magnitudes of these effects were close to zero (Fig. 2 C). Black-backed jackals may not be responding as strongly to bottom-up processes or variation in lion density as they do to human activity; however, lack of response to our index of lion density may indicate that jackal avoidance of lions occurs at smaller spatial and shorter temporal scales than those applied here.

Sparse data from distance sampling alone led to imprecise estimates of spatially explicit covariate effects. However, by integrating presence-only data into our analysis, we were able to demonstrate that jackals are likely benefiting from human disturbance and that their abundance patterns are influenced by a variety of landscape factors, including an interesting interaction with distance to the Reserve border.

Discussion

Using an integrated modeling framework, we combined presence-only and distance sampling data for the first time to accurately and precisely estimate species abundance and the effects of ecological covariates across space. Integrated point process models are useful in describing spatially varying population abundance and can be linked hierarchically to multiple data types such as presence-only, detection–nondetection, count (Miller et al. 2019, Isaac et al. 2020), and now distance sampling data. Our simulation study demonstrates how parameter estimates can be improved when supplementing presence-only data with distance sampling data. Regardless of whether low or high amounts of presence-only data were used, the addition of distance sampling data with fairly minimal coverage (10–15%) led to parameter identifiability of the intercept, λ_0, and hence the ability to estimate abundance (Fig. 1). In our case study, we augmented distance sampling data with opportunistic presence-only data leading to precise estimation of jackal abundance, including the effects of covariates, which was not possible using distance sampling data alone. The feasibility of our integrated framework depends on the availability of both data types and the ecology of the target species. Unlike data collected via structured sampling, presence-only data are not typically collected with the target species in mind as they are usually recorded as "incidental sightings." As we demonstrated in our case study, decisions on how to summarize and structure presence-only data (e.g., defining the spatiotemporal resolution) should be made in consideration of the ecological context of the study system. In some situations, a species' life history may be incompatible with the analysis and integration of presence-only data, such as in highly mobile species (i.e., violation of geographic closure) or species with high demographic turnover (i.e., violation of demographic closure).

Data integration is an increasingly popular analytical technique, but many challenges remain for widespread implementation (Fletcher et al. 2016, Isaac et al. 2020). Our simulation study revealed that parameters were identifiable in our integrated framework. However, parameter identifiability should be considered relative to the quantity and quality of available data, especially in cases where collinearity occurs between covariates on the observation and biological processes (Dorazio 2014). Our model estimated variable accuracy and precision of parameters across simulation scenarios with different amounts of data (Fig. 1C–F). Unstructured data, such as presence-only, is frequently collected opportunistically and researchers may have little control over their available quantity. However, a priori assessments, through power analyses or other simulation approaches, can help determine the optimal amount of required structured data to achieve specific estimation objectives and thereby optimize survey planning to reduce redundancies (Zipkin et al. 2021). Integrated point process models, such as the one we presented in this paper, make multiple assumptions about each data type (e.g., independence of data sources and individual points [Appendix S1]). Our simulation study followed the assumptions of our modeling structure. Researchers should carefully examine whether their systems meet these assumptions and conduct simulations to evaluate potential biases in cases when assumptions cannot be met completely (Renner et al. 2015).

The variability and complexity of biological systems, along with individual nuances of data collection, make it difficult to develop a universal integrated model, but general frameworks and guidelines for data integration can help ecologists tweak or customize models for their intended purposes (Saunders et al. 2019). Identifying the common currency (i.e., occurrence, abundance, point process) across data types and the parameters of interests will help narrow the choice of possible integration methods. We used a thinned spatial point process to link

different data types by describing abundance as a collection of individuals or points across space while accounting for differential observation processes between data sources. Point process models are a major advancement in the development of integrated modeling frameworks because they provide a straightforward statistical approach to unify data and an easy biological interpretation (Miller et al. 2019, Isaac et al. 2020). Such models can provide ecologists with valuable tools to tackle multidimensional problems with an array of data sources, now including distance sampling.

Acknowledgments

Development of the integrated modeling framework was supported by the National Science Foundation (NSF) grants EF-1702635 and DBI-1954406 to E. F. Zipkin. Collection of the case study data was supported by NSF grants OISE-1853934 to K. E. Holekamp and IOS-1755089 to K. E. Holekamp and E. F. Zipkin and the Lakeside Foundation and Kenya Wildlife Trust. We thank the Kenyan National Commission for Science, Technology and Innovation, the Narok County Government, The Mara Conservancy, and the Kenya Wildlife Service, and Brian Health for permission to do research and data collection. Daniel Greeson aided in processing the black-backed jackal data. Adam Duarte, Viviana Ruiz Gutiérrez, and two anonymous reviewers provided many useful comments on the paper. M. T. Farr and E. F. Zipkin conceived the idea for the research. K. E. Holekamp and D. S. Green coordinated the data collection for the case study. M. T. Farr and E. F. Zipkin led the analysis and wrote the manuscript with contributions from K. E. Holekamp and D. S. Green on multiple drafts. All co-authors accept responsibility of this work and have provided their approval for publication.

Literature Cited

Baddeley, A., M. Berman, N. I. Fisher, A. Hardegen, R. K. Milne, D. Schuhmacher, R. Shah, and R. Turner. 2010. Spatial logistic regression and change-of-support in Poisson point processes. Electronic Journal of Statistics 4:1151–1201.

Buckland, S. T., D. R. Anderson, K. P. Burnham, and J. L. Laake. 1993. Distance sampling: Estimating abundance of biological populations. Oxford University Press, Oxford, UK.

Chandler, R. B., J. A. Royle, and D. I. King. 2011. Inference about density and temporary emigration in unmarked populations. Ecology 92:1429–1435.

Cressie, N. A. C. 1993. Statistics for spatial data. John Wiley & Sons Inc., New York, New York, USA.

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

Farr, M. T., D. S. Green, K. E. Holekamp, G. J. Roloff, and E. F. Zipkin. 2019. Multispecies hierarchical modeling reveals variable responses of African carnivores to management alternatives. Ecological Applications 29:e01845.

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

Fletcher, R. J., R. A. McCleery, D. U. Greene, and C. A. Tye. 2016. Integrated models that unite local and regional data reveal larger-scale environmental relationships and improve predictions of species distributions. Landscape Ecology 31:1369–1382.

Green, D. S., L. Johnson-Ulrich, H. E. Couraud, and K. E. Holekamp. 2018. Anthropogenic disturbance induces opposing population trends in spotted hyenas and African lions. Biodiversity and Conservation 27:871–889.

Green, D. S., E. F. Zipkin, D. C. Incorvaia, and K. E. Holekamp. 2019. Long-term ecological changes influence herbivore diversity and abundance inside a protected area in the Mara-Serengeti ecosystem. Global Ecology and Conservation 20:e00697.

Isaac, N. J. B., et al. 2020. Data integration for large-scale models of species distributions. Trends in Ecology & Evolution. https://doi.org/10.1016/j.tree.2019.08.006.

Kellner, K. 2016. jagsUI: A wrapper around 'rjags' to streamline 'JAGS' analyses. R package version 1.4.2. https://cran.r-project.org/package=jagsUI

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

Maunder, M. N., and A. E. Punt. 2013. A review of integrated analysis in fisheries stock assessment. Fisheries Research 142:61–74.

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

Moehlman, P. D. 1979. Jackal helpers and pup survival. Nature 277:382–383.

Pacifici, K., B. Reich, D. Miller, B. Pease, and D. Huberman. 2019. Resolving misaligned spatiotemporal data with integrated distribution models. Ecology 100:e02709.

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

Plummer, M. 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *In* Proceedings of the 3rd International Workshop on Distributed Statistical Computing. Vienna, Austria.

R Core Team. 2017. R: A language and environment for statistical computing. R Project for Statistical Computing, Vienna, Austria. http://www.R-project.org/

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

Saunders, S. P., M. T. Farr, A. D. Wright, C. A. Bahlai, J. W. Ribeiro, S. Rossman, A. L. Sussman, T. W. Arnold, and E .F. Zipkin. 2019. Disentangling data discrepancies and deficiencies with integrated population models. Ecology 100:e02714.

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

Zipkin, E. F., and S. P. Saunders. 2018. Synthesizing multiple data types for biological conservation using integrated population models. Biological Conservation 217:240–250.

Zipkin, E. F., E. R. Zylstra, A. D. Wright, S. P. Saunders, A. Finley, M. C. Dietze, M. Itter, and M. W. Tingley. 2021. Addressing data integration challenges to link ecological processes across scales. Frontiers in Ecology and the Environment, in press.

Supporting Information

Additional supporting information may be found in the online version of this article at http://onlinelibrary.wiley.com/doi/10.1002/ecy.3204/suppinfo

Data Availability Statement

Code and data are available on Zenodo: https://doi.org/10.5281/zenodo.3981241