Running head: OCCUPANCY MODELS FOR HISTORICAL DATA

Utilizing occupancy-detection models with museum specimen data: promise and pitfalls

Vaughn Shirey¹, Rassim Khelifa², Leithen K. M'Gonigle³, Laura Melissa Guzman^{3,4}

1. Department of Biology
Georgetown University
37th and O Streets NW, Washington, DC, United States, 20057
2. Department of Zoology
University of British Columbia
6270 University Blvd, Vancouver, BC, Canada, V6T 1Z4
3. Department of Biological Sciences
Simon Fraser University
8888 University Drive, Burnaby, BC, Canada, V5A 1S6
4. Marine and Environmental Biology Section at the Department of Biological Sciences
University of Southern California
Allan Hancock Foundation Building, Los Angeles, CA 90089-0371

Abstract

Historical museum records provide potentially useful data for identifying drivers of change in species occupancy. However, because museum records are typically obtained via many collection methods, methodological developments are needed in order to enable robust inferences. Occupancy-detection models, a relatively new and powerful suite of methods, are a potentially promising avenue because they can account for changes in collection effort through space and time. Here we present a methodological road-map for using occupancy models to analyze historical museum records. We use simulated data-sets to identify how and when patterns in data and/or modelling decisions can bias inference. We focus primarily on the consequences of contrasting methodological approaches for dealing with species' ranges and inferring species' non-detections in both space and time. We find that not all data-sets are suitable for occupancy-detection analysis but, under the right conditions (namely, data-sets that span long durations and contain a high fraction of community-wide collections, or collection events that focus on communities of organisms), models can accurately estimate trends. Finally, we present a case-study on eastern North American odonates where we calculate long-term trends of occupancy by using our most robust workflow.

Keywords: global change, occupancy, museum specimens, hierarchical model

Introduction

Global change processes are contributing to the rapid restructuring of biodiversity across the planet (Araújo and Rahbek, 2006; Bellard et al., 2012). At the same time, data regarding species occurrence are becoming more widely available through the digitization of museum collections (Hedrick et al., 2020) and citizen science platforms (Dickinson et al., 2012). At the time of writing, two large databases of species occurrences, the Global Biodiversity Information Facility (GBIF) and Integrated Digitized Biocollections (iDigBio), contain over 1.5-billion records of species occurrences across the planet. These records may be rich in detailed information including date of collection, location, taxonomic determinations, and collector names. However, because museum data are aggregated records from many different collectors and time periods, they do not derive from a single sampling design and, further, the sampling design associated with each specimens is rarely noted. Consequently, they can contain numerous spatiotemporal and taxonomic biases (Isaac et al., 2014). For example, sampling of opportunistic (henceforth, "unstructured") data is often uneven in space and time and may be spatially concentrated around regions of high human population densities (Mair and Ruete, 2016; Tiago et al., 2017; Daru et al., 2018; Shirey et al., 2021). Additionally, charismatic taxa are often oversampled with respect to the diversity of their respective clades, leading to occurrence shortfalls, or fewer than expected occurrences given the diversity of a particular clade, in hyperdiverse groups such as arthropods (Troudet et al., 2017; Callaghan et al., 2021). In some groups, such as North American butterflies, undersampling is prolific in regions that are forecasted to experience the most dramatic changes in climate (Shirey et al., 2021). Given these biases, proper treatment of unstructured data can lead to misleading inferences (Guzman et al., 2021; Larsen and Shirey, 2021). Thus, it is imperative that we develop statistical

frameworks for analyzing unstructured data in order to afford researchers the ability to test hypotheses related to how global change is impacting species over long time-periods and large spatial extents.

Species observation is rarely perfect so, even when a given species is present at a site, the observer may fail to detect it during any given survey (e.g., animals with cryptic phenotype or behavior may be difficult to see). Occupancy-detection models are a powerful statistical framework for disentangling this imperfect processes of observation from actual species' occurrence or "occupancy" (MacKenzie et al., 2003) and these models have been widely used for species' distribution modeling in ecology (Kéry and Schaub, 2011). Occupancy models have been extended to model entire communities (so called, "multi-species occupancy models") and to model one or more species across multiple seasons (so called, "dynamic occupancy models") (Kéry and Schaub, 2011). These models have improved our ability to explore spatial and temporal patterns of biodiversity and aided in identification of drivers of global change (Kéry and Royle, 2020).

Occupancy models are potentially well-suited to handle the inherent biases in species observations present in unstructured data (Erickson and Smith, 2021). A major challenge when doing so is accounting for the fact that records of when and where historical collectors sampled are rarely available. Collections that did not yield specimens, or did but the specimens were never archived in a museum, are typically unreported, because no physical specimens exist in the museum collection. This shortcoming is further exacerbated by variable (and also unknown) individual collector behaviors. For example, some collectors focus on particular taxonomic groups while others collect entire communities (de Siracusa et al., 2020; Di Cecco et al., 2021). Species' traits may also influence the probability of detection or "allure" and, thus, collection outcomes

(Johnston et al., 2014). In all, physical specimens and their metadata represent a fragmented history of potential collection events that researchers must carefully piece together to reconstruct historical patterns.

Researchers have developed some best-practices for overcoming some of these challenges for presence-only data. For instance, it might be appropriate to infer non-detections of one (or more) species at a given site based on known detections of other species at that same site (Kéry, 2010; van Strien et al., 2013a; Kamp et al., 2016a). Constraining detection-non-detection records only to those sites that are plausibly within the range of a species' geographic distribution has also been shown to improve model performance (Guzman et al., 2021). Combined, these steps not only improve estimates of occupancy and detection, but can also reduce the computation time of analyses. However, we lack clear guidelines that highlight best practices for avoiding the potential biases introduced by variable observer behavior, temporal trends in visitation, and species distributions and traits.

Here, we use a simulation framework to explore the performance of occupancy-detection models when applied to data that contain many of the patterns likely present in natural history collections data. Specifically, we explore how changes in collector visit frequency, species' occupancy, and species' detection impact model performance for data-sets containing presence only records across a community of species, each with its own geographic range. We consider the consequences of 1) incorporating estimated species' ranges and 2) inferring non-detections of some species from recorded detections of others on estimated trends in occupancy and detection. We synthesize our findings by providing guidelines for steps that could be taken to reduce bias when applying similar occupancy-detection models to empirical data-sets. Finally, we apply our scenarios to a real-world data-set containing detection records for eastern

95 North American odonates (dragonflies and damselflies).

96 **Methods**

97 **Data simulation**

98 We simulate a community of N species, each inhabiting some or all of J potential sites
 99 over K time periods, which we call "eras." Eras could correspond to single years or could
 100 span multiple years. We suppose that each era is further broken into I time intervals,
 101 during which collectors may visit sites to sample some or all of the species present
 102 (commonly used terms in this manuscript are presented in the glossary of terms Box 1).

103 Because species often differ in their spatial extent, in some of our scenarios, we simulate
 104 a unique "geographic" range for each species. Specifically, species i potentially occupies
 105 sites $R_i = \{s_{i,1}, s_{i,2}, \dots, s_{i,r_i}\}$, where $r_i \in \{1, 2, \dots, J\}$. We refer to R_i as the "range" of
 106 species i . To simulate each species range, we generated polygons where size was
 107 determined using a Poisson distribution. Ranges were constructed by placing a series of
 108 random vertices on a defined grid and by drawing a concave polygon around these
 109 vertices. The size of the range was defined by a scaling factor, r_i which was drawn from
 110 a Poisson distribution as:

$$r_i \sim 1 - \text{Poisson}(0.1). \quad (1)$$

111 For site j within the range of species i , we simulate occupancy in era k by drawing from a
 112 Bernoulli distribution with success probability $\psi_{i,j,k}$, where

$$\psi_{i,j,k} = \text{expit}(\psi_0 + \psi_{\text{sp}}[i] + \psi_{\text{era}}[i] \times (k - 1)). \quad (2)$$

Here ψ_0 denotes the baseline occupancy (on the linear scale), $\psi_{\text{sp}}[i]$ is a random species-specific intercept, and $\psi_{\text{era}}[i]$ is a random species-specific effect of era (a positive value of $\psi_{\text{era}}[i]$ would indicate that species i is increasing in occupancy through time). We assume that $\psi_{\text{sp}}[i]$ and $\psi_{\text{era}}[i]$ are normally distributed, such that:

$$\begin{aligned}\psi_{\text{sp}}[i] &\sim \mathcal{N}(0, \sigma_{\psi, \text{sp}}) \\ \psi_{\text{era}}[i] &\sim \mathcal{N}(\mu_{\psi, \text{era}}, \sigma_{\psi, \text{era}}).\end{aligned}\tag{3}$$

Here, $\sigma_{\psi, \text{sp}}$ denotes the interspecific variability in occupancy, $\mu_{\psi, \text{era}}$ denotes the mean effect of era on species occupancy, and $\sigma_{\psi, \text{era}}$ denotes the variability in species' temporal occupancy trends.

Because sampling effort can change through time, we assume that the probability that a site is visited in a time interval can change across eras. Further, because some collectors may record all species that they detect, while others may record only a single species, all site visits are not equivalent. If a collector records all species (we call this "community sampling"), then a visit to a site in a given time interval is a relevant visit *for all species*. However, if a collector only records a single species, ignoring others (we call this "targeted sampling"), then a visit to a site in a given time interval is only a meaningful visit for the recorded species. In our data simulation, we suppose that, in era k , site j receives a "community sampling" visit in each time interval with probability

$$v_{j,k} = \rho_{\text{com}} \cdot \text{expit}[\eta_0 + \eta_{\text{era}} \times (k - 1)]\tag{4}$$

and when such a visit occurs, each species is recorded as either detected or not detected. Here, η_0 denotes the baseline probability of a community sampling visit (on the linear scale) and η_{era} is the effect of era on visitation probability. This latter term allows the

probability of site visitation to change systematically across eras. ρ_{com} lets us tune the fraction of community sampling visits relative to targeted sampling events; if $\rho_{\text{com}} = 1$, all visits are community sampling events, whereas, if $\rho_{\text{com}} = 0$, all sampling events are targeted sampling events. By parameterizing ρ_{com} as a function of model parameters, one could also allow collector behaviour to change across space or through time, however, that is beyond our scope here.

Only time intervals (at each site) that do not receive a community sampling visit can be subjected to targeted sampling visits (a targeted sampling visit will not add any additional information if a community sampling visit also occurs in the same time interval). Targeted sampling visits are, by their nature, species-specific. The probability that, in era k , site j receives a targeted sampling visit for species i is given by

$$v_{i,j,k} = (1 - \rho_{\text{com}}) \cdot \text{expit} [\nu_0[i] + \nu_{\text{era}}[i] \times (k - 1)] \quad (5)$$

where $\nu_0[i]$ denotes the baseline probability of a targeted sampling visit searching for species i (on the linear scale) and $\nu_{\text{era}}[i]$ denotes the effect of era on the probability of targeted sampling visits searching for species i . We assume that both $\nu_{\text{sp}}[i]$ and $\nu_{\text{era}}[i]$ are normally distributed, such that

$$\begin{aligned} \nu_0[i] &\sim \mathcal{N}(\mu_{\nu,0}, \sigma_{\nu,\text{sp}}) \\ \nu_{\text{era}}[i] &\sim \mathcal{N}(\mu_{\nu,\text{era}}, \sigma_{\nu,\text{era}}) \end{aligned} \quad (6)$$

where $\mu_{\nu,0}$ denotes the mean probability, across species, of a targeted sampling visit (in the first era), $\mu_{\nu,\text{era}}$ denotes the mean effect, also across species, of era on the probability of targeted sampling visits, and $\sigma_{\nu,\text{sp}}$ and $\sigma_{\nu,\text{era}}$ denote the interspecific variation in visitation probability and variation in temporal change in visitation, respectively. In our

simulation, we let η_0 equal $\mu_{v,0}$, and η_{era} equal $\mu_{v,\text{era}}$ so that both the probability of community and targeted sampling events increase or decrease simultaneously, but this does not need to be the case. For example, the probability of community sampling events could increase across eras, while the probability of targeted sampling events can go down. Exploring that scenario could be interesting but beyond the scope of this manuscript. Because η_{era} is equal to $\mu_{v,\text{era}}$, variations of $\mu_{v,\text{era}}$ (and therefore η_{era}) will simply be labelled as $\mu_{v,\text{era}}$.

Finally, for each species, we simulate detection during the time intervals where visits occurred. For site $j \in R_i$, we draw species' i detections across the visits that occurred and could have detected that species (e.g., community sampling visits or targetting sampling visits aimed at species i) in era k from a Bernoulli distribution with probability $p_{i,j,k}$ where

$$p_{i,j,k} = \text{expit} [p_0 + p_{\text{sp}}[i] + p_{\text{site}}[j, k] + p_{\text{era}} \times (k - 1)] , \quad (7)$$

where p_0 is the baseline detection probability (on the linear scale), $p_{\text{sp}}[i]$ is a random species-specific intercept, $p_{\text{site}}[j, k]$ is a random site-specific intercept that varies by era, and p_{era} is an overall effect of era. $p_{\text{site}}[j, k]$ allows spatiotemporal variability in detection probability (changing across sites and eras), and helps account for the variation that is inherent in sample effort across space and time in opportunistic historical data-sets. p_{era} allows detection to change systematically through time, as has likely occurred in many groups as sampling techniques have improved, for example. We assume that both p_{sp} and p_{site} are normally distributed, such that:

$$\begin{aligned} p_{\text{sp}}[i] &\sim \mathcal{N}(0, \sigma_{p,\text{sp}}) \\ p_{\text{site}}[j, k] &\sim \mathcal{N}(0, \sigma_{p,\text{site}}), \end{aligned} \quad (8)$$

where $\sigma_{p,\text{sp}}$ denotes the interspecific variation in detection and $\sigma_{p,\text{site}}$ the spatiotemporal

(site and era) variation in detection.

We considered (i) three different numbers of eras (K), (ii) five different fractions of community sampling events (ρ_{com}), (iii) three different trends in visitation probability ($\mu_{v,\text{era}}$; increasing, decreasing, and stable), and (iv) identical versus variable species ranges for a total of 135 simulation scenarios (specific parameter values are shown in Table S1 and Fig. 1). We conducted all data simulations in R (code is available at https://github.com/lmguzman/occ_historical).

Occupancy-detection models

We built an occupancy model that reflect the structure of the simulated data-sets, except that we do not incorporate site visitation history or community sampling information, as this information is typically not easily accessible (or accessible at all) for empirical data-sets. We provide a detailed descriptions of our occupancy models in the Supplementary Information.

Using our simulated data-sets, we contrast the performance of six different decision workflows one might follow when applying an occupancy model to a historical data-set. We provide an overview of these below and they are also shown graphically in Fig. 2.

- $WF_{\text{all,all}}$: we model all sites for all species across all time intervals, even when some of those time intervals did not contain any known species detections.
- $WF_{\text{all,detected}}$: we model all sites for all species, but each site is only modelled over the time intervals where at least 1 species was detected. Here, we infer non-detections of non-observed species when at least one other species was detected at a site in a given visit. This is a fair assumption when most collectors

sample the entire community ($\rho_{\text{com}} = 1$), but not when they mostly target individual species ($\rho_{\text{com}} = 0$).

- $WF_{\text{all,visits}}$: we model all sites for all species, but each site only over the time intervals where we know visits occurred (i.e., we model the true site visitation history which we are only able to do because we are working with simulated data). This case provides a benchmark for a 'best information' workflow.
- $WF_{\text{range,all}}$: we model each species over only the sites that fall within its range and each site is modelled across all time intervals (this is an analog to $WF_{\text{all,all}}$ above, but where species' ranges are estimated and then incorporated).
- $WF_{\text{range,detected}}$: we model each species over only the sites that fall within its inferred range and, for each of those sites, only over time intervals where at least 1 species was detected (this is an analog to $WF_{\text{all,detected}}$ above, but where species' ranges are estimated and then incorporated).
- $WF_{\text{range,visits}}$: we model each species over only the sites that fall within its inferred range, but each site only over the time intervals where we know visits occurred (this is an analog to $WF_{\text{all,visits}}$ above, but where species' ranges are estimated and then incorporated).

We compared these different processing workflows by assessing accuracy of model-estimated parameter values against their true values using root mean squared error (RMSE) across five replicates.

Case Study: Eastern North American Odonates

Study organism: Odonates are globally widespread aquatic insects that live in a wide range of freshwater systems. They are charismatic, easy to detect and identify, and popular among nature enthusiasts. Both larvae and adults are generalist predators and interact with a large number of aquatic and terrestrial taxa. In North America, there are 462 species (136 Zygoptera and 326 Anisoptera), with 336 species on the east coast (Paulson, 2011). Odonates are sensitive to environmental change, particularly climate warming, and evidence suggests that they respond by shifting their phenology and geographic distribution (Hickling et al., 2005; Hassall et al., 2007), making them interesting sentinels of climate change.

Odonate data-set: We selected 27 states in Eastern North America from GBIF (<https://doi.org/10.15468/dl.cabqrc>). The data comprise 152,863 records of 368 species observed during 1970 – 2020 and show a clear increase in the number of records through time, with 70.3% of the records occurring in the last two decades (2000 – 2020). We partitioned records into “sites” by overlaying a grid of $2^{\circ} \times 2^{\circ}$ cells and filtered to include only species with more than 100 occurrences (194 species) across 90 sites and five eras (decades), each further partitioned into five two year time intervals.

Analysis of the odonate data-set: We provide guidelines for researchers when analysing historical museum records with occupancy models (Box 2) and we demonstrate application of these steps using the above-described Odonate data-set. Specifically, we use a multispecies occupancy model, as described in the supplementary materials. For each species, we construct the geographic range by drawing a convex polygon around all occurrences records. We process the data under the $WF_{\text{range,detected}}$ (best case scenario on simulated data, excluding $WF_{\text{range,visits}}$ which is not possible for empirical data),

$WF_{\text{all,detected}}$ which ignores underlying species ranges, and $WF_{\text{range,all}}$ which models all visits, regardless of whether or not visits likely occurred. We also assessed how many of the records in our data-set likely originated via community sampling vs. targeted sampling visits. To do this, we grouped records collected on the same day and within 1km of one other. We assumed that a community sampling event occurred if more than one species was present in each grouping. While this is likely a rough approximation of the actual fraction of community sampling events, it lets us quickly approximate this quantity without having to use more time-consuming methods like a detailed analysis of collector identities, etc.

We performed all analyses with R4.1.1 (R Core Team, 2019). We used the nimble package to compute our models (de Valpine et al., 2017). All of the code utilized in our study is available at https://github.com/lmguzman/occ_historical.

Results

Simulation Study

The choice of data processing workflow directly affects inferences made by occupancy models. Not surprisingly, workflows that only include site $\times$ time interval combinations where visits actually occurred ($WF_{\text{all,visits}}$ and $WF_{\text{range,visits}}$) outperformed models that inferred site visitation and this was invariant to the fraction of visits that represent community sampling events (Fig. 3). The performance of workflows that infer species' non-detections by either modelling all time intervals ($WF_{\text{all,detected}}$) or only time intervals where at least one species was detected ($WF_{\text{all,all}}$) depends on the fraction of visits that are community sampling events. Specifically, when all site visits correspond to

community sampling events ($\rho_{\text{com}} = 1$), inferring species non-detections based on positive detection of other species is a sensible action and, consequently, $WF_{\text{all,detected}}$ performs as well as $WF_{\text{all,visits}}$ (Fig. 3c,d,e,f). In contrast, when all site visits are targeted sampling events, ($\rho_{\text{com}} = 0$), detections of one species provide little information about non-detections of other species and, consequently, both $WF_{\text{all,detected}}$ and $WF_{\text{all,all}}$ yield biased estimates of occupancy change through time (purple and blue curves are much higher than green curve for small values of ρ_{com} in Fig. 3).

We found that model error increased as the minimum fraction of visits that correspond to community sampling events decreased and, further, that the extent of this error depended on (i) the number of eras, (ii) whether species' ranges were incorporated into the analysis, (iii) the extent to which site visitation probability changed through time, and (iv) the extent to which occupancy and/or detection probability changed through time. We will discuss these processes next.

In general, estimated temporal changes in occupancy and/or detection probability through time are more accurate when data comprise a greater number of eras (larger K) and this conclusion is true across data processing workflows (Fig. 3). Model estimates for data-sets spanning ten eras were able to better estimate changes in occupancy and detection through time when community visits comprise a smaller fraction of the total site visits (lower ρ_{com}) than estimates for data-sets spanning five eras (Fig. 3c,d,e,f). Estimates for changes in detection through time were, however, less sensitive to both the fraction of visits that comprise community sampling events and to the workflow (Fig. 3, Fig. 4).

We simulated data-sets either by assuming that every species could potentially occupy every site or by assuming species were constrained to different (possibly overlapping)

geographic ranges. We subsequently differentiated our data processing workflows into those that ignore potential differences in species ranges ($WF_{all,all}$, $WF_{all,detected}$, and $WF_{all,visits}$, which we collectively label as $WF_{all,X}$) and those where each species is modelled over only the sites in its range ($WF_{range,all}$, $WF_{range,detected}$, and $WF_{range,visits}$, which we collectively label as $WF_{range,X}$). When all species could plausibly occupy all sites, $WF_{all,X}$ and $WF_{range,X}$ become equivalent. We found that workflows that consider underlying species' ranges (either $WF_{all,X}$ when all the species were simulated to be plausibly occupy all sites or $WF_{range,X}$ when species ranges were simulated) generally had lower error rates in changes of occupancy through time than the case where the data processing workflow mismatched the underlying species ranges (using $WF_{all,X}$ when species ranges were simulated) (Fig. 4a,c).

When the probability of site visitation decreased through time, we found that workflows that attempted to infer and model species' detections based on recorded presences of other species (e.g., $WF_{range,detected}$) yielded better occupancy and detection estimates than workflows that simply modeled all species across time intervals (e.g., $WF_{range,all}$). This pattern is most obvious when most visits correspond to community sampling events (high ρ_{com} , Fig. 5, top row). This effect disappeared when site visitation probability remained constant or increased through time (Fig. 5 bottom rows). Generally, an increasing probability of site visitation through time improves model inferences compared to decreasing probability of site visitation when a lower fraction of the visits correspond to community sampling events. For instance, models yield accurate estimates of occupancy and detection when only 25% of site visits are community sampling events, as long as the site visitation probability increased through time, the number of eras was high ($K = 10$), the data processing workflow took into account the underlying species ranges, and species' non-detections were only inferred

when other other species were detected (i.e., $WF_{\text{range,detected}}$).

Because real-world changes in quantities like site visitation rate, occupancy, detection, and collector behavior are potentially changing through time, we ran a full suite of occupancy models that included 3,750 workflow-simulation combinations with five replicates per combination. Root-mean squared errors (RMSEs) from this full set of workflow-simulation configurations revealed that overestimation of occupancy declines are likely to occur when the fraction of visits that are community samples is low ($0 - 0.25$) and this is especially true when detection and visit probability decrease through time (Supplemental Figures S1, S2, and S3).

Case Study: Eastern North American Odonates

To examine occupancy trends for eastern North American odonates from the 1970s through 2010s, we ran occupancy models on data prepared using each of the processing workflows $WF_{\text{range,detected}}$, $WF_{\text{all,detected}}$, and $WF_{\text{range,all}}$. We found that the proportion of site $\times$ era $\times$ time interval combinations that were likely community samples in this data was $\sim 39.7\%$. Because our conclusions from simulated data suggest that this might be too low for models to generate non-biased estimates, we re-filtered our data, applying stricter criteria. Specifically, we restricted our records to only those from site $\times$ era $\times$ time interval combinations where at least 50% of sampling events were inferred to be community sampling events (i.e., multiple species detected on the same day within 1km $\times$ 1km from one other). This led to an estimated global proportion of community samples equal to 76.6%. We then re-ran an occupancy model using our best performing workflow ($WF_{\text{range,detected}}$) on this restricted data-set.

In general, we found that mean occupancy increased over the past 4 decades. However,

species trends were highly variable, with many increasing while others decreased. By using a range-restricted workflow, our odonate results became less extreme in terms of overall, species-specific occupancy change from the 1970s to 2010s (Fig. 6). For example, using an $WF_{\text{all,detected}}$ workflow yields 65 species that are either increasing or decreasing by at least 25-percent. Restricting our workflow to $WF_{\text{range,detected}}$ reduces this figure to 22 species (24 species for $WF_{\text{range,detected}}$ with site/era/visit combinations that were clustered, community visits - 6d).

Discussion

We assessed the performance of different data-processing workflows when applying multi-species occupancy models to simulated data-sets in order to identify a set of best practices guidelines one might follow when analyzing historic museum records or opportunistic data. Our work here confirms the importance of restricting species-specific analyses across large geographic scales to those species' ranges. Further, inferring non-detections of species from other species' detections emerged as a better approach than simply modelling all possible time intervals at all sites. Our analysis also determines the extent to which patterns in data sampling, such as changing collector behaviour, can undermine our ability to infer trends using occupancy models from historic records. In addition, workflows that bin occurrence data into fewer eras are generally less accurate.

Inferring non-detections of some species based on recorded observations of other species is becoming a more frequent practice when working with presence-only occurrence data, but specific approaches differ. For example, van Strien et al. (2013b); Kamp et al. (2016b); Powney et al. (2019) inferred non-detections for a particular species if a different species

was observed at that same site on the same date. Outhwaite et al. (2019) used the same approach, but only inferred non-detections for species from the same taxonomic group as the observed individual(s). Analyses using these approaches have mostly taken place in Europe and with taxonomic groups where community science approaches yield structured data sets, for example those from regional monitoring schemes (van Swaay et al., 2008). For example, Johnston et al. (2021) found that well curated checklists combined with incomplete checklists were sufficient to enable reasonable estimates of occupancy, but not encounter rates. Here we have quantified the extent to which a data-set can comprise incomplete checklists or simply individual observations while still producing good parameter estimates. We found that, for the cases considered here, our workflows produced reliable estimates of occupancy trends through time when a least 50% of the samples derived from community sampling events.

Obtaining an approximate estimate of how frequently collectors sampled the entire community, versus only collecting or recording individual species, is a necessary step in assessing whether the methods we use here are likely to be appropriate. Unfortunately, direct measurement of this quantity is often impossible for large historic data-sets and, thus, we suggest that researchers approximate this metric to the best of their ability. In our Odonate analysis, we grouped records in space and time and assumed that specimens collected on the same day and nearby to one another were part of the same collection event. Based on our simulated data-sets, we concluded that reliable estimates for data-sets that span only two time periods require that greater than 50% of site visits must comprise community sampling events. In workflows that span a greater number of eras, this fraction could conceivably be as low as 25%, however, this likely depends on other data-specific trends and parameter estimates, and we would suggest that authors conducting such analyses always use simulation to guide their analyses. Our

examination of interactions between changes in occupancy and detection probability over time as well as changes in visitation probability revealed that when detection probability or when visit probability decrease over time, overestimation of occupancy declines will likely occur in data-sets where the probability of community visits is less than 0.5 (Supplemental Figures S1, S2, and S3).

In applying our different workflows to a large historic data-set for eastern North American odonates, we demonstrate that neglecting to incorporate estimates of species' ranges demonstrates how different workflow choices can impact downstream results with a real-world data-set. In particular, the results from this analysis show how data filtering can reduce the estimation of dramatic occupancy increases and declines (Fig. 6a vs. Fig. 6b,c,d). Overestimation of occupancy change is a known behavior of occupancy-detection models when heterogeneous observer effort is not accounted for (Kéry and Royle, 2020). Our findings generalize this conclusion to show that similar overestimation is possible when species are modeled outside of their ranges (using the Odonate example), when species are modelled in all time intervals, or when there is too few community sampling events (using simulated data).

Using the best workflow to analyze the Odonate dataset, we found that the majority of species had range expansions and a low number of species had range contractions in the eastern United States. This is mostly concordant with the assessment of the conservation status and population trends of the IUCN Red List at the global scale which lists all studied species as Least Concern with stable or increasing population trends. Odonates have a high ability to colonize new suitable habitats following the warming of cool areas (Hickling et al., 2005) or the creation of artificial sites (e.g. ponds for irrigation) (Simaika et al., 2016). The observed low range contractions might be due to their ability to rapidly recolonize areas where they were temporarily extirpated (Shiffer and White, 2014),

which is often observed at the local scale in the studied area (Shiffer et al., 2014, 2015).

Studies using occupancy models to assess the temporal pattern of the geographic distribution of odonates at a large scale are rare (Rocha-Ortega et al., 2020), however they may be possible under the occupancy-detection framework.

While we only use “single-season” occupancy models in our analyses here, we also explored (but do not present) dynamic occupancy models which link each time periods (or eras, in our terminology) via extinction-colonization dynamics (Royle and Kéry, 2007). In general, we found that these models performed poorly unless the probability of community visitation was high. Missing visits/eras in the dynamic framework created multiple possibilities for parameter estimation, where a single era with missing data could represent extinction followed by colonization, or persistence; multiple consecutive eras with missing visits simply compounds the number of possibilities. We expect that many historical data-sets likely do not have enough data to make dynamic models feasible. However, for many questions, single-season models parameterized with appropriate fixed and random temporal effects should suffice.

Processes which generate unstructured occurrence data are complex and it is likely that the fraction of community sampling events (vs. targeted sampling) varies across eras and through time. This may be especially relevant if researchers are interested in utilizing combined museum specimen and community science records for inference. Individual collector behaviors may be important in models using these combined data-sets given that collector behavior has a potentially large effect on community science outcomes (Di Cecco et al., 2021). Incorporating these processes is an increasingly important task, as community science records far outpace the digitization and collection of specimens via other methods in the last decade (Spear et al., 2017; Shirey et al., 2021).

While the current paper presents context to certain technical aspects of occupancy models that may have important consequences, there are many further complications that require additional attention. For example, restricting species' range in multispecies occupancy modeling is a good general practice when estimating occupancy and detection. However, if researchers are interested in shifts in species ranges (e.g., range expansion), the set of modeled sites for a given species needs to be sufficiently large to allow for such range expansions. Adding a buffer around a species' inferred range is a possible solution, but such decisions should be explored via simulation in more detail.

Conclusion

Our analyses of simulated data demonstrate that for some unstructured data, robust estimates of temporal trends in occupancy and/or detection can be obtained, particularly when (a) community sampling events comprise at least half of all sampling events and (b) when analyses span larger temporal eras. We strongly encourage researchers to "get to know" their data by assessing the ways in which it has been historically collected and curated and to perform, at a minimum, the guidelines we propose in this paper. Estimating the fraction of collection events that were community sampling events is critical for assessing how reliable occupancy-detection estimates may be. Overall, understanding when occupancy models are likely to fail and lead to incorrect inferences is a powerful tool when applying these models to historical museum records.

Acknowledgements

We acknowledge funding from Simon Fraser University (to LKM and LMG), the Natural Sciences and Engineering Research Council of Canada (NSERC) (Discovery Grant to LKM), Liber Ero Fellowship Program (LMG), Georgetown University (VS), National Science Foundation Graduate Research Fellowship (VS, grant number 1937959), and the Swiss National Science Foundation fellowship (RK, grant number P2ZHP2-175028). This research was enabled in part by support from West Grid (www.westgrid.ca) and Compute Canada (www.computecanada.ca).

Data Accessibility

The code used to generate simulated data is available via GitHub at https://github.com/lmguzman/occ_historical. A snapshot of the final code used at the time of publication is located in a Zenodo repository at xx.

References

- Araújo, M. B. and C. Rahbek, 2006. How does climate change affect biodiversity? *Science* 313:1396–1397.
- Bellard, C., C. Bertelsmeier, P. Leadley, W. Thuiller, and F. Courchamp, 2012. Impacts of climate change on the future of biodiversity. *Ecology letters* 15:365–377.
- Callaghan, C. T., A. G. Poore, M. Hofmann, C. J. Roberts, and H. M. Pereira, 2021.

Large-bodied birds are over-represented in unstructured citizen science data.

Scientific reports 11:1–11.

Daru, B. H., D. S. Park, R. B. Primack, C. G. Willis, D. S. Barrington, T. J. Whitfeld, T. G.

Seidler, P. W. Sweeney, D. R. Foster, A. M. Ellison, et al., 2018. Widespread sampling

biases in herbaria revealed from large-scale digitization. *New Phytologist* 217:939–955.

de Valpine, P., D. Turek, C. Paciorek, C. Anderson-Bergman, D. Temple Lang, and

R. Bodik, 2017. Programming with models: writing statistical algorithms for general

model structures with NIMBLE. *Journal of Computational and Graphical Statistics*

26:403–413.

Di Cecco, G., V. Barve, M. Belitz, B. Stucky, R. Guralnick, and A. Hurlbert, 2021.

Observing the obserers: how participants contribute data to inaturalist and

implications for biodiversity science. *BioScience* 71:1179–1188.

Dickinson, J. L., J. Shirk, D. Bonter, R. Bonney, R. L. Crain, J. Martin, T. Phillips, and

K. Purcell, 2012. The current state of citizen science as a tool for ecological research

and public engagement. *Frontiers in Ecology and the Environment* 10:291–297.

Erickson, K. D. and A. B. Smith, 2021. Accounting for imperfect detection in data from

museums and herbaria when modeling species distributions: combining and

contrasting data-level versus model-level bias correction. *Ecography* 44:1341–1352.

Guzman, L. M., S. A. Johnson, A. O. Mooers, and L. K. M’Gonigle, 2021. Using historical

data to estimate bumble bee occurrence: Variable trends across species provide little

support for community-level declines. *Biological Conservation* 257:109141.

Hassall, C., D. J. Thompson, G. C. French, and I. F. Harvey, 2007. Historical changes in

the phenology of british odonata are related to climate. *Global Change Biology*

13:933–941.

- Hedrick, B. P., J. M. Heberling, E. K. Meineke, K. G. Turner, C. J. Grassa, D. S. Park, J. Kennedy, J. A. Clarke, J. A. Cook, D. C. Blackburn, et al., 2020. Digitization and the future of natural history collections. *BioScience* 70:243–251.
- Hickling, R., D. B. Roy, J. K. Hill, and C. D. Thomas, 2005. A northward shift of range margins in british odonata. *Global change biology* 11:502–506.
- Isaac, N. J., A. J. van Strien, T. A. August, M. P. de Zeeuw, and D. B. Roy, 2014. Statistics for citizen science: extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution* 5:1052–1060.
- Johnston, A., W. M. Hochachka, M. E. Strimas-Mackey, V. Ruiz Gutierrez, O. J. Robinson, E. T. Miller, T. Auer, S. T. Kelling, and D. Fink, 2021. Analytical guidelines to increase the value of community science data: An example using ebird data to estimate species distributions. *Diversity and Distributions* .
- Johnston, A., S. E. Newson, K. Risely, A. J. Musgrove, D. Massimino, S. R. Baillie, and J. W. Pearce-Higgins, 2014. Species traits explain variation in detectability of uk birds. *Bird Study* 61:340–350.
- Kamp, J., S. Oppel, H. Heldbjerg, T. Nyegaard, and P. F. Donald, 2016a. Unstructured citizen science data fail to detect long-term population declines of common birds in denmark. *Diversity and Distributions* 22:1024–1035.
- , 2016b. Unstructured citizen science data fail to detect long-term population declines of common birds in denmark. *Diversity and Distributions* 22:1024–1035.
- Kéry, M., 2010. Introduction to WinBUGS for Ecologists: A Bayesian Approach to Regression, ANOVA and Related Analyses. Academic Press, Elsevier.

514 Kéry, M. and J. A. Royle, 2020. Applied Hierarchical Modeling in Ecology: Analysis of
515 distribution, abundance and species richness in R and BUGS: Volume 2: Dynamic and
516 Advanced Models. Academic Press.

517 Kéry, M. and M. Schaub, 2011. Bayesian population analysis using WinBUGS: a
518 hierarchical perspective. Academic Press.

519 Larsen, E. A. and V. Shirey, 2021. Method matters: pitfalls in analysing phenology from
520 occurrence records. Ecol. Lett. P. Early View.

521 MacKenzie, D. I., J. D. Nichols, J. E. Hines, M. G. Knutson, and A. B. Franklin, 2003.
522 Estimating site occupancy, colonization, and local extinction when a species is
523 detected imperfectly. Ecology 84:2200–2207.

524 Mair, L. and A. Ruete, 2016. Explaining spatial variation in the recording effort of citizen
525 science data across multiple taxa. PloS one 11:e0147796.

526 Outhwaite, C. L., G. D. Powney, T. A. August, R. E. Chandler, S. Rorke, O. L. Pescott,
527 M. Harvey, H. E. Roy, R. Fox, D. B. Roy, et al., 2019. Annual estimates of occupancy for
528 bryophytes, lichens and invertebrates in the uk, 1970–2015. Scientific data 6:1–12.

529 Paulson, D., 2011. Dragonflies and Damselflies of the East. Princeton University Press.

530 Powney, G. D., C. Carvell, M. Edwards, R. K. Morris, H. E. Roy, B. A. Woodcock, and
531 N. J. Isaac, 2019. Widespread losses of pollinating insects in Britain. Nature
532 communications 10:1–6.

533 R Core Team, 2019. R: A Language and Environment for Statistical Computing. R
534 Foundation for Statistical Computing, Vienna, Austria. URL
535 <http://www.R-project.org/>.

- 536 Rocha-Ortega, M., P. Rodríguez, J. Bried, J. Abbott, and A. Córdoba-Aguilar, 2020. Why
537 do bugs perish? range size and local vulnerability traits as surrogates of odonata
538 extinction risk. *Proceedings of the Royal Society B* 287:20192645.
- 539 Royle, J. A. and M. Kéry, 2007. A bayesian state-space formulation of dynamic
540 occupancy models. *Ecology* 88:1813–1823.
- 541 Shiffer, C. N., B. Leppo, and H. White, 2014. Odonata of black moshannon state park,
542 centre county, pennsylvania. *Argia* 26:8–15.
- 543 Shiffer, C. N., B. Leppo, and H. B. White, 2015. Odonata of beaver dam, huntingdon
544 county, pennsylvania: A record of faunal succession in a changing habitat. *Argia*
545 27:12–23.
- 546 Shiffer, C. N. and H. B. White, 2014. Dragonfly and damselfly colonization and
547 recolonization of a large, semi-permanent pennsylvania pond. *Northeastern*
548 *Naturalist* 21:630–651.
- 549 Shirey, V., M. W. Belitz, V. Barve, and R. Guralnick, 2021. A complete inventory of north
550 american butterfly occurrence data: narrowing data gaps, but increasing bias.
551 *Ecography* 44:537–547.
- 552 Simaika, J. P., M. J. Samways, and P. P. Frenzel, 2016. Artificial ponds increase local
553 dragonfly diversity in a global biodiversity hotspot. *Biodiversity and Conservation*
554 25:1921–1935.
- 555 de Siracusa, P. C., L. M. Gadelha, and A. Ziviani, 2020. New perspectives on analysing
556 data from biological collections based on social network analytics. *Scientific reports*
557 10:1–10.

- Spear, D., G. Pauly, and K. Kaiser, 2017. Citizen science as a tool for augmenting museum collection data from urban areas. *Frontiers in Ecology and Evolution* 5:86.
- van Strien, A. J., C. A. van Swaay, and T. Termaat, 2013a. Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. *Journal of Applied Ecology* 50:1450–1458.
- , 2013b. Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. *Journal of Applied Ecology* 50:1450–1458.
- van Swaay, C., P. Nowicki, J. Settele, and A. van Strien, 2008. Butterfly monitoring in europe: methods, applications and perspectives. *Biodiversity and Conservation* 17:3455–3469.
- Tiago, P., A. Ceia-Hasse, T. A. Marques, C. Capinha, and H. M. Pereira, 2017. Spatial distribution of citizen science casuistic observations for different taxonomic groups. *Scientific reports* 7:1–9.
- Troudet, J., P. Grandcolas, A. Blin, R. Vignes-Lebbe, and F. Legendre, 2017. Taxonomic bias in biodiversity data and societal preferences. *Scientific Reports* 7:9132.

Figure 1: Schematic depicting data simulation decisions. Visit, occupancy, and detection probabilities all have three levels: decreasing, stable, and increasing with time (-0.1, 0, 0.1). The probability that visits are community sampling events is: 0, 0.25, 0.50, 0.75, 1. All five parameters were varied sequentially, which resulted in 270 multi-species detection-non-detection data-sets ($N=2*3*3*3*5$). We also simulated data with two, five, and 10 eras.

Figure 2: Schematic depicting data processing workflows. Workflows differ in which sites and time intervals they model. Small black and white panels show, for species i , which site $\times$ time interval combinations from the larger panels are modelled in each workflow, with black denoting a modelled site $\times$ time interval combination and white a non-modelled combination. For example, $WF_{range,all}$ models all time intervals for every site within the range of species i , but none of the sites outside of it. Colours in large boxes behind small panels correspond to colours in subsequent figures.

Figure 3: When we only model time intervals where visits actually occurred (WF_{all,visits}, green line), estimates of change in occupancy ($\mu_{\psi,era}$) and detection (p_{era}) are generally more accurate (lower root mean square error RMSE) compared to workflows where we either model only visits with at least one species detected (WF_{all,detected}, blue line) or where we model all time intervals (WF_{all,all}, green line). This difference is largest when only the fraction of the visits that were community sampling events is low (small ρ_{com}). Across the board, models that span a greater number of eras generally yield more precise estimates. $\mu_{v,era} = 0$.

Figure 4: Changes in occupancy ($\mu_{\psi,era}$) and detection (p_{era}) through time have less error (root mean squared error RMSE) as the probability of community visits ρ_{com} increases. The coloured lines are the workflows regarding to the visitation history where X just stands in for any range workflow. The rows show the effect of the workflows with regards to the range and the underlying simulation. Sim_{all} is when all of the species were simulated to plausibly occupy all sites, while Sim_{range} is when species were simulated to have ranges. $WF_{all,X}$ is the data processing workflow where all sites are included in the occupancy model and $WF_{range,X}$ is the data processing workflow where the only sites that are included in the model are the ones that fall within a species range. $\mu_{v,era}$ is zero, Number of eras is five.

Figure 5: Changes in occupancy ($\mu_{\psi,era}$) and detection (p_{era}) through time have less error (root mean squared error RMSE) using the workflow $WF_{range,visits}$ than $WF_{range,detected}$ and $WF_{range,all}$. This error depends on the probability of community visits ρ_{com} and the change in visit history through time ($\mu_{v,era}$). Number of eras is ten.

Figure 6: Species-specific (colored lines) and community (black line) trends for eastern North American odonates under three data processing workflows: (a) $WF_{all,detected}$, (b) $WF_{range,all}$, (c-d) $WF_{range,detected}$. In (d), we further restrict analyses to only site $\times$ era $\times$ time interval combinations where the proportion of clustered collection events containing 2 or more species exceeds 0.5. Lines are coloured to indicate occupancy trend through time (difference from Era 1 to Era 5). Inset histograms show these occupancy differences. Black dashed lines denote 95% Bayesian Credible Intervals.

Glossary of terms
<p>Era: Period of time in which occupancy is estimated.</p> <p>Time interval: Units of time in which eras are further divided and in which potential site visits might occur.</p> <p>Visit: A species' observation at a site $\times$ time interval combination takes place when a collector searching for that species <i>visits</i> that sites in that time interval. A visit can be inferred to have occurred if at least one species record from a site $\times$ time interval combination exists.</p> <p>Site: We partition space into a spatial grid of <i>sites</i>, in which visits can then be inferred to have occurred.</p> <p>Species' range: For a given species, the collection of sites where that species could plausibly be found.</p> <p>Community Sampling: A sampling event where a collector samples the entire community of species. Under this type of sampling, one can infer non-detections of species not found from detections of species that are found.</p> <p>Targeted sampling: A sampling event where a collector primarily seeks a single species. Under this type of sampling, records of non-recorded species should not be inferred from records of recorded species. Targeted and community sampling events are two ends of a continuous spectrum. We do not consider intermediate combinations of these strategies here.</p>

Table 1: Box: Glossary of terms commonly used throughout this manuscript.

Steps for Analyzing Occurrence Data	Example with Odonate data-set
1. Check the probability of community sampling events.	The Odonate data-set had 7451 of unique dates with observations where multiple species were observed.
↓	↓
2. Decide on the spatial and temporal scale of the analysis and the size of the length of eras and time intervals	The temporal range of the analysis is from 1970 to 2020. An era is a period of 10 years, while a <i>time interval</i> is a period of 2 years.
↓	↓
3. Depending on the spatial scale of the analysis, check for overlap in species ranges
↓	↓
4. Check the taxonomic scale and sampling methodology consistent with the taxa in mind, should non-detections be inferred across genus or family level? Then, infer non-detections	We did not infer non-detections differently based on taxonomy.
↓	↓
5. Check whether the visit history increases, decreases or stays constant through eras	The average number of visits for our odonate dataset remained relatively stable in eras 1-4 and decreased slightly in era 5.
↓	↓
6. Remove sites that are only present in a single era. These sites contribute less to estimates of changes in occupancy and do skew the visit history though time.	We found that the Odonate data-set had 5 of sites that were visited in a single era.
↓	↓
7. If the visit history decreases through time and the probability of community sampling events is too low, then re-evaluate from step 2 onwards.	Based on spatiotemporal clustering, we found that our Odonate data-set included 39.7% "community visits" while the rest were single specimen events.

Table 2: Box: Steps that researchers can take to analyse historical record data using occupancy models, illustrated using a data-set of Odonate occurrences.

Supplementary Information

Occupancy Model

We assume that the probability that species i is detected at site j in era k , x_{ijk} , is drawn from a Bernoulli distribution (0 or 1) with probability (y_{ijk}),

$$x_{ijk} \sim \text{Bernoulli}(y_{ijk}) \quad (\text{S1})$$

where y_{ijk} is the product of detection probability (p_{ijk}) and the unknown, but true occupancy state, z_{ijk} ,

$$y_{ijk} = p_{ijk} * z_{ijk} \quad (\text{S2})$$

The true but unknown site occupancy for species i at site j , z_{ijk} is equal to 1 if that site is occupied and 0 if it is not. We assume that this true site occupancy is drawn from a Bernoulli distribution with mean equal to the species' occupancy probability at that site,

$$z_{ijk} \sim \text{Bernoulli}(\psi_{ijk}) \quad (\text{S3})$$

Both occupancy probability, ψ , and detection probability, p , are formulated as functions of time.

We model occupancy as:

$$\begin{aligned} \text{logit}(\psi_{ijk}) = & \psi_0 + \\ & \psi_{\text{sp}}[i] + \\ & \psi_{\text{era}}[i] \times (k - 1) \end{aligned} \tag{S4}$$

where ψ_0 denotes the baseline occupancy (on the linear scale), $\psi_{\text{sp}}[i]$ is a random species-specific intercept, and $\psi_{\text{era}}[i]$ is a random species-specific effect of era (a positive value of $\psi_{\text{era}}[i]$ would indicate that species i is increasing in occupancy through time). We assume that $\psi_{\text{sp}}[i]$ and $\psi_{\text{era}}[i]$ are normally distributed, such that:

$$\begin{aligned} \psi_{\text{sp}}[i] & \sim \mathcal{N}(0, \sigma_{\psi, \text{sp}}) \\ \psi_{\text{era}}[i] & \sim \mathcal{N}(\mu_{\psi, \text{era}}, \sigma_{\psi, \text{era}}) \end{aligned} \tag{S5}$$

Then, we model detection probability as:

$$\begin{aligned} \text{logit}(p_{ijk}) = & p_0 + \\ & p_{\text{sp}}[i] + \\ & p_{\text{site}}[j, k] + \\ & p_{\text{era}} \times (k - 1) \end{aligned} \tag{S6}$$

where p_0 is the baseline detection probability (on the linear scale), $p_{\text{sp}}[i]$ is a random species-specific intercept, $p_{\text{site}}[j, k]$ is a random site-specific intercept that varies by era, and p_{era} is an overall effect of era. $p_{\text{site}}[j, k]$ allows spatiotemporal variability in detection probability (changing across sites and eras), and helps account for the variation that is inherent in sample effort across space and time in opportunistic historical data-sets. p_{era}

allows detection to change systematically through time, as has likely occurred in many groups as sampling techniques have improved, for example. p_{sp} and p_{site} are assumed to be normally distributed, such that:

$$\begin{aligned} p_{\text{sp}}[i] &\sim \mathcal{N}(0, \sigma_{p,\text{sp}}) \\ p_{\text{site}}[j, k] &\sim \mathcal{N}(0, \sigma_{p,\text{site}}), \end{aligned} \tag{S7}$$

We used uninformative priors for all parameter values with normal distributions for ψ_0 , $\mu_{\psi,\text{era}}$, p_0 , p_{era} , and uniform distributions for $\sigma_{\psi,\text{sp}}$, $\sigma_{\psi,\text{era}}$, $\sigma_{p,\text{sp}}$, $\sigma_{p,\text{site}}$.

We ran all models for 100,000 iterations, with a burn in period of 1000 iterations, after burn in iterations were thinned every 100 iterations, and we ran 3 chains of the MCMC. This ensured that across all of our simulations, most parameters converged (r-hat < 1.1), and this convergence was independent of parameter space and model choice.

Parameters varied in the simulations	Values
Number of eras K	2, 5, 10
Probability of community visits ρ_{com}	0, 0.25, 0.5, 0.75, 1
Visitation probability $\mu_{v,\text{era}} / \eta_{\text{era}}$	-0.1, 0, 0.1
Species ranges	no species ranges (all species can be found in all sites), species ranges
Parameters values used in the simulations	Values
Number of species	50
Number of sites	100
Number of visits per era	3
Detection parameters	
p_0	-0.5
$\sigma_{p,\text{site}}$	1.5
$\sigma_{p,\text{sp}}$	0.5
p_{era}	0
Occupancy parameters	
ψ_0	0
$\sigma_{\psi,\text{sp}}$	0.5
$\sigma_{\psi,\text{era}}$	0.2
$\mu_{\psi,\text{era}}$	0
Visitation parameters	
$\mu_{v,0}, \eta_0$	0

Table S1: All of the parameter values used in the simulations of occurrence data.

605 **Supplementary Results**

Figure S1: Estimated change of occupancy through time $\mu_{p,era}$, as a function of changes in detection through time (p_{era}), visit history through time ($\mu_{v,era}$), and the probability of community visits (p_{com}). Points are means across five replicates and error bars are the standard deviation across the replicates. The true value of $\mu_{p,era}$ is represented by the dashed grey line and is equal to zero.

Figure S2: Estimated change of occupancy through time $\mu_{p,era}$, as a function of changes in detection through time (p_{era}), visit history through time ($\mu_{v,era}$), and the probability of community visits (p_{com}). Points are means across five replicates and error bars are the standard deviation across the replicates. The true value of $\mu_{p,era}$ is represented by the dashed grey line and is equal to -0.1.

Figure S3: Estimated change of occupancy through time $\mu_{v,era}$, as a function of changes in detection through time (p_{era}), visit history through time ($\mu_{v,era}$), and the probability of community visits (p_{com}). Points are means across five replicates and error bars are the standard deviation across the replicates. The true value of $\mu_{v,era}$ is represented by the dashed grey line and is equal to 0.1.

Figure S4: Distribution of the proportion of community visits by site/era/time interval combinations for a real-world odonate data-set. Only site/era/time interval with over half of visits as "community sampling events" (greater than 2 species detected based on a spatiotemporal clustering algorithm) are shown.