ECOGRAPHY

Research

Is more data always better? A simulation study of benefits and limitations of integrated distribution models

Emily G. Simmonds*, Susan G. Jarvis*, Peter A. Henrys, Nick J. B. Isaac and Robert B. O'Hara

E. G. Simmonds (https://orcid.org/0000-0002-3348-6153) ✉ (emilygsimmonds@gmail.com) and R. B. O'Hara (https://orcid.org/0000-0001-9737-3724), Dept of Mathematical Sciences and Centre for Biodiversity Dynamics, Norwegian Univ. of Science and Technology (NTNU), Trondheim, Norway. – S. G. Jarvis (https://orcid.org/0000-0001-5382-5135) and P. A. Henrys (https://orcid.org/0000-0003-4758-1482), UK Centre for Ecology & Hydrology, Bailrigg, Lancaster, UK. – N. J. B. Isaac (https://orcid.org/0000-0002-4869-8052), UK Centre for Ecology & Hydrology, Benson Lane, Crowmarsh Gifford, Wallingford, UK.

Ecography
43: 1413–1422, 2020
doi: 10.1111/ecog.05146

Subject Editor: Cory Merow
Editor-in-Chief: Miguel Araújo
Accepted 15 June 2020

Species distribution models are popular and widely applied ecological tools. Recent increases in data availability have led to opportunities and challenges for species distribution modelling. Each data source has different qualities, determined by how it was collected. As several data sources can inform on a single species, ecologists have often analysed just one of the data sources, but this loses information, as some data sources are discarded. Integrated distribution models (IDMs) were developed to enable inclusion of multiple datasets in a single model, whilst accounting for different data collection protocols. This is advantageous because it allows efficient use of all data available, can improve estimation and account for biases in data collection. What is not yet known is when integrating different data sources does not bring advantages. Here, for the first time, we explore the potential limits of IDMs using a simulation study integrating a spatially biased, opportunistic, presence-only dataset with a structured, presence–absence dataset. We explore four scenarios based on real ecological problems; small sample sizes, low levels of detection probability, correlations between covariates and a lack of knowledge of the drivers of bias in data collection. For each scenario we ask; do we see improvements in parameter estimation or the accuracy of spatial pattern prediction in the IDM versus modelling either data source alone? We found integration alone was unable to correct for spatial bias in presence-only data. Including a covariate to explain bias or adding a flexible spatial term improved IDM performance beyond single dataset models, with the models including a flexible spatial term producing the most accurate and robust estimates. Increasing the sample size of presence–absence data and having no correlated covariates also improved estimation. These results demonstrate under which conditions integrated models provide benefits over modelling single data sources.

Keywords: citizen science, data integration, integrated distribution models, simulations, species distribution models

NORDIC SOCIETY OIKOS

www.ecography.org

* joint first authors.

Introduction

Species distribution modelling has many applications in ecology and is now a mature discipline. In recent years, data mobilization, citizen science and a raft of new monitoring technologies have generated enormous growth in the data available for such models. Whilst these new data streams are welcome, they present challenges for species distribution modelling because each data source has different attributes, reflecting variation in protocols, spatial extent, sampling intensity and the time period over which they were collected. Confronted by this heterogeneity, modellers commonly face a choice over which data sources to use for a particular application.

Until recently, the standard approach was to favor one dataset and either discard others or use them in secondary analyses (e.g. model validation). Integrated distribution models (IDMs) have emerged as a way to avoid this choice (Fletcher et al. 2019, Isaac et al. 2019, Miller et al. 2019, Zipkin et al. 2019). The key feature of IDMs is that separate datasets are modelled in a way that is faithful to how they were generated. This is usually achieved by sharing parameters between datasets, often by treating each data source as a separate realisation of the true distribution (the 'joint-likelihood approach' (Pacifici et al. 2017)).

Integrated modelling has some obvious virtues, such as allowing us to make efficient use of all available data, but additional benefits are now becoming clear. Fithian et al. (2015) and Peel et al. (2019) have shown, in multi-species analyses, that information from a highly structured dataset (e.g. including non-detections or a standardized protocol) can be sufficient to overcome the spatial biases in presence-only data. Bowler et al. (2019) demonstrated that integrating datasets with different spatial footprints allowed predictions to be estimated over a wider spatial area, and with greater precision, than one dataset alone. Estimation of unbiased parameter values, higher precision of estimates and increased spatial coverage are some of the primary benefits suggested for IDMs. While there is no doubt that IDMs are a useful advance for species distribution modelling, it is not clear whether an integrated model is always better than analyzing datasets independently (Isaac et al. 2019).

Available evidence indicates that IDMs outperform models based on individual datasets under nearly all scenarios so far investigated (Fithian et al. 2015, Koshkina et al. 2017, Peel et al. 2019). However, the advantages of integrated models are unlikely to be universal, particularly when information from repeat surveys or multiple species are not available. We consider four challenges that might influence the performance of IDMs relative to single dataset models.

Firstly, there has been conflicting evidence on the usefulness of IDMs when high quality data (defined in terms of minimal error and/or minimal bias) is very limited relative to data with high error or bias. Although IDMs have been shown to be robust to small sample sizes (Peel et al. 2019), other studies have suggested the benefits of IDMs may be reduced when high quality data is limited, unless an appropriate weighting can be applied (Fletcher et al. 2019). In practical terms, we might assume that the benefit of integrating large quantities of opportunistic presence-only data with high quality data is most apparent when high quality data is low in volume. However, large discrepancies in the size of the datasets could also lead to domination of the results by a single source and reduce any meaningful gain from integration.

Secondly, if the probability of detection is very low in one data source then it may be more challenging to estimate model parameters correctly (Guillera-Arroita et al. 2014). Although IDMs have been shown to be robust to low detection compared to using solely presence-only data (Koshkina et al. 2017), it is not yet clear whether IDMs provide any advantage when detection probabilities are very low. We might expect low detection probabilities to be a particular challenge when data from repeat surveys are not available to estimate them explicitly.

Thirdly, if covariates that relate to the underlying state and those that affect the observation process, are correlated it may be impossible for IDMs to correctly identify both processes (Fithian et al. 2015).

Finally, many of the approaches demonstrated so far assume that any bias in species observations can be estimated with known covariates (Dorazio 2014) or is shared between species (Fithian et al. 2015). Here we consider that in some situations information on causes of bias may not be known or available for modellers, therefore it cannot be modelled explicitly.

We address these four challenges by conducting a simulation study integrating an unbiased presence–absence survey with a spatially biased presence-only dataset. We investigate under which conditions IDMs provide the greatest or least benefit over modelling each data source individually. We provide an indication of when IDMs are a useful tool, and when they may not improve inference over modelling datasets separately.

Material and methods

Scope of the simulation study

Our simulations assumed two classes of data are available to model the spatial distribution of a hypothetical species; the first dataset was spatially unbiased across the survey area, such as may arise from a spatially-balanced sampling design, and recorded both presences and absences (PA dataset). We assumed a second source was derived according to citizen–science type protocols where only species presence was recorded (PO dataset), detection was imperfect and detection probability was not uniformly distributed across the survey area causing a bias towards certain areas. Note that we assumed that variation in detection probability could arise from either sampling effort or changes in species detectability and did not consider these processes separately. The choice of only PA and PO data mirrors commonly available data in ecology and

allows our results to be comparable to previous simulation studies (Fithian et al. 2015, Koshkina et al. 2017, Peel et al. 2019). However, in contrast to these studies, we assumed data were only available for a single species and no repeat surveys were conducted. The number of locations surveyed in the PA dataset, the detection probability in the PO dataset and the correlation between the spatial bias in PO data and an environmental covariate, were all varied to address the challenges above.

The datasets were modelled independently and in an IDM. Additionally, for models including the PO data we optionally included a covariate explaining spatial variation in detection in the PO data. Excluding this covariate simulated a scenario where the source of spatial bias in PO data was unknown. Finally, we considered a model where the source of bias in PO data was unknown, but variation in detection could be modelled by a random spatial field.

Generating true species distributions

The first step was to generate a 'true' distribution of species presences from which to sample data used to parametrize our models. This was done in two phases. The first generated the 'true' intensity surface of species abundance. The second turned the 'true' intensity into a realisation of species presences, recreating actual locations of individuals. All simulations were conducted on a continuous square domain, D (where $D \subset \mathbb{R}^2$) with dimensions 300×300 (Fig. 1) and we define s to be the set of all locations within D.

The true species distribution was assumed to come from an inhomogeneous Poisson point process, a statistical model which describes the distribution of points over space with an intensity function (following; Cressie 1993, Dorazio 2014, Fithian et al. 2015, Koshkina et al. 2017, Peel et al. 2019). This intensity function describes the density of points (in our

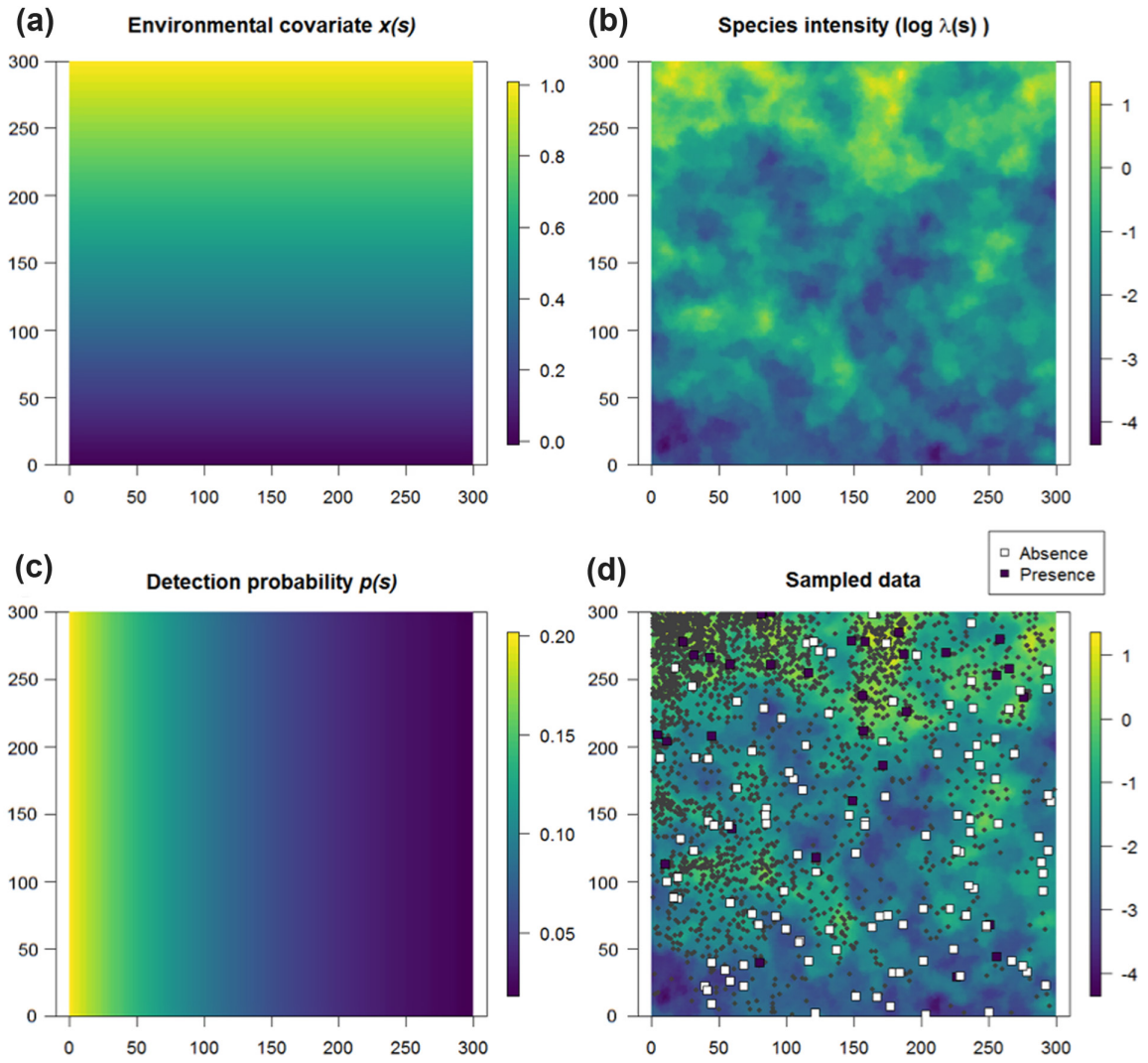

Figure 1. Layout of the simulated data showing (a) the environmental covariate, (b) a species intensity generated from a log-Gaussian Cox process, (c) spatial variation in detection probability and (d) simulated PO data (grey points) and PA data (white squares indicating absences and black squares indicating presences).

simulations we make the simplifying assumption that these are equivalent to individuals) in a given area. This is shown in equation 1, where the number of points in an area A comes from a Poisson distribution with the expectation being an integration of the intensity function $\lambda(s)$ over A. Conceptualizing species distributions as point processes is important to integrated modelling as it allows different currencies of species information e.g. counts, presence–absence or presence-only data, to be thought of as different ways of observing the same phenomena (Isaac et al. 2019, Miller et al. 2019).

Specifically, we assumed that the locations of individuals are governed by a log-Gaussian Cox process, which represents a doubly stochastic Poisson process as the intensity itself is stochastic. The intensity function was defined by Eq. 1:

$$N(A) \sim \text{Poisson}\left(\int_A \lambda(s)ds\right)$$

$$\log\left(\lambda(s)\right) = \alpha_0 + \beta_x x(s) + \xi(s) + \varepsilon(s) \quad (1)$$

where log of the intensity $\lambda(s)$ was defined by an intercept (α_0), a linear relationship ($\beta_x x(s)$) with an environmental covariate ($x(s)$), a Gaussian random field ($\xi(s)$) which simulated spatial variation not explained by the environmental covariate, and some random error ($\varepsilon(s)$). The single environmental covariate ($x(s)$) was defined as a continuous gradient from the bottom to the top of the grid, with values ranging from 0 to 1 (Fig. 1a). We used a very simple spatial structure for the environmental covariate here to facilitate the interpretation of results. We assumed the spatial variation in the random field had a Matérn covariance structure, which was governed by three parameters corresponding to the variance, scale (κ) and smoothness (ν).

The parameters used to generate the true species distribution did not vary between scenarios. The intercept (α_0) was set at -2 and an environmental covariate effect given a value of 2 (β_x). Therefore, at mean level of the environmental covariate ($x(s) = 0.5$) the intensity would have a mean of 0.37 individuals per unit area. The variance parameter of the covariance was set to 0.5, the scale parameter κ was set at 0.05 and the smoothness parameter ν was set to 1 (Fig. 1b). To generate a 'true' species intensity from this model we used the rLGCP function in the 'spatstat' package (Baddeley et al. 2015).

Simulation of sampling processes

After the 'true' intensity had been generated, the next step in the simulations was to sample from this truth, mimicking data collection in the field. Two sampling processes were simulated, one for the PO data and one for the PA data. A separate realisation was generated from the log-Gaussian Cox process for each sampling process to represent the fact that the individuals sampled by each method were unlikely to be the same i.e. data collection were at different points in time.

PO data were generated by thinning a realisation of the log-Gaussian Cox process to represent imperfect detection. This was done by creating a continuous gradient of detection probability $p(s)$, and sampling the individuals with their location-specific probability using a Bernoulli trial. The detection probability decreased across the spatial gradient to represent different levels of sampling effort or detectability e.g. relating to density of human populations or surveyor preferences. For most scenarios, the gradient of detection probability began at 0.2 and declined by a factor of ten (to 0.02) across the whole gradient and was perpendicular to the environmental covariate (Fig. 1c). All parameters were chosen to balance computational efficiency and sample size. Excessively large or small sample sizes led to slow computing times or inability to draw inference (e.g. no presences recorded). In these simulations the detection probability combined both the probability of visiting a location and the probability of seeing an individual if the location was visited. After thinning, an average of 3884 PO observations remained. Under the lowest detection scenario (Table 1), there were an average of 77 PO observations, for a full summary see Supplementary material Appendix 1.

It was assumed that detection probability was never known exactly, but was strongly correlated ($\rho = 0.99$) with a covariate, $z(s)$, available at all locations. As a result, the covariate on detection was a very good, but not a perfect, descriptor of the sampling process, reflecting the fact that it is unrealistic to have perfect knowledge of sampling bias. The difference between 0.99 and 1 correlation was detectable and is demonstrated in the Supplementary material Appendix 4.

To generate PA data the domain was split into 25 strata in a five by five grid to represent a stratified random sampling design and ensure equal coverage of each stratum. Within each stratum 6 'quadrats' (each 1×1 in dimension) were placed randomly (150 quadrats in total). If a point (assumed to represent an individual) was recorded in the quadrat, a presence was recorded (i.e. we assume perfect detection).

Table 1. Details of the scenarios run in this study.

Scenario	Number of PA samples	Maximum observation probability in PO data	Environmental covariate correlated with bias?
Baseline	150	0.2	No
Different sample sizes of PA data	50, 100, 150, 200, 250, 300, 350, 400, 450, 500	0.2	No
Different levels of observation probability in PO data	150	0.2, 0.16, 0.12, 0.8, 0.04, 0.02, 0.004	No
Environmental covariate correlated with bias	150	0.2	Yes

As the detection probability was the same in each of the 25 strata, the stratified design here was effectively the same as simple random sampling across the whole domain on average. None of our scenarios produced a PA sample size of less than 10 recorded presences (minimum was approximately 18 presences on average).

Details of scenarios

Each scenario was run 500 times to account for stochasticity in the data generation process. For each of the 500 simulations one 'true' intensity was generated and sampled to generate PO and PA datasets, all models were then run on the same datasets to ensure comparability of results.

The scenarios we use here were designed to address each of our four questions exploring the influence of; the size of the PA dataset, detection probability in the PO data, a correlation between an environmental covariate and detection probability in the PO data, and the way bias in PO data is modelled on the relative performance of different models. Table 1 presents an overview of the 4 scenarios and 17 parameter combinations used. Note that changing the maximum detection probability of the PO data is equivalent to changing the quantity of PO data.

Statistical modelling of the simulated data

To estimate the true intensity, as given in Eq. 1, we fitted six models with different properties to each data generation scenario. In this study, the models fell into two core model types; single and integrated. Single models included only a single dataset, either PA or PO (Table 2, models A, B and D). Integrated models modelled both PA and PO data simultaneously, using a joint likelihood (Table 2, models C, E and F). Within these model types we fit two groups of models; those that included information on bias in the PO data (Table 2, models D, E and F), and those that did not (Table 2, models B and C). We accounted for spatial bias present in the data by extending our models in two ways. The first included the covariate $z(s)$ to represent the spatial bias in observation probability (the combined probability of visiting an area and detection probability) in the PO data (Table 2, models D and E). The second included a second spatial field $\zeta(s)$ which was only informed by the PO data and should reflect spatial variation not explained by either the shared spatial field or environmental covariate (Illian 2017; Table 2, model F). Therefore, we investigate the ability of this second spatial field $\zeta(s)$ to account for spatial bias in the PO data.

All of the models assumed that the intensity surface ($\lambda(s)$) resulted in one or more Poisson point patterns which could have been observed in different ways (Bowler et al. 2019, Isaac et al. 2019). For each quadrat (i) where the PA data (Y_i) were sampled, presence was modelled as a single Bernoulli trial with probability of presence p_i. γ_i (within s) is the location of quadrat i. A cloglog link was used to link p_i to the log intensity of the Poisson process ($\lambda(s)$ evaluated at γ_i) (Eq. 2 (Kery and Royle 2016, Bowler et al. 2019)). The intercept is the baseline expected abundance when the environmental covariate equals zero.

$$\begin{aligned} Y_i &\sim \text{Bernoulli}\left(p_i\right) \\ \text{clog log}\left(p_i\right) &= \log\left(\lambda\left(\gamma_i\right)\right) \end{aligned} \tag{2}$$

We assume the PO data locations come from a Poisson point process as detailed in Eq. 3. Where the total number of presences in a sub-region A ($N(A)$) are Poisson distributed with intensity given by integrating $\lambda(s)$ over A. We modelled the locations of the observations of PO data as an log-Gaussian Cox point process following Renner et al. (2015) and Simpson et al. (2016). As a result, both data types could be modelled as originating from the same underlying state (defined by the intensity which we assume follows the form given in Eq. 1), but with different observation processes and therefore different intercepts. By fitting separate intercepts, we allow unexplained variation in observation

Table 2. Model types fit in this study parameters are indicated with a hat to distinguish them from the true parameters in Eq. 1. Here all parameters are model estimates of the true parameters. Integrated models include two predictors, one for the PA data and one for the PO data, each with their own intercept but with shared parameters. The cloglog link was used for PA data and log link for PO data to link the response to the predictor.

Model	Model description	Type	Response	Predictor
A	PA-only	Single	PA	$\alpha_{PA} + \hat{\beta}_x x(s) + \hat{\xi}(s)$
B	PO-only	Single	PO	$\alpha_{PO} + \hat{\beta}_x x(s) + \hat{\xi}(s)$
C	IDM	Integrated	PA, PO	(1) $\alpha_{PA} + \hat{\beta}_x x(s) + \hat{\xi}(s)$ (2) $\alpha_{PO} + \hat{\beta}_x x(s) + \hat{\xi}(s)$
D	PO-only with bias covariate	Single	PO	$\alpha_{PO} + \hat{\beta}_x x(s) + \hat{\beta}_z z(s) + \hat{\xi}(s)$
E	IDM with bias covariate	Integrated	PA, PO	(1) $\alpha_{PA} + \hat{\beta}_x x(s) + \hat{\xi}(s)$ (2) $\alpha_{PO} + \hat{\beta}_x x(s) + \hat{\beta}_z z(s) + \hat{\xi}(s)$
F	IDM with second spatial field	Integrated	PA, PO	(1) $\alpha_{PO} + \hat{\beta}_x x(s) + \hat{\xi}(s)$ (2) $\alpha_{PO} + \hat{\beta}_x x(s) + \hat{\xi}(s) + \hat{\zeta}(s)$

processes between datasets to be captured in each intercept term. We are therefore unable to estimate abundance from this integrated model as the intercepts capture more than the baseline intensity, however the estimated spatial pattern is unaffected.

$$N(A) \sim \text{Poisson}\left(\int_A \lambda(s)ds\right) \quad (3)$$

For estimating the true intensity $\lambda(s)$ all models had linear predictors (shown in Table 2) that included the environment gradient and a Gaussian Markov spatial random field (compare to Eq. 1). For the IDMs, these model terms were shared between data sets and predictions were made using the estimated intercept for the PA data as the PA data was assumed to have perfect detection. Note that the model fit to the PO data is slightly different to the data generation process. We do not model the thinning process explicitly (i.e. we do not directly estimate $p(s)$). This is because we allow each data source to have separate intercepts, meaning it is not possible to separate the intercept of the PO data from the detection probability. Instead of modelling the thinning process directly, additional components are added to the linear predictor of the intensity surface to account for spatial variation in observation probability: a covariate related to $p(s)$ is included in models D and E; and a second spatial field included in model F. This second spatial field is a latent spatial effect that should be a spatial representation of the observation probability of the PO data excluding residual environmental spatial structure.

We fit all models using approximate Bayesian inference through integrated nested Laplace approximation (INLA) (Rue et al. 2009) using the stochastic partial differential equation approach (Lindgren et al. 2011). INLA was chosen as it is an efficient method for modelling flexible spatial fields. Default priors were used for all parameters. Models were fit in R using the package R-INLA.

Evaluation of model performance

In this study, we assumed that using PA and PO data for a single species, without additional information, cannot return an accurate estimate of the intercept of the original 'true' intensity surface. We assumed this because our models did not estimate observation probability in either dataset, which is unlikely to be perfect in real ecological datasets. Instead, we estimated a relative pattern of occurrence, relative to the mean prediction.

Models were evaluated based on three metrics. Firstly, the accuracy and precision of the estimate of $\hat{\beta}_x x(s)$ was assessed to check how well model parameters were estimated. Secondly, the predicted intensity was compared to the true intensity by calculating the correlation between the two. This metric assessed how well each model captured the spatial pattern in species distributions. Lastly, the mean absolute error (MAE) between predicted and true intensities was calculated as the unsigned difference. Because all validation was conducted relative to the mean prediction, the MAE does not inform on the ability of the models to return the absolute intensity values. Rather, the MAE reflects the ability of the models to capture the variation in intensity.

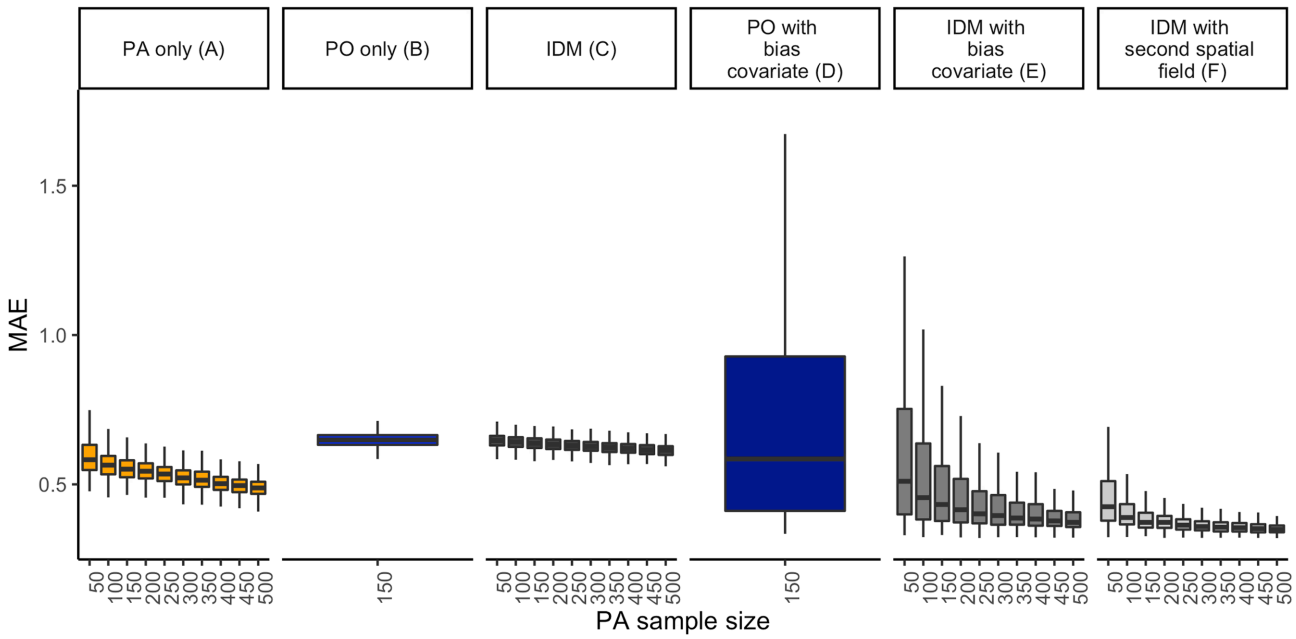

Figure 2. Boxplot of the MAE between predicted relative intensity and true relative intensity for all 500 simulations at each sample size of PA data. Each boxplot shows the interquartile range (25–75% quartiles), and the median of the simulated results. Whiskers of the boxplots extend to the largest or smallest point within 1.5 times the interquartile range from the edge of the boxplot. Outliers are not plotted.

Congruence between prediction and truth and MAE metrics were evaluated at 900-point locations in a regular 10 by 10 grid across the domain (effectively a stratified subsampling design).

Results

Sample size of PA data

There was little evidence that the IDM performed better than a model with PO data only in terms of mean absolute error (MAE) (Fig. 2, models B and C) and correlation (Supplementary material Appendix 2 Fig. A5) if no information on bias in PO data was included. This suggests that without repeat visits or multiple species, IDMs cannot compensate for bias in PO data simply by including unbiased PA data. The IDM (model C) always had higher MAE than the PA-only models (model A), and had lower correlation with the truth if the number of PA samples was over 200, indicating no advantage of joint models when reasonable amounts of PA data were available. The number of PA samples relative to the number of PO samples ranged from 1% to 13%, (PA sample size = 50 or 500, respectively).

Increasing quantities of PA data improved the accuracy of PA-only models as expected. IDMs only benefited from increased quantities of PA data if some information on bias in PO data was included. Both the IDM with a bias covariate (model E) and IDM with a second spatial field (model F) showed improvements in performance with more PA data. These integrated models performed better than PO-only models including the bias covariate (model D), but the improvement was quite small at low levels of PA data. This suggests integrated models do have improved performance over single dataset models in terms of accuracy of spatial predictions, but that the greatest improvement in performance comes from modelling the spatial bias in PO data (e.g. model B versus model E or model F) which requires additional information beyond the PA data alone (e.g. model C).

The environmental covariate was returned similarly in all models (Supplementary material Appendix 2 Fig. A6–A7). The PA data alone (model A) had the highest accuracy and precision in the estimate of the environmental covariate and both improved with more data. All other models did identify the correct direction of the covariate effect, but credible intervals frequently spanned zero. Although precision improved with greater amounts of PA data for models A, E and F, model C showed little change. All models except the PA only model tended to slightly overestimate $\hat{\beta}_x x(s)$ and this was largely unaffected by the amount of PA data.

Observation probability of PO data

The second scenario reduced the maximum observation probability of the PO data. In all models including the PO data the correlation between prediction and truth tended to decrease with lower maximum observation probability (Supplementary material Appendix 1 Fig. A1), as would be expected given fewer data points available to model. The MAE generally worsened with lower observation probability, but two unexpected patterns were detected (Fig. 3). Firstly, at very low observation probability the simple IDM (model C)

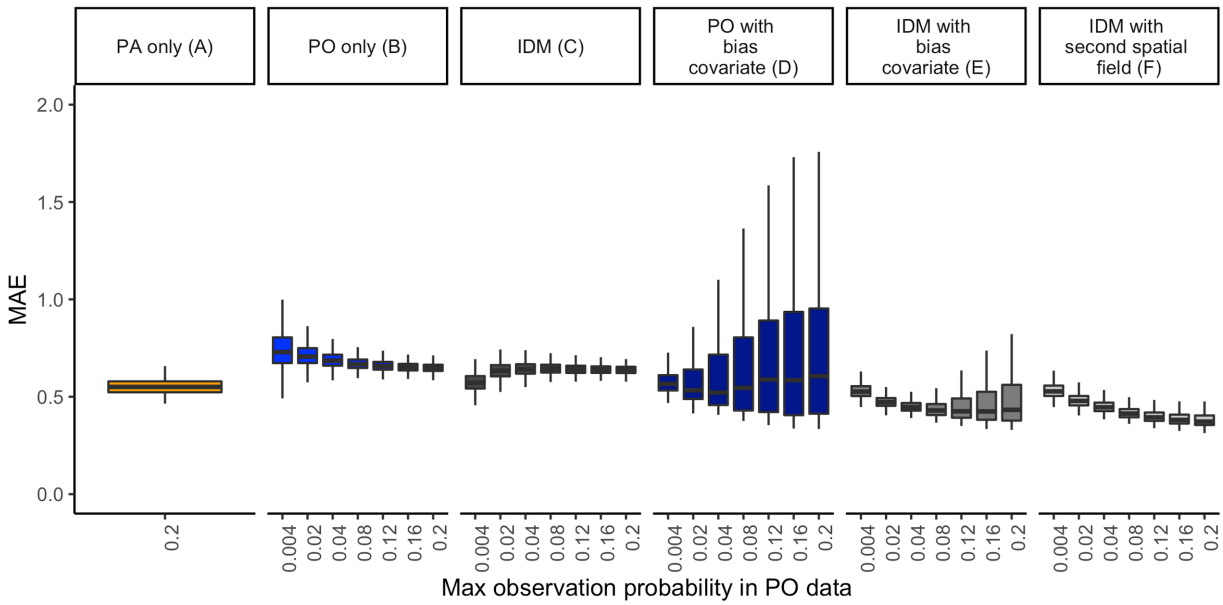

Figure 3. Boxplot of the MAE between predicted relative intensity and true relative intensity for all 500 simulations at each level of maximum observation probability. Each boxplot shows the interquartile range (25–75% quartiles), and the median of the simulated results. Whiskers of the boxplots extend to the largest or smallest point within 1.5 times the interquartile range from the edge of the boxplot. Outliers are not plotted.

without either $z(s)$ or $\zeta(s)$ performed slightly better on average than at higher observation probabilities. This is likely to reflect the PA data having a higher relative weighting in the joint likelihood when the amount of PO data is low, bringing estimates closer to the PA-only model.

Secondly, the variation in MAE tended to increase with higher observation probability for models including the covariate $z(s)$ to model spatial bias in PO data (models D and E). This was attributed to poorer estimation of the coefficient $\hat{\beta}_z z(s)$ with larger amounts of PO data (Supplementary material Appendix 1 Fig. A2), a counterintuitive result. Overall, IDMs again performed best in terms of accuracy of spatial prediction, but only when provided with information on bias. If bias was accounted for then increasing amounts of PO data provided improved model estimates compared to PA-only models. Similar to the first scenario, the maximum observation probability had little effect on estimation of the environmental covariate (Supplementary material Appendix 1 Fig. A3–A4), but the effect it did have was counterintuitive. As observation probability decreased the accuracy and precision of estimates from models B, C, D and E all increased. In contrast, model F showed little response to altering observation probability. For IDMs this pattern reflects the relative contribution of PA data, which is highest at the lowest observation probabilities when PA data makes up approximately two thirds of the data input (77 PO samples to 150 PA samples). For the PO only models, it could be driven by a stronger spatial pattern in bias when maximum probability of observation is higher, which could mask the environmental effect.

Correlated covariates

The final scenario investigated whether correlation between the spatial bias in PO data $p(s)$ and the environmental covariate $x(s)$ would affect model performance. All models except the PA-only model consistently underestimated the effect of the environmental covariate and a lower correlation when $p(s)$ and $x(s)$ were correlated (Fig. 4, Supplementary material Appendix 3). In this scenario, the IDM with a second spatial field had the highest correlation and lowest MAE, even though these models failed to correctly estimate $\beta_x x(s)$. So, IDMs with a second spatial field were able to retain robust spatial predictions, even if parameter estimation was poor.

Discussion

Our simulation study investigated whether IDMs always performed better than single models of PO and PA data under a range of scenarios. We found that IDMs outperformed single dataset models in some cases, but if bias in PO data was ignored then IDMs did not provide any benefits over modelling PA data alone.

Previous work has shown that bias in PO data can be accounted for by leveraging information in PA data (Fithian et al. 2015). These previous applications of IDMs have assumed there are covariates, or information, available to estimate sampling bias in the PO data. They did not consider cases where bias is unknown or poorly explained by the available covariates. Here, we demonstrate that if bias cannot

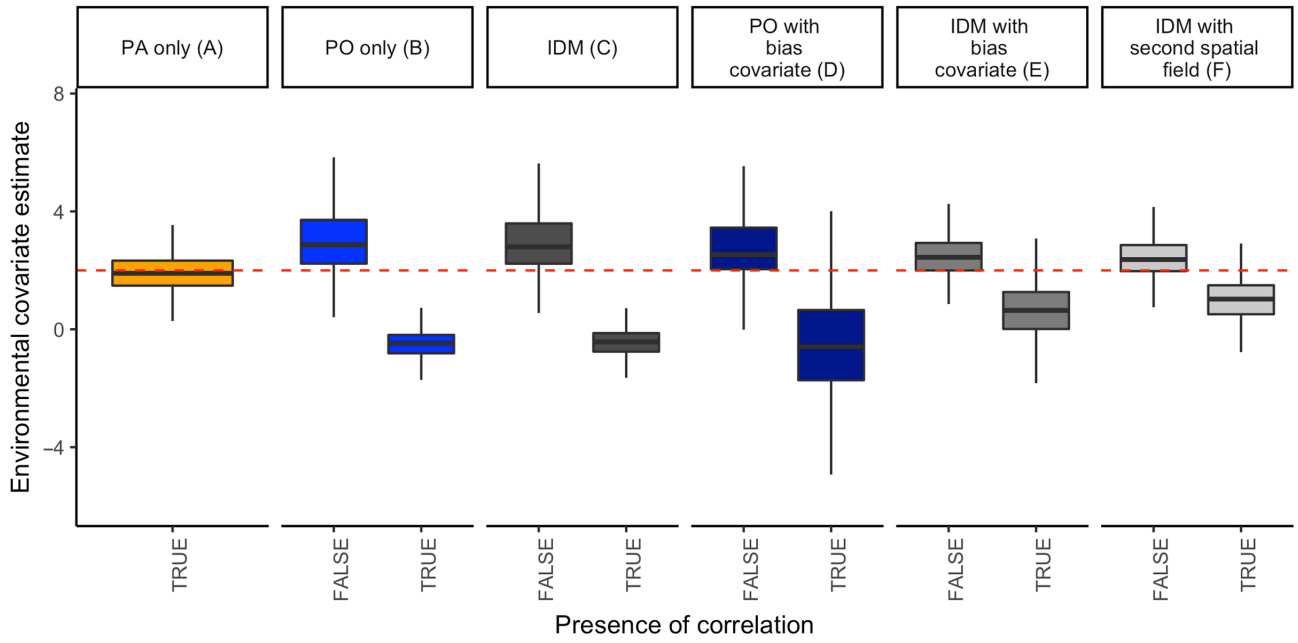

Figure 4. Boxplots of the mean estimate of the environmental covariate effect for all 500 simulations at each sample size of the structured data. Red horizontal line indicates simulated value of the environmental covariate effect. 95% of the results for each model type are plotted.

be accounted for then there is a risk that IDMs will provide estimates that are less accurate and more imprecise than analyzing PA data alone. However, the IDMs did produce an improvement over analyzing PO data alone. Analyzing PO data alone is equivalent to data pooling, when two datasets are combined by degrading the more detailed dataset (PA data) to match the less detailed one (PO data) (Fletcher et al. 2019, Isaac et al. 2019). One difference between our study and previous work is that we did not attempt to estimate $p(s)$, the thinning probability of PO data, explicitly.

Our simulations demonstrated that the IDM with a second spatial field performed well in all scenarios and was robust to correlations between bias and the environmental covariate (Supplementary material Appendix 3 Fig. A8–A10). Single PO models and the IDM without bias information or with a bias covariate all had imprecise and inaccurate estimates when bias in the PO data was correlated with the environmental covariate. The good performance of the model with two spatial fields may be due to two reasons. Firstly, there was evidence that in some cases the bias covariate was poorly estimated, particularly when observation probability was higher (Supplementary material Appendix 1 Fig. A2). This may have occurred because information related to bias was inappropriately included in the shared spatial field instead of attributed to the covariate. In particular, this may explain why poor estimation seemed to be more common with higher volumes of PO data. It would be useful to explore further whether variation due to spatial covariates could be attributed to the shared spatial field as this may also influence interpretation of environmental covariates. Secondly, high performance of the two spatial field model could reflect an overfitting of the highly flexible second spatial field. This might explain why the models with a second spatial field performed well in the correlated covariates scenario, even when the environmental covariate was incorrectly estimated. Default priors on the spatial fields were used in this simulation, but real-world applications would require careful selection of spatial priors (Illian et al. 2012) or use of penalized complexity priors (Fuglstad et al. 2019) to reduce overfitting.

Fitting a second field may be a useful approach when knowledge on potential sources of bias is limited or covariates are not available. In real world applications, understanding if bias is adequately captured by covariates is challenging. Therefore, the potential of using the second spatial field, instead of a known covariate, could be a mechanism to make use of the large amounts of unstructured data we have available, even without known bias information. Investigating the patterns in this field could even provide useful information on possible sources of bias (Neyens et al. 2019).

A key assumption we use here is that there is a spatial pattern in the bias associated with PO data which we can estimate by using the PA data to constrain the shared spatial pattern. In our simulation, the shape of bias in observation probability was necessarily simplistic and highly spatially structured (Fig. 1c), likely aiding the performance of the second spatial field. In reality, not all biases will have such strong spatial patterns. Although some sources of bias are spatially-patterned (e.g. human population density) others are not (e.g. time spent searching). Unknown non-spatial variation in effort or detection cannot be accounted for by a second spatial field and may lead to incorrect estimations of species distributions, particularly if the volume of PO data is large in relation to PA data. An alternative solution to integrate data with suspected biases might be to construct IDMs without using the joint likelihood framework. Adding the PO dataset via covariates, an informative prior or a correlation structure could be more robust to bias in opportunistic data whilst still allowing integration of different data types (Pacifici et al. 2017, Miller et al. 2019). Exploring the performance of the IDMs with a second spatial field against these alternative models for datasets where bias is not spatially structured, is an avenue that should be pursued.

The models in our study are far simpler than in most of the literature on IDMs. We restricted our simulations to mimic the minimal amount of data that might be available to ecologists seeking to combine datasets to estimate species distributions. Conversely, previous work has used multiple species (Fithian et al. 2015, Peel et al. 2019) or repeat surveys (Koshkina et al. 2017) to provide additional information to constrain estimates and reduce bias. In situations where only two datasets are available on a single species and no information can be provided to the model to determine which dataset to prioritize, the likelihood is swamped by the larger data source. As a result, in our simulations IDMs did not give a meaningful improvement over single models of PO data. Fletcher et al. (2019) proposed a weighted likelihood approach as a solution to swamping of the likelihood. However, this requires a choice as to which dataset should have the highest weight and exactly what that weight should be: research towards objective criteria for weighting datasets in IDMs is therefore a priority.

This study has demonstrated that integrated distribution models can outperform models of single datasets when spatial bias in PO data is explicitly included in the model. Second spatial fields seem like a potentially exciting tool that could quantify spatial bias in PO data, even when knowledge or the shape and drivers of this bias is unknown. We have included several simplifying assumptions in this study, such as perfect detection in PA data. Testing what happens if these assumptions are violated would give further insight into the utility of IDMs and deepen our understanding of the performance of these models.

Data availability statement

All of the code used in this study is openly available at: <https://github.com/NERC-CEH/IOFFsimwork>. This includes code to create all figures and tables.

Acknowledgements – We would like to thank Richard Chandler and Stephen Freeman for their contributions to the initial stages of planning this paper and Diana Bowler for useful comments on a previous version on this manuscript.

Funding – This work was supported by Natural Environment Research Council grant NE/R005133/1 and award NE/R016429/1 as part of the UK-SCAPE programme delivering National Capability.

Author contributions – EGS and SGJ (with equal contribution) came up with the design of the simulation study, conducted all analyses, wrote code, analysed outputs and wrote the manuscript. RBH and NJBI conceived of the initial idea for the manuscript, had input on the design and execution of the simulations, and assisted with writing of the manuscript. PAH developed one of the models used in this study, he also contributed to the methodological design and the writing of the manuscript.

References

Baddeley, A. et al. 2015. Spatial point patterns: methodology and application with R. – Chapman and Hall/CRC.

Bowler, D. E. et al. 2019. Integrating data from different survey types for population monitoring of an endangered species: the case of the Eld's deer. – Sci. Rep. 9: 7766.

Cressie, N. A. C. 1993. Statistics for spatial data, revised edition. – Wiley.

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

Fithian, W. et al. 2015. Bias correction in species distribution models: pooling survey and collection data for multiple species. – Methods Ecol. Evol. 6: 424–438.

Fletcher, R. J. et al. 2019. A practical guide for combining data to model species distributions. – Ecology 100: e02710.

Fuglstad, G. A. et al. 2019. Constructing priors that penalize the complexity of Gaussian random fields. – J. Am. Stat. Assoc. 114: 445–452.

Guillera-Arroita, G. et al. 2014. Ignoring imperfect detection in biological surveys is dangerous: a response to 'fitting and interpreting occupancy models'. – PLoS One 9: e99571.

Illian, J. B. 2017. Spatial and spatio-temporal point processes in ecological applications. – In: Gelfand, A. E. et al. (eds), Handbook of environmental and ecological statistics. CRC Press, Taylor & Francis Group, pp. 97–132.

Illian, J. B. et al. 2012. A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA). – Ann. Appl. Stat. 6: 1499–1530.

Isaac, N. J. B. et al. 2019. Data integration for large-scale models of species distributions. – Trends Ecol. Evol. 35: 56–67.

Kery, M. and Royle, J. 2016. Applied hierarchical modeling in ecology: analysis of distribution, abundance and species richness in R and BUGS. – Elsevier.

Koshkina, V. et al. 2017. Integrated species distribution models: combining presence-background data and site-occupany data with imperfect detection. – Methods Ecol. Evol. 8: 420–430.

Lindgren, F. et al. 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. – J. R. Stat. Soc. B 73: 423–498.

Miller, D. A. W. et al. 2019. The recent past and promising future for data integration methods to estimate species' distributions. – Methods Ecol. Evol. 10: 22–37.

Neyens, T. et al. 2019. Mapping species richness using opportunistic samples: a case study on ground-floor bryophyte species richness in the Belgian province of Limburg. – Sci. Rep. 9: 19122.

Pacifici, K. et al. 2017. Integrating multiple data sources in species distribution modeling: a framework for data fusion. – Ecology 98: 840–850.

Peel, S. L. et al. 2019. Reliable species distributions are obtainable with sparse, patchy and biased data by leveraging over species and data types. – Methods Ecol. Evol. 10: 1002–1014.

Renner, I. W. et al. 2015. Point process models for presence-only analysis. – Methods Ecol. Evol. 6: 366–379.

Rue, H. et al. 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. – J. R. Stat. Soc. B 71: 319–392.

Simpson, D. et al. 2016. Going off grid: computationally efficient inference for log-Gaussian Cox processes. – Biometrika 103: 49–70.

Zipkin, E. F. et al. 2019. Innovations in data integration for modeling populations. – Ecology 100: e02713.

Supplementary material (available online as Appendix ecog-05146 at <www.ecography.org/appendix/ecog-05146>). Appendix 1–4.